

DISSERTATION

SPATIAL PROBIT MODELS FOR MULTIVARIATE ORDINAL DATA:  
COMPUTATIONAL EFFICIENCY AND PARAMETER IDENTIFIABILITY

Submitted by

Erin M. Schliep

Department of Statistics

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Summer 2013

Doctoral Committee:

Advisor: Jennifer Hoeting

Daniel Cooley

Myung Hee Lee

Colleen Webb

UMI Number: 3593430

All rights reserved

INFORMATION TO ALL USERS

The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.



UMI 3593430

Published by ProQuest LLC (2013). Copyright in the Dissertation held by the Author.

Microform Edition © ProQuest LLC.

All rights reserved. This work is protected against unauthorized copying under Title 17, United States Code



ProQuest LLC.  
789 East Eisenhower Parkway  
P.O. Box 1346  
Ann Arbor, MI 48106 - 1346

Copyright by Erin M. Schliep 2013

All Rights Reserved

## ABSTRACT

### SPATIAL PROBIT MODELS FOR MULTIVARIATE ORDINAL DATA: COMPUTATIONAL EFFICIENCY AND PARAMETER IDENTIFIABILITY

The Colorado Natural Heritage Program (CNHP) at Colorado State University evaluates Colorado's rare and at-risk species and habitats and promotes conservation of biological resources. One of the goals of the program is to determine the condition of wetlands across the state of Colorado. The data collected are measurements, or metrics, representing landscape condition, biotic condition, hydrologic condition, and physiochemical condition in river basins statewide. The metrics differ in variable type, including binary, ordinal, count, and continuous response data. It is common practice to uniformly discretize the metrics into ordinal values and combine them using a weighted-average to obtain a univariate measure of wetland condition. The weights assigned to each metric are based on best professional judgement.

The motivation of this work was to improve on the user-defined weights by developing a statistical model to estimate the weights using observed data. The challenges of creating a model that fulfills this requirement are many. First, the observed data are multivariate and consist of different variable types which we wish to preserve. Second, the multivariate response data are not independent across river basin because wetlands at close proximity are correlated. Third, we want the model to provide a univariate measure of wetland condition that can be compared across the state. Lastly, it is of interest to the ecologists to predict the univariate measure of wetland condition at unobserved locations requiring covariate information to be incorporated into the model.

We propose a multivariate multilevel latent variable model to address these challenges. Latent continuous response variables are used to model the different types of response variables. An additional latent variable, or common factor, is used as a univariate measure of

wetland condition. The mean of the common factor contains observable covariate data in order to predict at unobserved locations. The variance of the common factor is defined by a spatial covariance function to account for the dependence between wetlands.

The majority of the metrics reported by the CNHP are ordinal. Therefore, our primary focus is modeling multivariate ordinal response data where binary data is a special case. Probit linear models and probit linear mixed models are examples of models for ordinal response data. Probit models are attractive in that they can be defined in terms of latent variables.

Computational efficiency is a major issue when fitting multivariate latent variable models in a Bayesian framework using Markov chain Monte Carlo (MCMC). There is also a high computation cost for running MCMC when fitting geostatistical spatial models. Data augmentation and parameter expansion are both modeling techniques that can lead to optimal iterative sampling algorithms for MCMC. Data augmentation allows for simpler and more feasible simulation from a posterior distribution. Parameter expansion is a method for accelerating convergence of iterative sample algorithms and can enhance data augmentation algorithms. We propose data augmentation and parameter-expanded data augmentation algorithms for fitting MCMC to spatial probit models for binary and ordinal response data.

Parameter identifiability is another challenge when fitting multivariate latent variable models due to the multivariate response data, number of parameters, unobserved latent variables, and spatial random effects. We investigate parameter identifiability for the common factor model for multivariate ordinal response data. We extend the common factor model to include covariates and spatial correlation so we can predict wetland condition at unobserved locations. The partial sill and range parameter of a spatial covariance function are difficult to estimate because they are near-nondetectable. We propose a new parameterization for the covariance function of the spatial probit model that leads to better mixing and faster convergence of the MCMC.

Whereas our spatial probit model for ordinal response data follows the common factor model approach, there are other forms of the spatial probit model. We give a comprehensive comparison of two types of spatial probit models, which we refer to as the first-stage and second-stage spatial probit model. We discuss the implications of fitting each model and compare them in terms of their impact on parameter estimation and prediction at unobserved locations. We propose a new approximation for predicting ordinal response data that is both accurate and efficient.

We apply the multivariate multilevel latent variable model to data collected in the North Platte and Rio Grande River Basins to evaluate wetland condition. We obtain statistically derived weights for each of the response metrics with confidence limits. Lastly, we predict the univariate measure of wetland condition at unobserved locations.

## ACKNOWLEDGEMENTS

There are many people whom I must thank for their expertise and encouragement throughout my graduate work at Colorado State University. First and foremost, I extend great thanks to my advisor and friend, Jennifer Hoeting. There were many times that Jennifer had more confidence in me than I had in myself and without her, I wouldn't have completed a PhD in statistics.

I would like to thank my Master's advisor and PhD committee member, Dan Cooley, who offered me a summer research fellowship at NCAR introducing me to environmental statistics. I thank Colleen Webb for being my PhD committee member and honorary adviser while Jennifer was on sabbatical. She gave me her time and support as well as the opportunity to be a member of the Webb lab as a research assistant on a collaborative project on infectious disease modeling.

The motivation for this dissertation came from the Basinwide Wetland Profile and Condition Assessment project with Joanna Lemly and Laurie Gilligan of the Colorado Natural Heritage Program. I thank them for posing interesting scientific questions and for teaching me about wetlands in Colorado. Part of this work was supported by U.S. Environmental Protection Agency (EPA-1605-09) awarded to the Colorado Heritage Program at Colorado State University.

Lastly, I thank my friends and family. I would specifically like to thank my Tuesday night running friends, Amy Hoseth, Kim Herman, and Whitney Achter, for assisting me in "Keeping America Beautiful" while finishing my dissertation, my great friend Emma Siedschlag for, among many things, making life fun, and my cousins Amy, Paul, Harold, and Guinevere for constant perspective on the important things in life. Finally, I thank my parents and sister for their love and support and for being my number one fans.

## DEDICATION

*This dissertation is dedicated to those who taught me to “work hard, play hard” and supported me through both.*

## TABLE OF CONTENTS

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Motivation for this work	1
1.2	Modeling ordinal data	4
1.3	Generalized and probit linear models	10
<b>2</b>	<b>Data Augmentation and Parameter Expansion for Ordinal, Spatial Data</b>	<b>13</b>
2.1	Introduction	13
2.2	Data Augmentation	15
2.3	Parameter expanded data augmentation	26
2.4	Data augmentation via over-centering	33
2.5	Data augmentation and parameter expansion for spatial probit models	44
2.6	Discussion	57
<b>3</b>	<b>Identifiability</b>	<b>60</b>
3.1	Identifiability, as it applies to frequentists and Bayesians	62
3.2	Identifiability of non-spatial latent variable models	65
3.3	Identifiability of spatial random effect models	79
3.4	Fitting spatial models using MCMC	95
3.5	Discussion	112
<b>4</b>	<b>Comparing and Contrasting First-Stage and Second-Stage Spatial Probit Models</b>	<b>114</b>
4.1	First-stage and second-stage spatial probit models	114
4.2	Comparing first-stage and second-stage model likelihoods via simulation	119
4.3	Latent variable approach to prediction of ordinal response data	125
4.4	Limiting prediction distributions of first-stage and second-stage spatial models	131
4.5	Discussion	140

<b>5 Multilevel Latent Gaussian Process Model for Mixed Discrete and Continuous Multivariate Response Data . . . . .</b>	<b>142</b>
5.1 Introduction . . . . .	142
5.2 Motivating example . . . . .	144
5.3 Model and inference . . . . .	145
5.4 Posterior inference . . . . .	153
5.5 Assessing wetland condition . . . . .	156
5.6 Discussion . . . . .	163
<b>6 Conclusion and Future Work . . . . .</b>	<b>166</b>
6.1 Computational efficiency: Conclusion and future work . . . . .	167
6.2 Identifiability: Conclusion and future work . . . . .	172
6.3 Conclusion . . . . .	175
<b>References . . . . .</b>	<b>177</b>
<b>A Appendix . . . . .</b>	<b>182</b>
A.1 Description of data used in non-spatial probit models . . . . .	182
A.2 Posterior inference . . . . .	184
A.3 Observed Data . . . . .	189
A.4 Squared error loss . . . . .	189
A.5 Multiple correlation . . . . .	191
A.6 Simulation Study . . . . .	197

## CHAPTER 1

### INTRODUCTION

#### 1.1 Motivation for this work

One of the goals of the Colorado Natural Heritage Program (CNHP) is to evaluate the condition of wetlands across the state of Colorado. Many different ecoregions are found in Colorado, and wetland types vary between and within an ecoregion. Field ecologists have spent countless hours collecting measurements, or metrics, representing different ecological categories of wetland condition in river basins statewide. These categories include landscape condition, biotic condition, hydrologic condition, and physiochemical condition. The measurements collected within each of these categories differ in variable type, including binary, ordinal, count, and continuous response data. Many of the metrics are then converted to an ordinal scale for comparison. It is common practice to combine the metrics within each ecological category using a weighted average to get an estimate of each ecological category at each observed location. Then, a second weighted average is computed combining the four ecological categories to obtain an overall wetland condition score. The weights are assigned in both weighted-average computation using best professional judgement.

Our goal was to improve on these user-defined weights by developing a statistical model to estimate the weights using the observed data. There are many challenges in creating a statistical model to meet the needs of the ecologists of the CNHP. First, the observed metrics, or response data are multivariate and may consist of different variable types. Second, the multivariate response data are not independent across wetland and river basin. That is, often wetlands at close proximity within a river basin are correlated. Third, similar to the overall wetland condition score, we want the model to provide a univariate measure of

wetland condition that can be compared across the state. Lastly, it is of interest to the ecologists to predict the univariate wetland condition value at unobserved locations. This entails incorporating covariate information into the model.

We address these challenges in a few ways. First, we use latent variables to incorporate the different types of response data. We represent each observed metric by a continuous latent variable in order to create uniformity in the response variables across metric. We model the univariate measure of wetland condition using a spatially correlated random effect in order to account for dependence between wetland locations. In order to make predictions at new locations, we assume the mean of the univariate measure of wetland condition to be a linear combination of covariates that we are able to obtain over the entire spatial domain.

When fitting the proposed multivariate latent variable model with spatially correlated random effects, we were met with further statistical challenges. Binary, count, and continuous response data are common in the literature and fit within the generalized linear model (GLM) framework. Ordinal data, however, is far less common and require careful consideration when fitting statistical models. The measurements recorded by the CNHP field ecologists for evaluating wetland condition are predominantly ordinal. Therefore, our primary focus in this work is modeling multivariate ordinal response data where binary data is a special case. Probit linear models are one type of model for binary or ordinal response data and will be discussed in detail in Section 1.2.

Computational efficiency is a major issue when fitting multivariate latent variable models in a Bayesian framework using Markov chain Monte Carlo (MCMC). For example, MCMC can result in poor mixing of the threshold parameters when modeling ordinal data. There is also a high computation cost for running MCMC when fitting a geostatistical spatial model. Data augmentation and parameter expansion are both modeling techniques that can lead to optimal iterative sampling algorithms for MCMC. Data augmentation allows for simpler and more feasible simulation from a posterior distribution by conditioning on latent, or augmented data (Tanner and Wong, 1987). Parameter expansion was first introduced by

Liu et al. (1998) as a way to accelerate the convergence of the EM algorithm by increasing the variability between iterations and has been adopted by Bayesians for sampling from the posterior distribution. Parameter-expanded data augmentation further enhances iterative sampling and is an advancement of data augmentation algorithms. We propose data augmentation and parameter-expanded data augmentation algorithms for fitting MCMC to spatial probit models for binary and ordinal response data (Chapter 2).

Parameter identifiability is another issue when fitting the latent variable model due the multivariate response data, number of parameters, unobserved latent variables, and spatial random effects. A model using multiple response variables to obtain inference on a univariate measure closely resembles a *common factor model* (Spearman, 1904). We assume the univariate measure of wetland condition is a latent variable, or common factor, that relates to the multivariate observed data using metric-specific factor loadings. We investigate parameter identifiability for the common factor model for multivariate ordinal response data (Chapter 3). We extend the common factor model to include covariates and spatial correlation so we can predict wetland condition at unobserved locations. The partial sill and range parameter of a spatial covariance function are difficult to estimate. Therefore, we continue our exploration of parameter identifiability to include spatial parameters of the spatial probit model for binary and ordinal response data.

Our spatial probit model for binary and ordinal response data follows the common factor model approach in order to encompass multivariate response data and provide univariate measure for wetland condition across space. However, there are other forms of the spatial probit model (e.g. De Oliveira, 2000). We discuss two different spatial probit model structures and the implications of fitting each model (Chapter 4). We compare the models in terms of their impact on parameter estimation and prediction at unobserved locations. This work includes a new approximation for predicting binary and ordinal response data that is both accurate and efficient.

We return to the data set that motivated this work in Chapter 5. We proposed a multivariate multilevel latent variable model for evaluating the condition of wetlands in Colorado. The model is for mixed ordinal and continuous multivariate data to evaluate a latent spatial Gaussian process. It can be used in many contexts where mixed continuous and discrete multivariate response data are observed in an effort to quantify and unobservable continuous measurement. The model is a modified common factor model in that the common factor is a latent spatial Gaussian process that includes covariate information and spatial correlation. The latent process gives a univariate measure of wetland condition and allows for prediction at unobserved locations. The model is also able to quantify the relationship between the latent process and the response variables allowing us to establish model-inferred weights for each metric. We apply the model to multivariate data collected in the North Platte and Rio Grande River Basins. We conclude the dissertation with a conclusion and future work (Chapter 6).

The remainder of this chapter discusses modeling ordinal response data and its challenges. Section 1.2 gives a general overview of ordinal data and modeling approaches. We define the probit linear model (PLM) that can be used for ordinal response data and is comparable to the GLM. In Section 1.3 we discuss modeling spatially correlated ordinal response data. We discuss generalized linear mixed models (GLMMs) with spatial random effects and define the probit linear mixed model (PLMM) that can be used to model spatially correlated ordinal response data.

## 1.2 Modeling ordinal data

Ordinal data are a type of categorical data where the possible classes of the variable have a distinct order. Some examples of ordinal data include the classification of a university student, (freshman, sophomore, junior, senior), a response to a survey question (strongly disagree, disagree, neutral, agree, strongly agree), or fire danger in a national park (low, moderate, high). Whereas these variables are categorical and could be modeled as such, there

is additional information in knowing that the categories have an inherent order. Therefore, we want to choose from models that are able to account for the order in the response variable to improve on both parameter estimation and prediction.

### 1.2.1 Multinomial distribution approach to ordinal data

We begin with a simple example. Let  $Y$  be an observable ordinal random variable that takes on values in the set  $\{1, \dots, K\}$ . For  $k = 1, \dots, K$ , we can define  $\pi_k$  as

$$\pi_k = P(Y = k)$$

where

$$\sum_{k=1}^K \pi_k = 1.$$

We could assume the random variable  $Y$  follows a multinomial distribution such that

$$Y \sim \text{Multinomial}(1, \pi_1, \dots, \pi_K).$$

For ease of notation, define the vector of probabilities as  $\boldsymbol{\pi} = (\pi_1, \dots, \pi_K)$ .

One of the fundamental difficulties in modeling ordinal data in general is in determining which functional form defines the relationship between the probability of being in each category and the explanatory variable(s). Agresti (2002, Chapter 7) suggests various link functions and models for ordinal and multinomial response data. Models that assume there exist explanatory variables that are driving the ordinal response are known as ordinal regression models. The function used to link  $\boldsymbol{\pi}$  to the explanatory variables is known as a link function. In our example, let  $\mathbf{X}$  be a vector of explanatory variables for random variable  $Y$ . We wish to link  $\mathbf{X}$  to the probabilities of each category,  $(\pi_1, \dots, \pi_K)$ , in a meaningful way. There are many different link functions that can be chosen to model categorical data and yet there are rarely conclusive reasons for choosing one over another. Therefore, it is

common practice to model the data using multiple link functions and to choose the one that results in the best fit of the model to the data (Johnson and Albert, 1999, Chapter 3).

Before describing the different link functions, it is important to note that often ordinal regression models are defined in terms of cumulative probabilities as opposed to the  $K$  individual category probabilities,  $\pi_k$ . This maintains the ordering between ordinal categories. Cumulative probabilities are defined as the probability that an observable random variable is in category  $k$  or below. That is, let

$$\eta_k = P(Y \leq k) = \sum_{j=1}^k \pi_j.$$

Using the cumulative probabilities,  $\boldsymbol{\eta}$ , we can compute the individual category probabilities,  $\boldsymbol{\pi}$ , as

$$\begin{aligned}\pi_k &= P(Y = k) \\ &= P(Y \leq k) - P(Y \leq k-1) \\ &= \eta_k - \eta_{k-1}.\end{aligned}\tag{1}$$

Define the link function,  $F$ , that relates  $\mathbf{X}$  to  $\eta_k$ , for  $k = 1, \dots, K$ , by

$$\eta_k = F(\mathbf{X}).$$

Two common link functions are the logistic link and probit link. In the GLM framework, the logistic link function, also referred to as logit, is defined as

$$\text{logit} \eta_k = \log \left( \frac{\eta_k}{1 - \eta_k} \right) = \lambda_k - \mathbf{X}'\boldsymbol{\beta}\tag{2}$$

where  $\lambda_k$  is a category-specific parameter such that  $\lambda_0 = -\infty$ ,  $\lambda_K = \infty$ , and  $\lambda_k \leq \lambda_{k'}$  for  $k < k'$ . Here,  $\boldsymbol{\beta}$  is a parameter vector of coefficients of the explanatory variables. The logistic

link function stems from logistic regression where the response variable is binary as opposed to ordinal with  $K$  categories. By solving for  $\eta_k$ , we get

$$\eta_k = \frac{\exp\{\lambda_k - \mathbf{X}'\boldsymbol{\beta}\}}{1 + \exp\{\lambda_k - \mathbf{X}'\boldsymbol{\beta}\}}.$$

Note that the GLM is linear in the covariates. When other functional forms of the covariates are assumed, the model fits within the class of generalized additive models (GAMs).

The properties of the cumulative probabilities make cumulative distributions functions a natural class of link functions for ordinal data. These properties are

1.  $\eta_k \in [0, 1]$  for all  $k \in \{1, \dots, K\}$
2.  $\eta_k \leq \eta_{k'}$  for  $k < k'$ ,  $k$  and  $k' \in \{1, \dots, K\}$
3.  $\eta_K = 1$

Let  $F$  define the cumulative distribution function and  $f$  the probability density function of a continuous random variable. We can express the cumulative probability of random variable  $Y$  as

$$P(Y \leq k) = \eta_k = \int_{-\infty}^{\lambda_k} f(z - \mathbf{X}'\boldsymbol{\beta}) dz.$$

The standard normal distribution is a common choice of cumulative distribution function to use as a link function. This link function is referred to as the probit link and the model containing covariate information is called a probit regression model. Here,

$$\eta_k = \Phi(\lambda_k - \mathbf{X}'\boldsymbol{\beta}) \tag{3}$$

where  $\Phi$  is the standard normal CDF and  $\lambda_k$  is a category-specific parameter as defined in 2. Similarly, it can be written as  $\Phi^{-1}(\eta_k) = \lambda_k - \mathbf{X}'\boldsymbol{\beta}$ . When  $Y$  is assumed to be drawn from a multinomial distribution with probability vector,  $\boldsymbol{\pi}$ , the probit regression model assigns probabilities to each of the  $K$  categories by integrating the standard normal CDF.

For example,

$$P(Y = 1) = \Phi(\lambda_1 - \mathbf{X}'\boldsymbol{\beta}) = \int_{-\infty}^{\lambda_1} f(z - \mathbf{X}'\boldsymbol{\beta}) dz$$

where  $f$  is the probability density function of the standard normal distribution defined as

$$f(z) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2\pi}z^2\right\} \quad \text{for } z \in (-\infty, \infty).$$

Similarly, for  $k = 2, \dots, (K - 1)$ ,

$$P(Y = k) = \Phi(\lambda_k - \mathbf{X}'\boldsymbol{\beta}) - \Phi(\lambda_{k-1} - \mathbf{X}'\boldsymbol{\beta}) = \int_{\lambda_{k-1}}^{\lambda_k} f(z - \mathbf{X}'\boldsymbol{\beta}) dz.$$

Lastly,

$$P(Y = K) = 1 - \Phi(\lambda_{J-1} - \mathbf{X}'\boldsymbol{\beta}) = 1 - \int_{-\infty}^{\lambda_{K-1}} f(z - \mathbf{X}'\boldsymbol{\beta}) dz = \int_{\lambda_{K-1}}^{\infty} f(z - \mathbf{X}'\boldsymbol{\beta}) dz.$$

### 1.2.2 Latent variable approach to ordinal data

A convenient alternative parameterization of the multinomial distribution of  $Y$  is through latent variables. Latent variables are unobserved quantities that are often functions or transformations of the variables of interest. Defining a model in terms of latent variables can sometimes ease computation. This is an example of data augmentation which we discuss further in Chapter 2. In ordinal regression, latent variables can be extremely useful and efficient for modeling cumulative probabilities. Another argument for latent variable models for the probit model is that it is reasonable to argue that a latent continuous random variable,  $Z$ , generated the observable ordinal data,  $Y$ . Define  $Z$  as a latent continuous random variable corresponding to the ordinal random variable,  $Y$ . Given the threshold vector,  $\boldsymbol{\lambda}$ , assume there exists a deterministic relationship between  $Y$  and  $Z$  such that

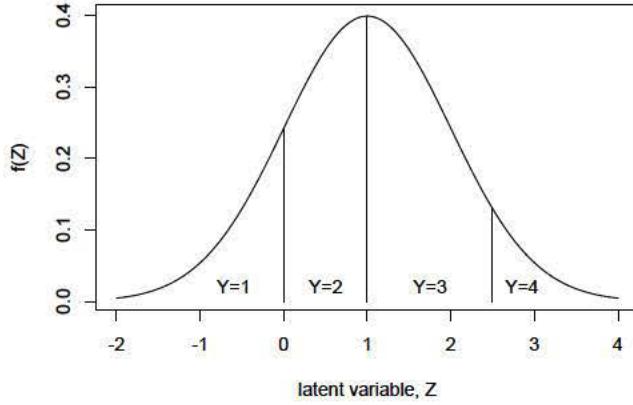


Figure 1.1: Modeling ordinal data with a latent variable.

$$Y = \begin{cases} 1 & \lambda_0 < Z < \lambda_1 \\ 2 & \lambda_1 < Z < \lambda_2 \\ \vdots & \vdots \\ K & \lambda_{K-1} < Z < \lambda_K \end{cases}$$

where

$$-\infty = \lambda_0 \leq \lambda_1 \leq \dots \lambda_K = \infty.$$

Figure 1.1 shows an example of the relationship between  $Y$  and  $Z$  for  $K = 4$  and  $\lambda = (-\infty, 0, 1, 2.5, \infty)$ .

In the latent variable framework, we link the explanatory variables,  $\mathbf{X}$ , to the ordinal response variable through the latent variable. If we assume that  $Z$  is normally distributed with a mean  $\mathbf{X}'\boldsymbol{\beta}$  and variance fixed to 1, the latent variable approach is equivalent to the multinomial regression model with probit link function. That is,

$$P(Y = 1) = P(Z \leq \lambda_1) = \Phi(\lambda_1 - \mathbf{X}'\boldsymbol{\beta}) = \int_{-\infty}^{\lambda_1} f(z - \mathbf{X}'\boldsymbol{\beta}) dz.$$

In general, for  $k \in \{1, \dots, K\}$ ,

$$P(Y = k) = P(Z \leq \lambda_k) - P(Z \leq \lambda_{k-1}) = \Phi(\lambda_k - \mathbf{X}'\boldsymbol{\beta}) - \Phi(\lambda_{k-1} - \mathbf{X}'\boldsymbol{\beta}).$$

One of the advantages of the probit link function is that the latent variable,  $Z$ , is Gaussian. In Chapter 2 we will show that this allows for conjugate updates for some model parameters when modeled in a Bayesian framework. Further, we will show that the latent variable approach to modeling ordinal data using the probit regression model can lead to simpler iterative sampling algorithms for parameter estimation and inference. For each  $k$ ,  $P(Y = k)$  is not unique to the link function or the values of the threshold vector  $\boldsymbol{\lambda}$ . That is, a different link function and different threshold values can preserve the probability of the random variable  $Y$  being in each ordinal category. Therefore, we discuss parameter identifiability for the probit model in Chapter 3.

### 1.3 Generalized and probit linear models

Generalized linear models (GLMs) are a large class of models that generalizes ordinary linear regression to allow for response variables that are not normally distributed. The generalization assumes that the response variable is from a distribution in the exponential family where the mean of the distribution,  $\mu$ , depends on the explanatory variables,  $\mathbf{X}$ , via a link function. That is,

$$E(Y) = g(\mu) = \mathbf{X}'\boldsymbol{\beta} \quad (4)$$

for some link function  $g$ .

GLM models are similar to the probit regression model (3) introduced in Section 1.2. The probit model, however, is only within the class of GLMs for binary data, or ordinal data with two categories. The probit regression model for ordinal data with  $K \geq 3$  is not within the class of generalized linear models. We illustrate the difference between the GLM and probit regression model through an example. Assume  $Y$  is an ordinal random variable with

three possible outcomes, namely,  $Y \in \{1, 2, 3\}$ . We write the expected value of  $Y$  as

$$\begin{aligned} E(Y) &= \sum_{k=1}^3 kP(Y = k) \\ &= 1\Phi(\lambda_1 - \mathbf{X}'\boldsymbol{\beta}) + 2(\Phi(\lambda_2 - \mathbf{X}'\boldsymbol{\beta}) - \Phi(\lambda_1 - \mathbf{X}'\boldsymbol{\beta})) + 3(1 - \Phi(\lambda_2 - \mathbf{X}'\boldsymbol{\beta})) \\ &= 3 - \Phi(\lambda_2 - \mathbf{X}'\boldsymbol{\beta}) - \Phi(\lambda_1 - \mathbf{X}'\boldsymbol{\beta}). \end{aligned}$$

Letting  $E(Y) = \mu$ , there does not exist a function  $g$  such that  $g^{-1}(\mathbf{X}'\boldsymbol{\beta}) = \mu$ . Therefore, the probit regression model with 3 ordinal categories does not fit within the GLM framework. This generalizes to ordinal data with  $K \geq 3$ . When  $K = 2$  such that the data is binary, however, the probit regression model does fit within the GLM framework. Letting  $Y \in \{0, 1\}$ , we can write

$$E(Y) = 0P(Y = 0) + 1P(Y = 1) = P(Y = 1) = 1 - \Phi(\lambda_1 - \mathbf{X}'\boldsymbol{\beta}).$$

For identifiability of the probit model, let  $\lambda_1 = 0$  (see Chapter 3 for further discussion of probit model identifiability). Therefore,

$$E(Y) = \mu = 1 - \Phi(-\mathbf{X}'\boldsymbol{\beta}) = \Phi(\mathbf{X}'\boldsymbol{\beta}).$$

Defining the link function,  $g$ , as  $g = \Phi^{-1}$ , then  $g(\mu) = \mathbf{X}'\boldsymbol{\beta}$ .

Generalized linear mixed models (GLMMs) are an extension to GLMs in that they allow the linear predictor to contain random effects as well as fixed effects. A GLMM with one random effect,  $\alpha$ , can be written as

$$E(Y) = g(\mu) = \mathbf{X}'\boldsymbol{\beta} + \alpha \tag{5}$$

and can extend to include additional random effects. Similarly, we define the probit regression mixed model as an extension to the probit regression model (3) where the cumulative

probability,  $\eta_k$ , for  $k = 1, \dots, K$ , now contains random effects as well as fixed effects. A probit regression mixed model with one random effect,  $\alpha$ , can be written as

$$\eta_k = P(Y \leq k) = \Phi(\lambda_k - \mathbf{X}'\boldsymbol{\beta} - \alpha). \quad (6)$$

Notice that the cumulative probabilities,  $\eta_k$ , for  $k = 1, \dots, K$  are defined as a linear function of the covariates in the probit regression model and a linear function of the covariates and random effect in the probit regression mixed model. Therefore, we refer to the probit regression model and probit regression mixed model for both binary and ordinal response data as a probit linear model (PLM) and probit linear mixed model (PLMM), respectively. In the binary case, the PLM and PLMM are in the class of the GLM and GLMM, respectively.

The remainder of this work focuses on modeling multivariate mixed discrete and continuous response data with emphasis on ordinal response data using the PLM and spatial PLMM. We fit both the PLM and spatial PLMM in the Bayesian framework. We assign prior distributions to the model parameters and use MCMC sampling algorithms to estimate the posterior distribution. In Chapter 2, we present a set of data augmentation and parameter-expanded data augmentation algorithms for spatial and non-spatial binary and ordinal response data. Chapter 3 defines parameter identifiability as it applies to both frequentist and Bayesian inference. We propose parameter constraints for a multivariate latent variable model for ordinal response data. We investigate identifiability and estimability of the parameters of the spatial PLMM for binary and ordinal response data. In Chapter 4 we compare two different forms of spatial probit model for binary and ordinal response data in terms of parameter estimation and prediction. In Chapter 5 we develop a multilevel latent variable model for multivariate response data. The model is applied to mixed ordinal and continuous response data collected as part of the Colorado Department of Wildlife (CDOW) Wetlands Program to create a Basinwide Wetland Profile. Chapter 6 concludes with a discussion and future work.

## CHAPTER 2

# DATA AUGMENTATION AND PARAMETER EXPANSION FOR ORDINAL, SPATIAL DATA

### 2.1 Introduction

The primary goal of data augmentation is constructing an optimal and iterative sampling algorithm by introducing latent or unobserved variables into the model. The approach first became popular within deterministic algorithms for maximizing likelihood functions or posterior densities using the expectation-maximization (EM) algorithm (Dempster et al., 1977). The work of Tanner and Wong (1987) popularized data augmentation within the literature of stochastic algorithms by developing the method for posterior sampling. The schemes were used to make simulating from the posterior distribution simpler and more feasible. The method is known in the physics literature as the method of auxiliary variables. As shown in Swendsen and Wang (1987), the method improves the speed of iterative simulation. In “The Art of Data Augmentation,” Van Dyk and Meng state:

*“Constructing data augmentation schemes that result in both simple and fast algorithms is a matter of art in that successful strategies vary greatly with the observed-data models being considered” (1).*

We agree that constructing a data augmentation algorithm is in fact an art, seeing that there exists many different sampling schemes for even the simplest models.

With the desire to make computations easier and more effective, much work has gone into the advancement of data augmentation algorithms. Liu and Wu (1999) developed what they call a *parameter-expanded data augmentation* (PX-DA) algorithm that further enhances the iterative conditional sampling of Tanner and Wong (1987). Parameter expansion was first

Table 2.1: Outline of data augmentation and parameter-expanded data augmentation algorithms used in fitting spatial and non-spatial probit models to binary and ordinal response data.

	Binary data	Ordinal data
Non-spatial models		
Data augmentation	Algorithm 1: Albert and Chib (1993)	Algorithm 3: Albert and Chib (1993) Algorithm 4: Cowles (1996) Albert and Chib (1997) <b>Algorithm 5</b>
Parameter-expanded data augmentation	Imai and Van Dyk (2005) <b>Algorithm 8</b>	<b>Algorithm 6</b> <b>Algorithm 9</b>
First-stage spatial models		
Data augmentation	Oliveira (2000)	
Parameter-expanded data augmentation	Berrett and Calder (2012)	
Second-stage spatial models		
Data augmentation	<b>Algorithm 10</b>	<b>Algorithm 13</b>
Parameter-expanded data augmentation	<b>Algorithm 11</b> <b>Algorithm 12</b>	<b>Algorithm 14</b> <b>Algorithm 15</b>

introduced by Liu et al. (1998) as a way to accelerate the convergence of the EM algorithm. The idea is that when incorporating missing or latent data into the model, the parameter space of the data model is expanded which can lead to an increased convergence rate of the algorithm. They show that the extra parameters can be introduced in the model without distorting the original observed data model.

In this work, we first outline the idea of data augmentation in the context of Bayesian analysis with the desire to draw iteratively from a posterior distribution. We then define data augmentation as it applies to modeling binary and ordinal response data. Table 2.1 summarizes data augmentation and parameter-expanded data augmentation algorithms for spatial and non-spatial, binary and ordinal response models proposed over the past 20 years. Algorithms in **bold** are those developed in this work. We advance the data augmentation

Table 2.2: Descriptions of acronyms used in data augmentation and parameter-expanded data augmentation algorithms in this chapter.

Acronym	Description
DA	Data augmentation
PX	Parameter expansion
PDA	Partial data augmentation
PX-DA	Parameter-expanded data augmentation
PX-PDA	Parameter-expanded partial data augmentation
PX <sup>2</sup> -PDA	Twice parameter-expanded data augmentation
RV	Random variance
RT	Random threshold

scheme by proposing a set of parameter-expanded approaches for binary and ordinal data. Lastly, we extend the sampling algorithms to probit linear mixed models (PLMMs) for spatially correlated binary and ordinal data. Refer to Table 2.2 for a list of acronyms used in this chapter.

## 2.2 Data Augmentation

The goal of data augmentation strategies within the Bayesian framework are to weaken the dependence between draws from the posterior distribution within a Markov chain Monte Carlo algorithm (Liang et al., 2011). Chains with lower parameter dependence have better mixing and faster convergence. Therefore, models with an elaborate hierarchy structure or a high-dimensional parameter space can greatly benefit from the approach. Data augmentation techniques have been shown to increase the conditional variability of the parameters of interest given the observed (or augmented) data. This leads to larger jumps between draws of the parameters within the chain, and therefore, the parameter space can be explored faster and more efficiently. A second benefit of data augmentation is that it can also lead to drawing from posterior distributions that are known in closed form. For a Gibbs sampling algorithm, this can alleviate the need for Metropolis-Hastings steps in some cases, also leading to a decrease in computation time.

Data augmentation strategies can be used in many varieties in Bayesian analysis. They were first introduced by Tanner and Wong (1987) as a method for dealing with missing values. Royle et al. (2007) developed a data augmentation scheme for multinomial models with unknown population sizes. They augment the data by including a known number of all-zero entries for those subjects not detected (or observed) in the survey and model the augmented dataset as a zero-inflated version of the complete data model. Cauchemez et al. (2004) developed a data-augmented model for estimating transmission characteristics of infectious disease. Here, the observed data are the dates in which new cases of disease are observed and the augmented data include the unobserved dates of the start and end of the infectious period. The augmented data are sampled at each iteration of the MCMC from their posterior distribution given the observed data and parameters. Therefore, data augmentation is a versatile tool as it applies whenever the data can be augmented in such a way that it is easy to both analyze and generate the augmented data given the parameters.

We outline the data augmentation algorithm in the context of a Bayesian analysis where there is missing observable data. This missing data can alternatively be thought of as “latent variables” introduced in latent variable modeling. For example, latent variable models are beneficial when modeling discrete response data, such as binary or multinomial data. In this work we focus on latent variable models for binary and ordinal response data. We begin by defining data augmentation for binary data since it is the simplest case of ordinal data having only two categories. We will then generalize the data augmentation strategy for ordinal data with 3 or more categories.

### 2.2.1 General framework for binary data

Data augmentation is beneficial when you have a complicated likelihood function that is difficult to maximize when doing maximum likelihood estimation. In Bayesian analysis, data augmentation is beneficial when the likelihood function creates a posterior distribution that is not available in closed form. For example, we may be interested in drawing from

the posterior distribution of the parameter,  $\beta$ , which can be a scalar or vector, given the observed data,  $\mathbf{y}$ . The posterior distribution is written as  $p(\beta|\mathbf{y}) \propto p(\mathbf{y}|\beta)p(\beta)$  where  $p(\mathbf{y}|\beta)$  is the likelihood and  $p(\beta)$  is the prior distribution.

When modeling ordinal data, it is difficult to draw from the posterior distribution because it is not known in closed form. This makes data augmentation strategies extremely beneficial. The probit model is a common approach for modeling ordinal data. Here, we consider a two-class ordinal model as a simplified example of ordinal data. Assume the observable data,  $\mathbf{Y}$ , are binary taking on one of two possible outcomes which we denote 0 and 1. Under the probit model,  $P(Y_i = 1|\mathbf{X}_i, \beta) = \Phi(\mathbf{X}'_i \beta)$  where  $\mathbf{X}_i$  is a vector of observable covariates and  $\Phi$  is the CDF of the standard normal distribution. Defining the observed data as  $\mathbf{y} = [y_1, \dots, y_n]$ , the posterior distribution of  $\beta$  given the data,  $\mathbf{y}$ , is given by

$$p(\beta|\mathbf{y}) \propto p(\mathbf{y}|\beta)p(\beta) = \prod_{i=1}^n (\Phi(\mathbf{X}'_i \beta))^{y_i} (1 - \Phi(\mathbf{X}'_i \beta))^{1-y_i} \times p(\beta). \quad (7)$$

This posterior distribution is not available in closed form because the likelihood function contains  $\Phi$ , the standard normal CDF. Therefore, the Gibbs sampler requires a Metropolis-Hastings step. This motivates the desire for an easier scheme for sampling from the posterior distribution.

### 2.2.2 Data augmentation for binary data

Alternatively, the probit model can be defined using latent variables and is an application of data augmentation (Albert and Chib, 1993). The latent variable, defined by  $Z_i$ , is such that  $Z_i = \mathbf{X}'_i \beta + \epsilon_i$  where  $\epsilon_i \sim N(0, 1)$ . The relationship between the latent variable,  $Z_i$ , and the observable random variable  $Y_i$  is such that

$$P(Y_i = 1|\beta) = P(Z_i > 0) = P(Z_i - \mathbf{X}'_i \beta > -\mathbf{X}'_i \beta) = \Phi(\mathbf{X}'_i \beta). \quad (8)$$

Therefore, the model is equivalent to that in (7).

**Algorithm 1.** Two-step Gibbs sampler for binary data:

1. Draw  $\mathbf{Z}^t$  from  $p(\mathbf{Z}|\mathbf{y}, \boldsymbol{\beta}^{t-1})$ .
2. Draw  $\boldsymbol{\beta}^t$  from  $p(\boldsymbol{\beta}|\mathbf{y}, \mathbf{Z}^t)$ .

The data-augmented model (8) has advantages over the original model (7) because we can first write the observed data as

$$p(\mathbf{y}|\boldsymbol{\beta}) = \int_Z p(\mathbf{y}, \mathbf{Z}|\boldsymbol{\beta}) d\mathbf{Z}.$$

In the Bayesian framework using Markov chain Monte Carlo (MCMC), we need to sample  $\boldsymbol{\beta}$  from the posterior distribution  $p(\boldsymbol{\beta}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\beta})p(\boldsymbol{\beta})$ . Notice that the joint distribution of the parameter,  $\boldsymbol{\beta}$ , and the latent variable  $\mathbf{Z}$  can be written as

$$p(\boldsymbol{\beta}, \mathbf{Z}|\mathbf{y}) \propto p(\mathbf{y}, \mathbf{Z}|\boldsymbol{\beta})p(\boldsymbol{\beta}).$$

Therefore, we would like to sample from

$$p(\boldsymbol{\beta}|\mathbf{y}) \propto \int_z p(\mathbf{y}, \mathbf{Z}|\boldsymbol{\beta}) d\mathbf{Z} p(\boldsymbol{\beta}).$$

In general, data augmentation is only beneficial when the conditional distributions of the data-augmented model are easier to sample from than the conditional distributions of the model containing only the observed data. For the probit model, it requires that the two conditional distributions,  $p(\mathbf{Z}|\mathbf{y}, \boldsymbol{\beta}) \propto p(\mathbf{y}, \mathbf{Z}|\boldsymbol{\beta})$  and  $p(\boldsymbol{\beta}|\mathbf{y}, \mathbf{Z}) \propto p(\mathbf{y}, \mathbf{Z}|\boldsymbol{\beta})p(\boldsymbol{\beta})$ , are known in closed form. In such a case, the data augmentation algorithm can follow a two-step Gibbs sampler.

In the case where  $p(\mathbf{Z}|\mathbf{y}, \boldsymbol{\beta})$ ,  $p(\boldsymbol{\beta}|\mathbf{y}, \mathbf{Z})$ , or both, are not known in closed form, the data augmentation approach only further complicates the problem. Therefore, augmenting the

model in a way that helps without hindering can take creativity. When modeling binary observed data,  $\mathbf{y}$ , with the probit model,  $p(\mathbf{Z}|\mathbf{y}, \boldsymbol{\beta}^{t-1})$  is a truncated-normal distribution with mean  $\mathbf{X}\boldsymbol{\beta}$  and variance of 1. The truncation is such that  $Z_i \leq 0$  when  $y_i = 0$  and  $Z_i > 0$  when  $y_i = 1$ . The variance of  $\mathbf{Z}$  is fixed to 1 since it is not identifiable given the data. Fixing the variance to 1 is referred to as conditional data augmentation and will be discussed in Section 2.3.1. When conditioning on the augmented data  $(\mathbf{y}, \mathbf{Z})$ , the multivariate normal distribution is a conjugate prior for the parameter  $\boldsymbol{\beta}$ . Assuming a non-informative prior, we let  $p(\boldsymbol{\beta}) \sim N(\mathbf{0}, \Sigma_{\boldsymbol{\beta}})$  where  $\Sigma_{\boldsymbol{\beta}} = \sigma_{\boldsymbol{\beta}}^2 I$ ,  $\sigma_{\boldsymbol{\beta}}^2$  is large, and  $I$  is the  $(p \times p)$  identity matrix and  $p$  is the length of the vector  $\boldsymbol{\beta}$ . The conditional distribution of  $p(\boldsymbol{\beta}|\mathbf{y}, \mathbf{Z}^t)$  is

$$p(\boldsymbol{\beta}|\mathbf{y}, \mathbf{Z}^t) \sim N((\mathbf{X}'\mathbf{X} + \Sigma_{\boldsymbol{\beta}}^{-1})^{-1}\mathbf{X}'\mathbf{Z}^t, (\mathbf{X}'\mathbf{X} + \Sigma_{\boldsymbol{\beta}}^{-1})^{-1}). \quad (9)$$

As a result, sampling from both conditional distributions in steps 1 and 2 of Algorithm 1 are easy, computationally.

Modeling ordinal data using the probit link function and the data augmentation strategy has other advantages as well. Since the latent variable,  $\mathbf{Z}$ , introduced in the model is normally distributed, it allows for conjugate priors and closed-form posterior distributions for many parameters. The model easily fits within the framework of latent Gaussian models (LGMs). LGMs are a flexible class of models that are easily-interpretable and commonly used in many statistical modeling applications.

### 2.2.3 Data augmentation for ordinal data

The binary response model can be generalized to allow the ordinal response variable to be of the set  $\{1, 2, \dots, K\}$ . Whereas above we used 0 as a threshold for the standard normal CDF to classify  $Y_i$  as either 0 or 1, we now must introduce additional parameters as cut-points (thresholds) into the model. We define the density of observable  $Y_i$ , in terms of the

**Algorithm 2.** *Hybrid Gibbs algorithm for ordinal data:*

1. Sample  $\boldsymbol{\lambda}^t$  from the conditional distribution  $p(\boldsymbol{\lambda}^t | \mathbf{y}, \boldsymbol{\beta}^{t-1})$ .
2. Sample  $\boldsymbol{\beta}^t$  from the conditional distribution  $p(\boldsymbol{\beta}^t | \mathbf{y}, \boldsymbol{\lambda}^t)$ .

covariates,  $\mathbf{X}_i$ , coefficients  $\boldsymbol{\beta}$ , and thresholds,  $\boldsymbol{\lambda}$ , as

$$P(Y_i = k) = \Phi(\lambda_k - \mathbf{X}'_i \boldsymbol{\beta}) - \Phi(\lambda_{k-1} - \mathbf{X}'_i \boldsymbol{\beta}) \quad (10)$$

where  $\boldsymbol{\lambda}$  is the vector of cut points such that  $-\infty = \lambda_0 < \lambda_1 \leq \dots \lambda_{K-1} < \lambda_K = \infty$ . For identifiability of the intercept term of the coefficient vector,  $\boldsymbol{\beta}$ , we fix the first cut-point,  $\lambda_1 = 0$ .

To draw inference on the parameters, we again need to assign prior distributions to both  $\boldsymbol{\beta}$  and  $\boldsymbol{\lambda}$ . Our goal is to sample from the posterior distribution  $p(\boldsymbol{\beta}, \boldsymbol{\lambda} | \mathbf{y})$ . Using Gibbs sampling, we are able to sample iteratively from the conditional distributions of  $\boldsymbol{\beta}$  and  $\boldsymbol{\lambda}$  separately. Therefore, we wish to sample according to Algorithm 2.

To apply data augmentation methods outlined above, we need to introduce the latent variable  $\mathbf{Z}$ . We define the deterministic relationship between the latent variable and observable ordinal response data using (4) where

$$P(Y_i = k) = P(\lambda_{k-1} < Z_i \leq \lambda_k)$$

for  $k = 1, \dots, K$  and  $\boldsymbol{\lambda} = (-\infty = \lambda_0 \leq \lambda_1 \leq \dots \lambda_K = \infty)$ .

**Algorithm 3.** *Data-augmented Gibbs sampler for ordinal data:*

1. Draw  $\mathbf{Z}^t$  from  $p(\mathbf{Z}|\mathbf{y}, \boldsymbol{\beta}^{t-1}, \boldsymbol{\lambda}^{t-1})$ .
2. Draw  $\boldsymbol{\lambda}^t$  from  $p(\boldsymbol{\lambda}|\mathbf{y}, \mathbf{Z}^t, \boldsymbol{\beta}^{t-1})$ .
3. Draw  $\boldsymbol{\beta}^t$  from  $p(\boldsymbol{\beta}|\mathbf{y}, \mathbf{Z}^t, \boldsymbol{\lambda}^t)$ .

Once again, we assume  $Z_i = \mathbf{X}'_i \boldsymbol{\beta} + \epsilon_i$  where  $\epsilon_i \sim N(0, 1)$ . Following the notation from the binary response model, we can write out the density of  $Y_i$  in terms of  $Z_i$  as

$$\begin{aligned} P(Y_i = k) &= P(\lambda_{k-1} < Z_i \leq \lambda_k) = P(Z_i \leq \lambda_k) - P(Z_i \leq \lambda_{k-1}) \\ &= \Phi(\lambda_k - \mathbf{X}'_i \boldsymbol{\beta}) - \Phi(\lambda_{k-1} - \mathbf{X}'_i \boldsymbol{\beta}). \end{aligned} \tag{11}$$

Therefore, the data-augmented likelihood is equivalent to the likelihood for the observed data model given in (10).

Using the augmented data approach, there are several sampling algorithms for drawing inference. One approach would be to use a Gibbs sampler where  $\mathbf{Z}, \boldsymbol{\lambda}, \boldsymbol{\beta}$  are all drawn from their full conditional distributions (Albert and Chib, 1993). This algorithm, referred to as the *data-augmented Gibbs sampler*, is given in Algorithm 3.

Whereas under the observed-data model where both  $\boldsymbol{\lambda}$  and  $\boldsymbol{\beta}$  would require Metropolis-Hastings steps using Algorithm 2, this augmented data approach has a major advantage in that the conditional distribution of  $\boldsymbol{\beta}$  is known in closed form. Since the latent data,  $\mathbf{Z}$ , is multivariate normal, this extends beyond the ease of drawing  $\boldsymbol{\beta}$  as it also allows for conjugate prior distributions for other variables, such as random effects, one might be interested in including in the model.

Unfortunately, in practice there is also a significant limitation to the data-augmented Gibbs sampler. The Markov chain for  $\boldsymbol{\lambda}$  can be extremely slow to mix due to the constraints on the parameters. For example, not only is  $\lambda_k$  forced to be greater than  $\lambda_{k-1}$  and less than

$\lambda_{k+1}$ , but when conditional on  $\mathbf{Z}$ ,  $\lambda_k$  must be within the interval

$$[\max\{\lambda_{k-1}, \{Z_i : Y_i = k\}\}, \min\{\lambda_{k+1}, \{Z_i : Y_i = k + 1\}\}].$$

This range can be tremendously limiting, therefore, restricting the movement between the parameter vector  $\boldsymbol{\lambda}$  at iteration  $t$  and iteration  $t + 1$ . Notice that the restrictive space is an issue regardless of the prior distribution assigned to  $\boldsymbol{\lambda}$ . Therefore, implementing the Gibbs sampler can be ineffective in practice.

Cowles (1996) extends the data augmentation approach of Albert and Chib (1993) by using the continuous latent variable approach for fitting the ordinal probit model. She demonstrates that the convergence of Algorithm 3 may be slow when the sample size is large. Therefore, she proposes a multivariate Hastings-within-Gibbs step that accelerates the convergence of the Markov chain. The sampling scheme encompasses a “grouping” or “blocking” approach that usually improves the efficiency of a Gibbs sampler (Roberts and Sahu, 1997). The blocking is applied by updating  $\mathbf{Z}$  and  $\boldsymbol{\lambda}$  jointly. Then,  $\boldsymbol{\beta}$  is updated from its complete conditional distribution as before.

The blocking scheme for  $(\mathbf{Z}, \boldsymbol{\lambda})$  is implemented by first writing the joint distribution as

$$p(\mathbf{Z}, \boldsymbol{\lambda} | \mathbf{y}, \boldsymbol{\beta}) \propto p(\mathbf{Z} | \boldsymbol{\lambda}, \mathbf{y}, \boldsymbol{\beta}) p(\boldsymbol{\lambda} | \mathbf{y}, \boldsymbol{\beta}).$$

The posterior density  $p(\boldsymbol{\lambda} | \mathbf{y}, \boldsymbol{\beta})$  is the original posterior density of the observed data model in Algorithm 2. The target density can be written as  $p(\boldsymbol{\lambda} | \mathbf{y}, \boldsymbol{\beta}) \propto p(\mathbf{y} | \boldsymbol{\lambda}, \boldsymbol{\beta}) p(\boldsymbol{\lambda})$ , assuming prior independence between  $\boldsymbol{\lambda}$  and  $\boldsymbol{\beta}$ . This updating step will require a Metropolis-Hastings step within the Gibbs sampler since

$$p(\mathbf{y} | \boldsymbol{\lambda}, \boldsymbol{\beta}) = \prod_{i=1}^n (\Phi(\lambda_k - \mathbf{X}'_i \boldsymbol{\beta}) - \Phi(\lambda_{k-1} - \mathbf{X}'_i \boldsymbol{\beta})).$$

**Algorithm 4.** *Cowles algorithm for ordinal data:*

1. Draw  $(\mathbf{Z}^t, \boldsymbol{\lambda}^t)$  from  $p(\mathbf{Z}, \boldsymbol{\lambda}|y, \boldsymbol{\beta}^{t-1})$ .
  - (a) Draw  $\boldsymbol{\lambda}^c$  from  $g(\boldsymbol{\lambda}^c|\boldsymbol{\lambda}^{t-1})$ .
  - (b) Compute the Metropolis-Hastings ratio
 
$$a = \min \left\{ 1, \frac{f(\boldsymbol{\lambda}^c)g(\boldsymbol{\lambda}^{t-1}|\boldsymbol{\lambda}^c)}{f(\boldsymbol{\lambda}^{t-1})g(\boldsymbol{\lambda}^c|\boldsymbol{\lambda}^{t-1})} \right\}. \quad (12)$$
  - (c) With probability  $a$ ,
    - i. Set  $\boldsymbol{\lambda}^t = \boldsymbol{\lambda}^c$
    - ii. Draw  $\mathbf{Z}^t$  from  $p(\mathbf{Z}|y, \boldsymbol{\beta}^{t-1}, \boldsymbol{\lambda}^t)$ .
  - (d) With probability  $1 - a$ , set  $\boldsymbol{\lambda}^t = \boldsymbol{\lambda}^{t-1}$  and  $\mathbf{Z}^t = \mathbf{Z}^{t-1}$ .
2. Draw  $\boldsymbol{\beta}^t$  from  $p(\boldsymbol{\beta}|y, \mathbf{Z}^t, \boldsymbol{\lambda}^t)$ .

To give a general idea of this approach, we let  $f$  be the target density we wish to sample from and  $g$  the proposal distribution for  $\boldsymbol{\lambda}$ . The *Cowles algorithm* and block update is outlined in Algorithm 4.

In Algorithm 4, the complete conditional distribution of  $Z_i$ , where  $Z_i$  and  $Z_j$ , for  $i \neq j$ , are conditionally independent given parameters and data, is the truncated-normal. Letting  $\text{TN}(\mu, \sigma^2, \lambda_{lower}, \lambda_{upper})$  specify a normal distribution with mean  $\mu$  and variance  $\sigma^2$  truncated between  $\lambda_{lower}$  and  $\lambda_{upper}$ , the second part of (c) draws  $Z_i^t$  from  $TN(\mathbf{X}'_i \boldsymbol{\beta}^{t-1}, 1, \lambda_{y_{i-1}}^t, \lambda_{y_i}^t)$  for  $i \in 1, \dots, N$ . Lastly, the algorithm samples  $\boldsymbol{\beta}$  from its complete conditional distribution. The blocking scheme increases the variability within the chain for the threshold parameter vector,  $\boldsymbol{\lambda}$ , by not including  $\mathbf{Z}$  in the conditional distribution in which  $\boldsymbol{\lambda}$  is updated. The algorithm is called a multivariate Hastings-within-Gibbs algorithm because  $\boldsymbol{\lambda}$  and  $\mathbf{Z}$  are updated jointly based on acceptance probability,  $a$ . Even though the overall acceptance probability for the latent parameter in the Cowles algorithm is less than that of the Gibbs algorithm (assuming  $a < 1$  for at least 1 iteration), the convergence of the chain is still improved when  $\boldsymbol{\lambda}$  has

better mixing. Therefore, the Cowles algorithm (A.4) is highly recommended over the data-augmented Gibbs algorithm (A.3).

The Cowles algorithm, and more specifically, the mixing of  $\boldsymbol{\lambda}$ , is enhanced further in the work of Albert and Chib (1997) by applying a transformation to the thresholds. Constraining the threshold parameter vector,  $\boldsymbol{\lambda}$ , such that  $\lambda_{k-1} \leq \lambda_k$  leads to poor mixing of the Markov chain for this parameter. Therefore, Albert and Chib (1997) propose transforming  $\boldsymbol{\lambda}$  by setting  $\boldsymbol{\alpha} = g(\boldsymbol{\lambda})$ , where the function  $g(\boldsymbol{\lambda})$  is such that  $\alpha_1 = \lambda_1 = 0$  and  $\alpha_k = \log(\lambda_k - \lambda_{k-1})$  for  $k = 2, \dots, K - 1$ . The unconstrained threshold parameter vector  $\boldsymbol{\alpha}$  is modelled  $p(\boldsymbol{\alpha}) \sim N(\boldsymbol{a}_0, \boldsymbol{A}_0)$ . As shown above, sampling  $\boldsymbol{\lambda}^t$  given  $\mathbf{y}$  and  $\boldsymbol{\beta}$  requires a Metropolis-Hastings step since the posterior is not known in closed form. We first transform  $\boldsymbol{\lambda}$  to  $\boldsymbol{\alpha}$  as described above. By defining the likelihood, proposal, and prior distributions all in terms of  $\boldsymbol{\alpha}$ , we don't need a Jacobian in the M-H ratio. The M-H ratio contains the densities  $p(\mathbf{y}|\boldsymbol{\alpha}, \boldsymbol{\beta})$  and  $p(\boldsymbol{\alpha})$ , as well the the proposal distribution,  $f(\boldsymbol{\alpha}^c|\boldsymbol{\alpha})$ , where  $\boldsymbol{\alpha}^c$  is the candidate of  $\boldsymbol{\alpha}$ . At step  $t$ , the algorithm moves to the candidate parameter,  $\boldsymbol{\alpha}^c$ , from the current value,  $\boldsymbol{\alpha}$ , with transition probability

$$a^* = \min \left\{ \frac{p(\mathbf{y}|\boldsymbol{\alpha}^c, \boldsymbol{\beta}^{t-1})f(\boldsymbol{\alpha}|\boldsymbol{\alpha}^c)p(\boldsymbol{\alpha}^c)}{p(\mathbf{y}|\boldsymbol{\alpha}, \boldsymbol{\beta}^{t-1})f(\boldsymbol{\alpha}^c|\boldsymbol{\alpha})p(\boldsymbol{\alpha})}, 1 \right\}. \quad (13)$$

We transform back from  $\boldsymbol{\alpha}^t$  to  $\boldsymbol{\lambda}^t$  via  $\boldsymbol{\lambda}^t = g^{-1}(\boldsymbol{\alpha}^t)$ . This acceptance probability,  $a^*$ , can replace the acceptance probability in (12) and has been shown to further improve convergence of the Markov chain.

Thus far, the main benefit of data augmentation is that it allows us to easily draw the parameter  $\boldsymbol{\beta}$  from its full conditional distribution. This will become even more useful as the model is made more complex by incorporating spatial or temporal correlation between observations. Unfortunately, we also know that the mixing of the chain for the threshold parameter,  $\boldsymbol{\lambda}$ , is much more efficient using the un-augmented data approach. If we were to include the latent variable  $\mathbf{Z}$  in the conditional distribution of  $\boldsymbol{\lambda}$ , as in the Gibbs sampler, we

**Algorithm 5.** *Partial data-augmented algorithm (PDA) for ordinal data:*

1. Draw  $\boldsymbol{\lambda}^t$  from  $p(\boldsymbol{\lambda}|\mathbf{y}, \boldsymbol{\beta}^{t-1})$ .
2. Draw  $\boldsymbol{\beta}^t$  from  $p(\boldsymbol{\beta}|\mathbf{y}, \boldsymbol{\lambda}^t)$ .
  - (a) Draw  $\mathbf{Z}^t$  from  $p(\mathbf{Z}|\mathbf{y}, \boldsymbol{\beta}^{t-1}, \boldsymbol{\lambda}^t)$ .
  - (b) Draw  $\boldsymbol{\beta}^t$  from  $p(\boldsymbol{\beta}|\mathbf{y}, \mathbf{Z}^t, \boldsymbol{\lambda}^t)$ .

need to sample  $\boldsymbol{\lambda}^t$  from  $p(\boldsymbol{\lambda}^t|\mathbf{y}, \mathbf{Z}^t, \boldsymbol{\beta}^{t-1})$ . Even when transforming the thresholds vector into the unconstrained vector  $\boldsymbol{\alpha}$ , this would require the M-H algorithm to include the conditional density  $p(\mathbf{Z}|\mathbf{y}, \boldsymbol{\alpha}^c, \boldsymbol{\beta}^{t-1})$  and  $p(\mathbf{Z}|\mathbf{y}, \boldsymbol{\alpha}, \boldsymbol{\beta}^{t-1})$  in the numerator and denominator, respectively. This is an issue because  $p(\mathbf{Z}|\mathbf{y}, \boldsymbol{\alpha}^c, \boldsymbol{\beta}^{t-1})$  can easily be 0 when the candidate threshold vector shifts causing the current value of  $\mathbf{Z}$  to be in the incorrect class according to the observed data  $\mathbf{y}$ . When this happens, the acceptance probability goes to 0 and the chain can get stuck for a long duration of iterations. Therefore, the ideal sampling algorithm would:

1. Update  $\boldsymbol{\lambda}$  using the transformation  $g(\boldsymbol{\lambda}) = \boldsymbol{\alpha}$ .
2. Sample  $\boldsymbol{\lambda}$  from  $p(\boldsymbol{\lambda}|\mathbf{y}, \boldsymbol{\beta})$  as opposed to  $p(\boldsymbol{\lambda}|\mathbf{y}, \mathbf{Z}, \boldsymbol{\beta})$ .
3. Avoid updating  $\boldsymbol{\lambda}$  and  $\mathbf{Z}$  as a block because this forces  $\mathbf{Z}$  to be updated with the Metropolis-Hastings step of  $\boldsymbol{\lambda}$ .
4. Sample  $\boldsymbol{\beta}$  conditional on both the observed data  $\mathbf{y}$  and the augmented data  $\mathbf{Z}$ .

We proposed the *partial data-augmented* (PDA) algorithm to achieve these goals. It combines the un-augmented sampling scheme given in Algorithm 2 with the Gibbs scheme in Algorithm 3. This leads to the following algorithm.

The first step of Algorithm 5 consists of updating  $\boldsymbol{\lambda}$  using the original likelihood of the observed data given in Algorithm 2. We can apply the transformation approach and update  $\boldsymbol{\lambda}$  using the Metropolis-Hastings acceptance probability in equation (13). In the second step,

we introduce the latent variable,  $\mathbf{Z}$ , and implement the data-augmented two-step Gibbs sampler (Tanner and Wong, 1987). We use the two-step Gibbs sampling algorithm to first sample  $\mathbf{Z}^t$  from  $p(\mathbf{Z}|\mathbf{y}, \boldsymbol{\beta}^{t-1}, \boldsymbol{\lambda}^t)$  and then sample  $\boldsymbol{\beta}^t$  from  $p(\boldsymbol{\beta}|\mathbf{y}, \mathbf{Z}^t, \boldsymbol{\lambda}^t)$ . The distribution of  $p(\mathbf{Z}|\mathbf{y}, \boldsymbol{\beta}^{t-1}, \boldsymbol{\lambda}^t)$  is a truncated-normal distribution with the new threshold values from step 1. Assuming a multivariate normal prior distribution, the posterior distribution of  $\boldsymbol{\beta}$  given the observed data, latent variable, and threshold parameter is the same as that given in (9).

This sampling scheme was employed in Albert and Chib (1997). However, in that work, they said the scheme used the blocking approach for updating  $\mathbf{Z}$  and  $\boldsymbol{\lambda}$ . Therefore, their evidence of both increased speed and mixing of the chain is not the result of only the transformation of  $\boldsymbol{\lambda}$ . It also is the result of sampling  $\mathbf{Z}$  at every iteration, independent of the acceptance probability of the Metropolis-Hastings algorithm computed for moving from  $\boldsymbol{\alpha}^{t-1}$  to the candidate value  $\boldsymbol{\alpha}^*$ , and thus, outside of the block update of  $\mathbf{Z}$  and  $\boldsymbol{\lambda}$ . We want to be clear that in the partial data augmentation algorithm, none of the parameters are being updated in a block. The PDA algorithm has further benefit as we introduce parameter-expanded schemes for data augmentation in Section (2.3).

## 2.3 Parameter expanded data augmentation

In certain settings, the Gibbs sampler can be slow to converge. One cause can be high dependence between realizations from the conditional posterior distribution in a Gibbs sampler. As discussed above, data augmentation schemes are motivated by the need for alternative distributions that allow sampling in closed form as compared to the original posterior distribution. Parameter expansion algorithms are an extension of data augmentation algorithms. They are shown to improve the convergence of the Markov chain by increasing the size of the parameter space. The two approaches fit together nicely in that the parameters that are introduced through parameter expansion are not identified by the observed data but they are identified by the latent variables introduced in data augmentation. It is important to

note that in order to implement parameter-expanded data augmentation (PX-DA), knowledge about the identifiability of the model parameters is extremely important. There are a considerable number of parameter expansion schemes available and determining which is the best for the data and model at hand is as much an art as data augmentation. Imai and Van Dyk (2005) present a set of data augmentation schemes using parameter expansion for multinomial response data. Their model is modified by Berrett and Calder (2012) to handle spatially correlated binary data. In this work, we explore various PX-DA approaches for ordinal data. Our goal is to determine which PX-DA scheme is the most suitable for probit regression for spatially correlated ordinal data.

### 2.3.1 Variance parameter approach

In the data augmentation algorithm for ordinal data, we defined the latent variable  $\mathbf{Z}$ , such that  $Z_i = \mathbf{X}'_i \boldsymbol{\beta} + \epsilon_i$  where  $\epsilon_i \stackrel{iid}{\sim} N(0, 1)$  for  $i = 1, \dots, n$ . The reason we have set  $\text{var}(\epsilon_i) = 1$  is because the variance is not identifiable from ordinal data. This can be seen by writing out the density of the observed value,  $Y_i$ . Assume  $Z_i^* = X'_i \boldsymbol{\beta} + \epsilon_i^*$  where  $\epsilon_i^* \sim N(0, \sigma^2)$

$$\begin{aligned}
P(Y_i = k) &= P(\lambda_{k-1} < Z_i^* \leq \lambda_k) = P(Z_i^* \leq \lambda_k) - P(Z_i^* \leq \lambda_{k-1}) \\
&= \Phi\left(\frac{\lambda_k - \mathbf{X}'_i \boldsymbol{\beta}}{\sigma}\right) - \Phi\left(\frac{\lambda_{k-1} - \mathbf{X}'_i \boldsymbol{\beta}}{\sigma}\right) \\
&= \Phi(\lambda_k^* - \mathbf{X}'_i \boldsymbol{\beta}^*) - \Phi(\lambda_{k-1}^* - \mathbf{X}'_i \boldsymbol{\beta}^*) \\
&= P(Z_i \leq \lambda_k) - P(Z_i \leq \lambda_{k-1}).
\end{aligned} \tag{14}$$

where  $\lambda_k^* = \lambda_k/\sigma$ ,  $\lambda_{k-1}^* = \lambda_{k-1}/\sigma$ , and  $\boldsymbol{\beta}^* = \frac{1}{\sigma} \boldsymbol{\beta}$ . Therefore, the thresholds and coefficients are both only identifiable up to a constant.

Since  $\sigma^2$  is not identifiable, we can condition on any value for the variance of  $\epsilon_i$ . Meng and Van Dyk (1999) proved that a model that marginalizes over  $\sigma^2$  is more diffuse than a model that conditions on a fixed value for  $\sigma^2$ . The optimality of marginalizing over  $\sigma^2$  has been shown in both the binomial and multinomial model. The data augmentation strategies

introduced by Imai and Van Dyk (2005) and Berrett and Calder (2012) both allow  $\sigma^2$  to vary throughout the MCMC algorithm, while the unidentified parameter is integrated out at each iteration. We note that both Imai and Van Dyk (2005) and Berrett and Calder (2012) propose two different parameter-expanded schemes for fitting multinomial and spatially correlated binary data, respectively. The first draws  $\sigma^2$  once within the MCMC while the second algorithm draws  $\sigma^2$  twice. We do not extend their first parameter-expanded algorithm to the ordinal, spatial setting since they showed the second to be optimal in terms of convergence and autocorrelation. Therefore, we develop our first PX-PDA algorithm that generalizes our PDA approach by allowing  $\epsilon \sim N(0, \sigma^2)$ . Following the same notation as the earlier work, we model

$$\tilde{\mathbf{Z}} \sim N(\mathbf{X}\tilde{\boldsymbol{\beta}}, \sigma^2\mathbf{I})$$

where  $\tilde{\cdot}$  represents the unidentified form of the latent variable or parameter. The identified model is written as

$$\mathbf{Z} \sim N(\mathbf{X}\boldsymbol{\beta}, I).$$

We define the following relationships between the identified and unidentified parameter sets:

$$\mathbf{Z} = \tilde{\mathbf{Z}}/\sigma, \quad \boldsymbol{\beta} = \tilde{\boldsymbol{\beta}}/\sigma, \quad \text{and} \quad \boldsymbol{\lambda} = \tilde{\boldsymbol{\lambda}}/\sigma.$$

The work of Imai and Van Dyk (2005) and Berrett and Calder (2012) suggests that we run our Markov chain using  $\sigma^2$  and the unidentified parameters  $\tilde{\mathbf{Z}}$ ,  $\tilde{\boldsymbol{\beta}}$ , and  $\tilde{\boldsymbol{\lambda}}$ . The estimates of the identified parameters are then obtained by marginalizing over  $\sigma^2$ . Since the major advantage of parameter expansion is to increase the variation between draws in the Markov chain, we maximize efficiency by running the entire chain using the unidentified parameters. For the threshold parameter,  $\boldsymbol{\lambda}$ , this suggests that at iteration  $t$  we draw the unidentified parameter,  $\tilde{\boldsymbol{\lambda}}^t$ , and then obtain  $\boldsymbol{\lambda}^t$  by setting  $\boldsymbol{\lambda}^t = \tilde{\boldsymbol{\lambda}}^t/\sigma^t$ . However, recall from Section 2.2.3 and Algorithm 5, that we do not want to sample the threshold vector  $\boldsymbol{\lambda}$  conditional on the

**Algorithm 6.** *RV-PX-PDA algorithm for ordinal data:*

1. Draw  $\boldsymbol{\lambda}^t$  from  $p(\boldsymbol{\lambda}|\mathbf{y}, \boldsymbol{\beta}^{t-1})$ .
2. Draw  $\boldsymbol{\beta}^t$  from  $p(\boldsymbol{\beta}|\mathbf{y}, \boldsymbol{\lambda}^t)$ .
  - (a) Draw  $(\tilde{\mathbf{Z}}^t, (\sigma^2)^*)$  from  $p(\tilde{\mathbf{Z}}, \sigma^2|\mathbf{y}, \boldsymbol{\beta}^{t-1}, \boldsymbol{\lambda}^t)$ .
    - i. Draw  $(\sigma^2)^*$  from  $p(\sigma^2|\mathbf{y}, \tilde{\boldsymbol{\beta}}^{t-1}, \boldsymbol{\lambda}^t) \sim p(\sigma^2)$ .
    - ii. Draw  $\tilde{\mathbf{Z}}^t$  from  $p(\tilde{\mathbf{Z}}|\mathbf{y}, \boldsymbol{\beta}^{t-1}, \boldsymbol{\lambda}^t, (\sigma^2)^*)$   
 $\sim TN(\mathbf{X}(\boldsymbol{\beta}^{t-1}\sigma^*), (\sigma^2)^*, \sigma^*\boldsymbol{\lambda}_{\mathbf{y}-1}^t, \sigma^*\boldsymbol{\lambda}_{\mathbf{y}}^t)$ . Set  $\mathbf{Z}^t = \tilde{\mathbf{Z}}^t / \sigma^*$ .
  - (b) Draw  $(\tilde{\boldsymbol{\beta}}^t, (\sigma^2)^t)$  from  $p(\tilde{\boldsymbol{\beta}}, \sigma^2|\mathbf{y}, \tilde{\mathbf{Z}}^t, \boldsymbol{\lambda}^t)$ .
    - i. Draw  $(\sigma^2)^t$  from  $p(\sigma^2|\mathbf{y}, \tilde{\mathbf{Z}}^t, \boldsymbol{\lambda}^t)$ .
    - ii. Draw  $\tilde{\boldsymbol{\beta}}^t$  from  $p(\tilde{\boldsymbol{\beta}}|\mathbf{y}, \tilde{\mathbf{Z}}^t, \boldsymbol{\lambda}^t, (\sigma^2)^t) \sim N(\mathbf{b}_p, \mathbf{B}_p)$  where  
 $\mathbf{b}_p = (\mathbf{X}'\mathbf{X} + \Sigma_{\boldsymbol{\beta}}^{-1})^{-1}\mathbf{X}'\tilde{\mathbf{Z}}^t$  and  $\mathbf{B}_p = (\mathbf{X}'\mathbf{X} + \Sigma_{\boldsymbol{\beta}}^{-1})^{-1}$ .  
Set  $\boldsymbol{\beta}^t = \tilde{\boldsymbol{\beta}}^t / \sigma^t$ .

augmented data. Therefore, we modify the framework of Imai and Van Dyk (2005) and Berrett and Calder (2012) in order to develop a PDA approach that is preferable for ordinal data. Parameter expansion for ordinal data is simplified using PDA because we already have one transformation of the threshold vector  $\boldsymbol{\lambda}$  to  $\boldsymbol{\alpha}$  and it eliminates the need for a second transformation to the unidentified thresholds,  $\tilde{\boldsymbol{\lambda}}$  and  $\tilde{\boldsymbol{\alpha}}$ .

We will refer to our first parameter-expanded data augmentation algorithm as the random variance parameter-expanded partial data augmentation algorithm (RV-PX-PDA). We will outline the sampling algorithm for drawing from the parameter-expanded posterior distribution. The RV-PX-PDA strategy is once again a hybrid Gibbs sampler with Metropolis-Hastings steps. For clarity, we diverge from the labels marginal and conditional data augmentation used by Imai and Van Dyk (2005) and Berrett and Calder (2012) since our sampling scheme is a combination of both data augmentation and parameter expansion. However, where possible, we will make connections between our algorithm and their marginal augmentation schemes.

The first step in RV-PX-PDA is the same as the first step in Algorithm 5. Since RV-PX-PDA diverges from the previous algorithm in step two, we will outline this step in more detail. Recall that the goal of the second step is to draw  $\beta^t$  from  $p(\beta|\mathbf{y}, \lambda^t)$ . In the PDA algorithm, this is extended to the two-step Gibbs sampler by drawing  $\mathbf{Z}^t$  from  $p(\mathbf{Z}|\mathbf{y}, \beta^{t-1}, \lambda^t)$  and  $\beta^t$  from  $p(\beta|\mathbf{y}, \mathbf{Z}^t, \lambda^t)$ . Here, we want to include the variance parameter  $\sigma^2$  in our updating scheme to increase the variability within the Markov chain. Therefore, using our augmented data in unidentified parameter form, we want to jointly draw  $(\tilde{\mathbf{Z}}, (\sigma^2)^*)$  and  $(\tilde{\beta}, \sigma^2)$ . Since  $\sigma^2$  is drawn in both steps 2(a) and 2(b), we denote the first draw as  $(\sigma^2)^*$  and the second as  $(\sigma^2)^t$ , where  $(\sigma^2)^t$  is the value that is being carried forward to the next iteration. The reason we make note of this is because both Imai and Van Dyk (2005) and Berrett and Calder (2012) define two different sampling algorithms for drawing  $(\sigma^2)^*$ . We are adopting their “Marginal Augmentation Scheme 1” as it was shown to be optimal. In contrast, at iteration  $t$ , their Scheme 2 sets  $(\sigma^2)^* = (\sigma^2)^{t-1}$  as opposed to drawing a new value.

We write the conditional posterior distribution required for part (a) of step 2 in Algorithm 6 as

$$p(\tilde{\mathbf{Z}}, (\sigma^2)^* | \mathbf{y}, \beta, \lambda) \propto p(\tilde{\mathbf{Z}} | \mathbf{y}, \beta, \lambda, (\sigma^2)^*) p((\sigma^2)^* | \mathbf{y}, \lambda, \beta).$$

Therefore, we can update  $\mathbf{Z}^t$  by first drawing  $(\sigma^2)^*$  from  $p((\sigma^2)^* | \mathbf{y}, \lambda^t, \beta^{t-1})$ . Assuming  $\lambda, \beta$ , and  $\sigma^2$  are independent,  $p((\sigma^2)^* | \mathbf{y}, \lambda, \beta) = p(\sigma^2)$  where  $p(\sigma^2)$  is the prior distribution of  $\sigma^2$  since  $\sigma^2$  is unidentifiable in the observed-data model. Next we are able to draw  $\tilde{\mathbf{Z}}^t$  from its complete conditional distribution which can be shown to follow a truncated-normal distribution. For  $i, j \in 1, \dots, n, i \neq j$ ,  $\tilde{Z}_i$  and  $\tilde{Z}_j$  are conditionally independent given  $\mathbf{X}$  and have distribution

$$\tilde{Z}_i^t = TN(\mathbf{X}_i \beta^{t-1} \sigma^*, (\sigma^2)^*, \sigma^* \lambda_{y_i-1}^t, \sigma^* \lambda_{y_i}^t).$$

We finish the step by setting  $\mathbf{Z}^t = \tilde{\mathbf{Z}}^t / \sigma^*$ .

In part (b) of step 2, we wish to draw  $(\boldsymbol{\beta}^t, (\sigma^2)^t)$ . The conditional posterior distribution required in this update can be written as

$$p(\boldsymbol{\beta}, \sigma^2 | \mathbf{y}, \tilde{\mathbf{Z}}, \boldsymbol{\lambda}) \propto p(\boldsymbol{\beta} | \mathbf{y}, \tilde{\mathbf{Z}}, \boldsymbol{\lambda}, \sigma^2) p(\sigma^2 | \mathbf{y}, \tilde{\mathbf{Z}}, \boldsymbol{\lambda}).$$

Once again, we will begin by first drawing  $\sigma^2$  given  $\mathbf{y}$ ,  $\tilde{\mathbf{Z}}$ , and  $\boldsymbol{\lambda}$ . Using Bayes' Theorem, we have

$$\begin{aligned} p(\sigma^2 | \mathbf{y}, \tilde{\mathbf{Z}}, \boldsymbol{\lambda}) &\propto p(\mathbf{y} | \tilde{\mathbf{Z}}, \boldsymbol{\lambda}, \sigma^2) p(\tilde{\mathbf{Z}} | \boldsymbol{\lambda}, \sigma^2) p(\boldsymbol{\lambda} | \sigma^2) p(\sigma^2) \\ &\propto p(\tilde{\mathbf{Z}} | \boldsymbol{\lambda}, \sigma^2) p(\sigma^2) \\ &= \int p(\tilde{\mathbf{Z}}, \tilde{\boldsymbol{\beta}} | \boldsymbol{\lambda}, \sigma^2) p(\sigma^2) d\tilde{\boldsymbol{\beta}} \\ &= p(\sigma^2) \int p(\tilde{\mathbf{Z}} | \tilde{\boldsymbol{\beta}}, \boldsymbol{\lambda}, \sigma^2) p(\tilde{\boldsymbol{\beta}} | \boldsymbol{\lambda}, \sigma^2) d\tilde{\boldsymbol{\beta}} \\ &= p(\sigma^2) \int p(\tilde{\mathbf{Z}} | \tilde{\boldsymbol{\beta}}, \sigma^2) p(\tilde{\boldsymbol{\beta}} | \sigma^2) d\tilde{\boldsymbol{\beta}}. \end{aligned} \tag{15}$$

Under other sampling algorithms (i.e., when  $\boldsymbol{\lambda}$  is not updated outside the data augmentation),  $p(\tilde{\boldsymbol{\lambda}} | \sigma^2)$  will not drop out of the derivation adding further complications to the desired posterior distribution. Here, the integral  $\int p(\tilde{\mathbf{Z}} | \tilde{\boldsymbol{\beta}}, \sigma^2) p(\tilde{\boldsymbol{\beta}} | \sigma^2) d\tilde{\boldsymbol{\beta}}$  can be evaluated by completing the square of the multivariate normal distribution. The resulting distribution is normal with  $\sigma^2$  as a scaling parameter of the variance-covariance matrix. Therefore, we can write the posterior density in (15) as

$$p(\sigma^2 | \mathbf{y}, \tilde{\mathbf{Z}}, \boldsymbol{\lambda}) \propto p(\sigma^2) \frac{1}{(\sigma^2)^{n/2}} \exp \left[ -\frac{1}{2\sigma^2} \left( (\tilde{\mathbf{Z}})' \tilde{\mathbf{Z}} - \tilde{\mathbf{Z}} \mathbf{X} (\mathbf{X}' \mathbf{X} + \Sigma_{\boldsymbol{\beta}}^{-1})^{-1} \mathbf{X}' \tilde{\mathbf{Z}} \right) \right].$$

The conjugate prior distribution is  $p(\sigma^2) \sim \text{Inv. Gamma}(\alpha_s, \beta_s)$ . We draw  $(\sigma^2)^t$  from the resulting posterior distribution:

$$p(\sigma^2 | \mathbf{y}, \tilde{\mathbf{Z}}^t, \boldsymbol{\lambda}^t) \sim \text{Inv. Gamma} \left( \alpha_s + \frac{n}{2}, \beta_s + \frac{1}{2} \left( (\tilde{\mathbf{Z}}^t)' \tilde{\mathbf{Z}}^t - \tilde{\mathbf{Z}}^t \mathbf{X} (\mathbf{X}' \mathbf{X} + \Sigma_{\beta}^{-1})^{-1} \mathbf{X}' \tilde{\mathbf{Z}}^t \right) \right).$$

Further details on this posterior derivation are included in Appendix A.2.1. We can then draw  $\tilde{\boldsymbol{\beta}}^t$  given  $(\mathbf{y}, \tilde{\mathbf{Z}}^t, \boldsymbol{\lambda}^t, (\sigma^2)^t)$  from its complete conditional distribution,

$$p(\tilde{\boldsymbol{\beta}} | \mathbf{y}, \tilde{\mathbf{Z}}^t, \boldsymbol{\lambda}^t, (\sigma^2)^t) \sim N(\mathbf{b}_p, \mathbf{B}_p)$$

where  $\mathbf{b}_p$  and  $\mathbf{B}_p$  are given in Algorithm 6. The algorithm finishes by setting  $\boldsymbol{\beta}^t = \tilde{\boldsymbol{\beta}}^t / \sigma^t$ .

We apply PDA (A.5) and RV-PX-PDA (A.6) to ordinal response data with  $K = 5$  categories. Therefore, we have three identified threshold parameters,  $\lambda_2$ ,  $\lambda_3$ , and  $\lambda_4$ , to estimate, assuming  $\lambda_1 = 0$ . We also include two covariates as fixed effects in the model, resulting in an intercept term,  $\beta_0$ , and coefficients  $\beta_1$  and  $\beta_2$ . The data consist of  $n = 232$  observations (See Section A.1 for further details of the data). We run Algorithms 5 and 6 for 100,000 iterations each, disregarding the first 10,000 as burn-in. The effective sample size is calculated to approximate the number of independent posterior draws of the parameter from the MCMC algorithm (Givens and Hoeting, 2012). Table 2.3 reports the effective sample size for the two algorithms, indicating that the RV-PX-PDA algorithm outperforms the PDA algorithm for all parameters. Autocorrelation plots are shown in Figures 2.1 and 2.2 for the threshold vector,  $\boldsymbol{\lambda}$ , and coefficient vector,  $\boldsymbol{\beta}$ , respectively. Lower autocorrelation values within the chains indicate that both  $\boldsymbol{\lambda}$  and  $\boldsymbol{\beta}$  are able to more easily move around when being initially drawn in the nonidentifiable parameter space than in the identifiable parameter space. This illustrates that the increased variation within the chains of the nonidentifiable parameters,  $\tilde{\mathbf{Z}}$  and  $\tilde{\boldsymbol{\beta}}$ , increases the variation within the identified parameters. As a result, the RV-PX-PDA algorithm will converge more quickly. It is a particularly interesting result

Table 2.3: Effective sample size for  $\beta$  and  $\lambda$  using PDA (A.5) and RV-PX-PDA (A.6) for ordinal data with five categories for 90,000 iterations.

Parameter	Algorithms	
	PDA (A.5)	RV-PX-PDA (A.6)
$\beta_0$	3,807	26,263
$\beta_1$	17,114	40,824
$\beta_2$	19,564	30,690
$\lambda_2$	4,906	7,656
$\lambda_3$	3,317	6,548
$\lambda_4$	3,009	7,165

that the threshold parameter,  $\lambda$ , has lower autocorrelation using the RV-PX-PDA algorithm since it is not updated using the augmented data.

## 2.4 Data augmentation via over-centering

Our second method of parameter-expanded data augmentation adopts the over-centering approach given in Liu and Wu (1999). In their simple example, they define the observed-data model  $p(y|\theta) \sim N(\theta, 1 + D)$  where  $D$  is known. Whereas the observed data are multidimensional random variables in general, they are kept as one-dimension in this example. The complete data model (i.e., data-augmented model), where  $y$  is the observed response and  $Z$  is the latent variable, is defined as

$$p(y|\theta, Z) \sim N(\theta + Z, 1), \quad p(Z|\theta) \sim N(0, D). \quad (16)$$

Forcing  $Z$  to have mean 0 causes high correlation between  $\theta$  and  $Z$ , thus, slowing down convergence of the MCMC algorithm. Over-parameterizing the model such that

$$p(y|\theta, Z, \alpha) \sim N(\theta - \alpha + Z, 1), \quad \text{and} \quad p(Z|\theta, \alpha) \sim N(\alpha, D) \quad (17)$$

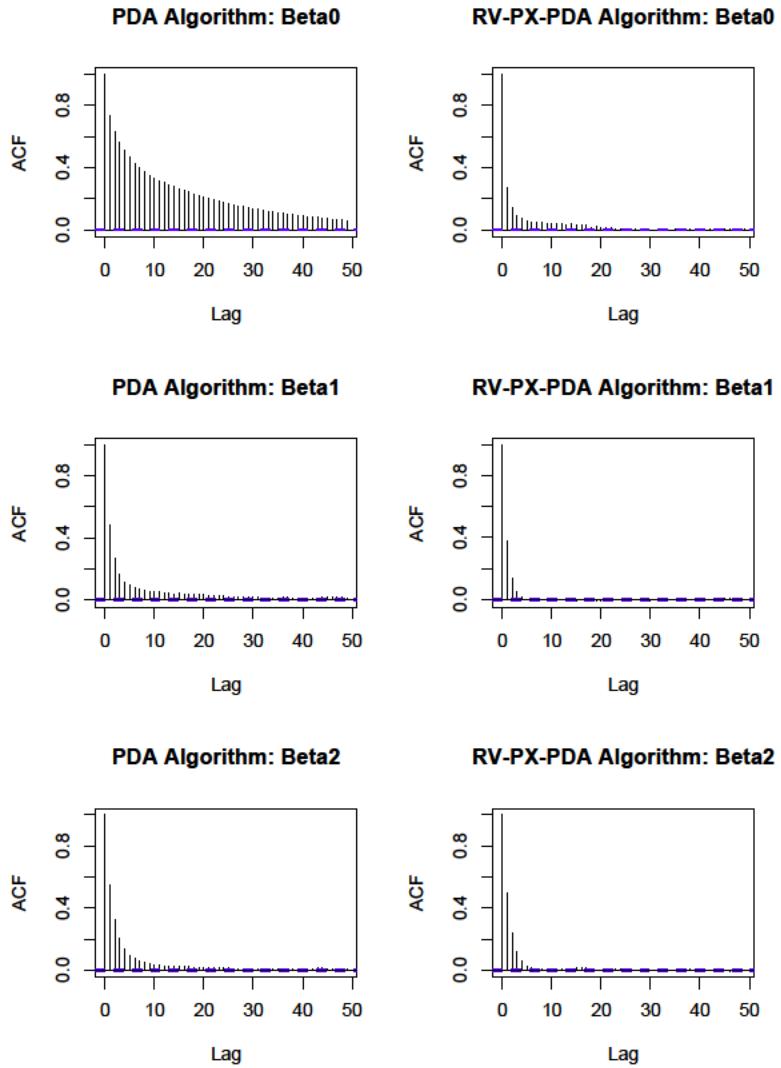


Figure 2.1: Autocorrelation plots for the coefficient vector,  $\beta$ , using PDA (A.5) and RV-PX-PDA (A.6).

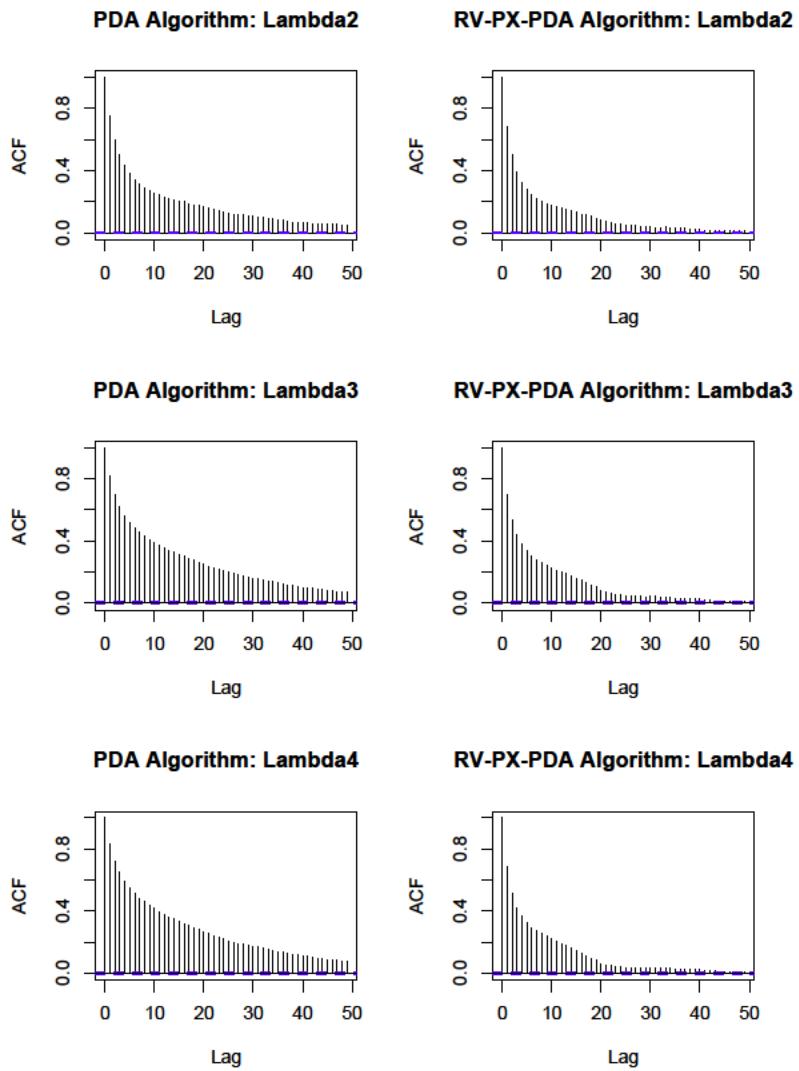


Figure 2.2: Autocorrelation plots for the threshold vector,  $\lambda$ , using PDA (A.5) and RV-PX-PDA (A.6).

**Algorithm 7.** Simple over-centering algorithm for Gaussian data:

1. Draw  $(\mathbf{Z}^t, \alpha^*)$  conditional on  $(\theta^{t-1}, \mathbf{y})$ .
  - (a) Draw  $\alpha^*$  from  $p(\alpha|\mathbf{y}, \theta^{t-1}) \sim N(0, A)$ .
  - (b) Draw  $\mathbf{Z}^t$  from  $p(\mathbf{Z}|\mathbf{y}, \theta^{t-1}, \alpha^*) \sim N\left(\frac{\mathbf{y} - \theta^{t-1}}{1+D^{-1}} + \alpha^*, \frac{1}{1+D^{-1}}\right)$ .
2. Draw  $(\theta^t, \alpha^t)$  conditional on  $(\mathbf{y}, \mathbf{Z}^t)$ .
  - (a) Draw  $\alpha^t$  from  $p(\alpha|\mathbf{y}, \mathbf{Z}^t) \sim N\left(\frac{A\mathbf{Z}^t}{A+D}, \frac{1}{A^{-1}+D^{-1}}\right)$ .
  - (b) Draw  $\theta^t$  from  $p(\theta|\mathbf{y}, \mathbf{Z}^t, \alpha^t) \sim N(\mathbf{y} - \mathbf{Z}^t + \alpha^t, 1)$ .

decreases the correlation between  $Z$  and  $\theta$  while still preserving the observed-data model.

The expansion parameter,  $\alpha$ , is identifiable only for the complete data  $(y, Z)$ . Assuming  $\alpha$  has prior distribution  $\alpha \sim N(0, A)$ , they define the PX-DA algorithm (A.7).

This sampling scheme is shown to greatly improve the rate of convergence and yet is equivalent to the original model.

#### 2.4.1 Random threshold approach for binary data

We will use the idea of over-centering to improve the convergence of the ordinal data model. First, we outline the data-augmented model and sampling algorithm using the binary data model outlined above. Recall that the complete data consists of the observed data,  $\mathbf{y}$ , and latent data,  $\mathbf{Z}$ . In our model, we assume  $Z_i = \lambda + \mathbf{X}'_i \boldsymbol{\beta} + \epsilon_i$  where  $\epsilon_i \sim N(0, 1)$ . The complete conditional distribution is truncated-normal with mean  $\lambda + \mathbf{X}'_i \boldsymbol{\beta}$ , variance of 1, and truncation point,  $\lambda$ . This modifies the original data augmentation model in Section (2.2.3) where the truncation point was fixed at 0. Notice that the complete-data model given in

**Algorithm 8.** *RT-PX-DA algorithm for binary data:*

1. Draw  $(\mathbf{Z}^t, \lambda^*)$  conditional on  $(\boldsymbol{\beta}, \mathbf{y})$ .
  - (a) Draw  $\lambda^*$  from  $p(\lambda|\mathbf{y}, \boldsymbol{\beta}^{t-1}) \sim TN(0, L)$ .
  - (b) Draw  $\mathbf{Z}^t$  from  $p(\mathbf{Z}|\mathbf{y}, \boldsymbol{\beta}^{t-1}, \lambda^*) \sim TN(\lambda^* + \mathbf{X}\boldsymbol{\beta}, 1, \lambda_{\mathbf{y}}^*, \lambda_{\mathbf{y}+1}^*)$ .
2. Draw  $(\boldsymbol{\beta}^t, \lambda^t)$  conditional on  $(\mathbf{y}, \mathbf{Z}^t)$ .
  - (a) Draw  $\lambda^t$  from  $p(\lambda|\mathbf{y}, \mathbf{Z}^t) \sim TN(\mu_\lambda, \tau_\lambda, l_\lambda, u_\lambda)$ , where  
 $\tau_\lambda = (L^{-1} + n - \mathbf{1}'\mathbf{X}(\mathbf{X}'\mathbf{X} + \Sigma_\beta^{-1})^{-1}\mathbf{X}'\mathbf{1})^{-1}$ ,  
 $\mu_\lambda = \tau_\lambda (\mathbf{1}'\mathbf{Z}^t - \mathbf{1}'\mathbf{X}(\mathbf{X}'\mathbf{X} + \Sigma_\beta^{-1})^{-1}\mathbf{X}'\mathbf{Z}^t)$ ,  
 $l_\lambda = \max\{Z_i^t : i \in C_0\}$ , and  $u_\lambda = \min\{Z_i^t : i \in C_1\}$ .
  - (b) Draw  $\boldsymbol{\beta}^t$  from  $p(\boldsymbol{\beta}|\mathbf{y}, \mathbf{Z}^t, \lambda^t) \sim N((\mathbf{X}'\mathbf{X} + \Sigma_\beta^{-1})^{-1}\mathbf{X}'(\mathbf{Z}^t - \mathbf{1}\lambda^t), (\mathbf{X}'\mathbf{X} + \Sigma_\beta^{-1})^{-1})$

equation (8) is now written as

$$P(Y_i = 1|\boldsymbol{\beta}) = P(Z_i > \lambda) = P(Z_i - \lambda - \mathbf{X}'_i\boldsymbol{\beta} > \lambda - \lambda - \mathbf{X}'_i\boldsymbol{\beta}) = \Phi(\mathbf{X}'_i\boldsymbol{\beta}),$$

which is the same as the observed data model  $p(\mathbf{y}|\boldsymbol{\beta})$ . Therefore, the model has been over-centered by letting  $\lambda$  be both in the mean of  $\mathbf{Z}$  and the threshold value that classifies the binary observable response variable,  $\mathbf{y}$ . Assign normal prior distributions to both  $\lambda$  and  $\boldsymbol{\beta}$  such that  $\lambda \sim N(0, L)$  and  $\boldsymbol{\beta} \sim N(\mathbf{0}, \Sigma_\beta)$ . Let  $C_0 = \{i : Y_i = 0\}$  and  $C_1 = \{i : Y_i = 1\}$  for  $i \in 1, \dots, n$ . We define the random threshold parameter-expanded data augmentation algorithm (RT-PX-DA) for binary response data in Algorithm 8.

Using a modified form of the ordinal response data (see Section A.1), we run Algorithms 1 and 8 for 10,000 iterations each and disregard the first 1,000 for burn-in. For this example, the effective sample sizes (Table 2.4) and autocorrelation plots (Figure 2.3) are very similar for the RT-PX-DA algorithm and the two-step Gibbs algorithm for the mixing of the Markov chain for  $\boldsymbol{\beta}$ . This indicates that there is no improvement in convergence for the random threshold parameter-expanded data augmentation algorithm.

Table 2.4: Effective sample size for  $\beta$  for the two-step Gibbs sampler (A.1) and RT-PX-DA (A.8) for 9,000 MCMC iterations.

Parameter	Algorithms	
	Two-step Gibbs sampler (A.1)	RT-PX-DA (A.8)
$\beta_0$	1,481	1,473
$\beta_1$	1,287	1,261
$\beta_2$	981	810

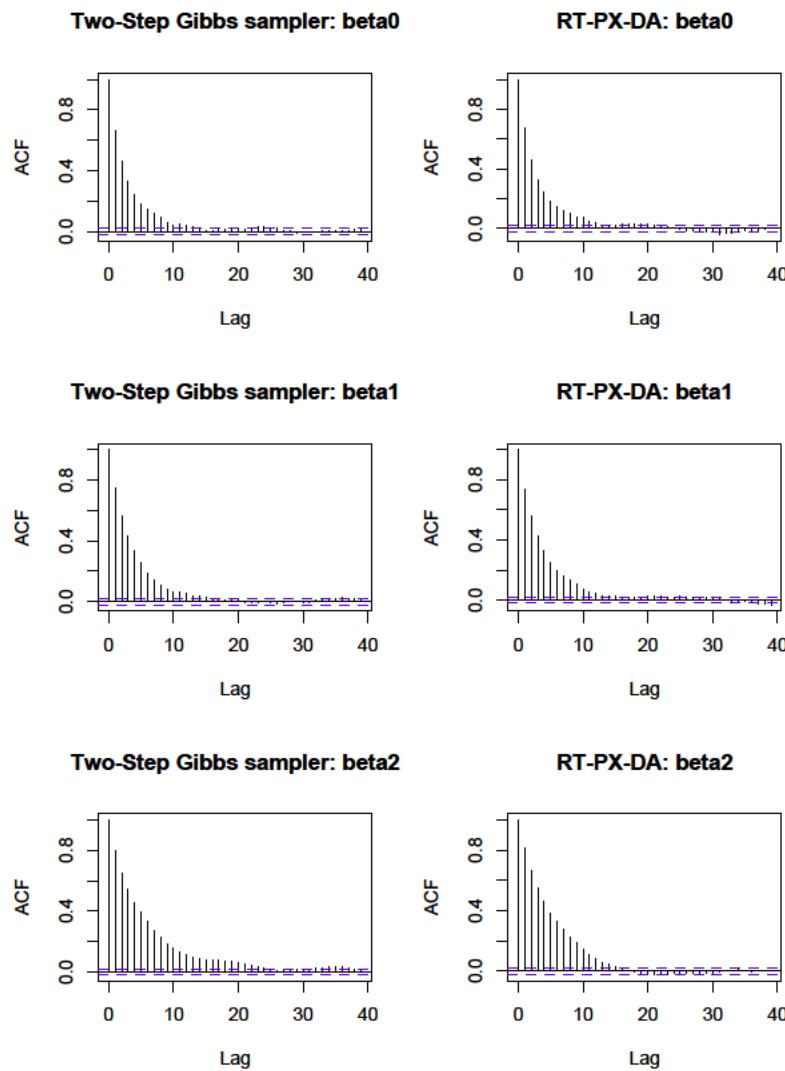


Figure 2.3: Autocorrelation plots for the coefficient vector,  $\beta$ , for the two-step Gibbs sampler (A.1) and RT-PX-DA (A.8) for 9,000 MCMC iterations.

### 2.4.2 Unconstrained threshold approach for ordinal data

We extend the varying threshold model in Algorithm 8 to allow for ordinal data with more than two categories. Assume, without loss of generality, that the response variable  $\mathbf{y} \in \{1, \dots, 5\}$ . In this case, we need four threshold parameters. In the PDA algorithm (Section 2.2.3), we defined the threshold vector  $\boldsymbol{\lambda}$  such that  $\lambda_0 = -\infty$ ,  $\lambda_1 = 0$ , and  $\lambda_5 = \infty$ . Therefore, we were left to estimate  $\lambda_2$ ,  $\lambda_3$ , and  $\lambda_4$ . We now would like to define a random threshold algorithm for ordinal data that allows  $\lambda_1$  to vary. Our new algorithm, which we refer to as RT-PX-PDA, is constructed as a combination of RT-PX-DA (A.8) and PDA (A.5).

Since  $\lambda_1$  is a parameter only in the augmented data model but not the observed data model, we ignore this unidentifiable parameter in step 1 by drawing  $[\lambda_2, \lambda_3, \lambda_4]$  given the observed data model. This step follows the transformation approach and Metropolis-Hastings step given in (13). We can then bring  $\lambda_1$  into the sampling algorithm in step 2 when we augment the model by introducing the latent response data,  $\mathbf{Z}$ .

Whereas the previous algorithms have been outlined in terms of the threshold parameter  $\boldsymbol{\lambda}$ , we define RT-PX-PDA in terms of the transformed parameter vector  $\boldsymbol{\alpha}$ . Recall that in the identified model where  $\lambda_1$  is fixed at 0,  $\boldsymbol{\alpha} = g(\boldsymbol{\lambda})$  with  $\alpha_1 = \lambda_1$  and  $\alpha_k = \log(\lambda_k - \lambda_{k-1})$  for  $k = 2, 3, 4..$ . Therefore, we have a deterministic relationship between  $\boldsymbol{\lambda}$  and  $\boldsymbol{\alpha}$  in the identified model. In allowing  $\lambda_1$  to vary, we have the same functional form where  $\alpha_1 = \lambda_1$  and  $\alpha_k$  is the log-distance between  $\lambda_k$  and  $\lambda_{k-1}$  for  $k = 2, 3, 4$ . For ease of notation, we will refer to  $\alpha_1$  as the varying threshold in the augmented model. The transformed threshold vector  $\boldsymbol{\alpha}$  will consist of the three identifiable threshold values such that  $\boldsymbol{\alpha} = [\alpha_2, \alpha_3, \alpha_4]$ . Therefore,  $(\alpha_1, \boldsymbol{\alpha}) = g(\boldsymbol{\lambda})$  where  $\boldsymbol{\lambda} = [\lambda_1, \lambda_2, \lambda_3, \lambda_4]$ .

Reported in Table 2.5 are the effective sample sizes for RT-PX-PDA compared to RV-PX-PDA and PDA from Table 2.3. RV-PX-PDA is the superior of the three algorithms in terms of effective sample size, but RT-PX-PDA also outperforms PDA for all model parameters.

**Algorithm 9.** *RT-PX-PDA algorithm for ordinal data:*

1. Draw  $\boldsymbol{\alpha}^t = [\alpha_2, \dots, \alpha_K]$  from the conditional distribution  $p(\boldsymbol{\alpha}|\mathbf{y}, \boldsymbol{\beta}^{t-1})$ .
2. Draw  $\boldsymbol{\beta}^t$  from the conditional distribution  $p(\boldsymbol{\beta}|\mathbf{y}, \boldsymbol{\alpha}^t)$ .
  - (a) Draw  $(\mathbf{Z}^t, \alpha_1^*)$  conditional on  $(\mathbf{y}, \boldsymbol{\beta}^{t-1}, \boldsymbol{\alpha}^t)$ .
    - i. Draw  $\alpha_1^*$  from  $p(\alpha_1|\mathbf{y}, \boldsymbol{\beta}^{t-1}, \boldsymbol{\alpha}^t) \sim N(0, L)$ .
    - ii. Draw  $\mathbf{Z}^t$  from  $p(\mathbf{Z}|\mathbf{y}, \boldsymbol{\beta}^{t-1}, \alpha_1^*, \boldsymbol{\alpha}^t) \sim TN(\alpha_1^* + X\boldsymbol{\beta}^{t-1}, 1, \lambda_{\mathbf{y}-1}^*, \lambda_{\mathbf{y}}^*)$  where  $\boldsymbol{\lambda}^* = g^{-1}(\alpha_1^*, \boldsymbol{\alpha}^t)$ .
  - (b) Draw  $(\beta^t, \alpha_1^t)$  conditional on  $(\mathbf{y}, \mathbf{Z}^t, \boldsymbol{\alpha}^t)$ .
    - i. Draw  $\alpha_1^t$  from  $p(\alpha_1|\mathbf{y}, \mathbf{Z}^t, \boldsymbol{\alpha}^t) \sim TN(\mu_\alpha, \tau_\alpha, l_\alpha, u_\alpha)$  where  $\tau_\alpha = (L^{-1} + n - \mathbf{1}'\mathbf{X}(\mathbf{X}'\mathbf{X} + \Sigma_\beta^{-1})^{-1}\mathbf{X}'\mathbf{1})^{-1}$ ,  $\mu_\alpha = \tau_\alpha (\mathbf{1}'\mathbf{Z}^t - \mathbf{1}'\mathbf{X}(\mathbf{X}'\mathbf{X} + \Sigma_\beta^{-1})^{-1}\mathbf{X}'\mathbf{Z}^t)$ ,  $l_\alpha = \max\{Z_i^t : i \in C_1\}$ , and  $u_\alpha = \min\{Z_i^t : i \in C_2\}$ .
    - ii. Draw  $\boldsymbol{\beta}^t$  from  $p(\boldsymbol{\beta}|\mathbf{y}, \mathbf{Z}^t, \alpha_1^t, \boldsymbol{\alpha}^t) \sim N((\mathbf{X}'\mathbf{X} + \Sigma_\beta^{-1})^{-1}\mathbf{X}'(\mathbf{Z}^t - \mathbf{1}\alpha_1^t), (\mathbf{X}'\mathbf{X} + \Sigma_\beta^{-1})^{-1})$ .

This is because extra variability is introduced in sampling the first threshold parameter,  $\alpha_1$ . Figures 2.4 and 2.5 give the autocorrelation plots for the coefficient and threshold vectors, respectively. We conclude that RV-PX-PDA has better mixing, and thus, faster convergence than both alternative sampling schemes. Table 2.6 gives the autocorrelations of the sample paths for  $\beta_1$  for the three algorithms we are comparing. The autocorrelations are very similar between the two lower-performing algorithms, PDA and RT-PX-PDA. RV-PX-PDA has the lowest autocorrelations across all lags, which coincides with it being the better of the algorithms in terms of effective sample size.

Table 2.5: Effective sample size for  $\beta$  and  $\lambda$  for PDA (A.5) and both PX-PDA algorithms, (A.6 and 9), for ordinal data with five categories.

Parameter	Algorithms		
	PDA (A.5)	RV-PX-PDA (A.6)	RT-PX-PDA (A.9)
$\beta_0$	3,807	26,263	4,415
$\beta_1$	17,114	40,824	17,407
$\beta_2$	19,564	30,690	20,786
$\lambda_2$	4,906	7,656	5,642
$\lambda_3$	3,317	6,548	3,627
$\lambda_4$	3,009	7,165	3,363

Table 2.6: Autocorrelations of the sample paths of  $\beta_1$  using PDA (A.5), RV-PX-PDA (A.6), and RT-PX-PDA (A.9).

Lag	Algorithms		
	PDA (A.5)	RV-PX-PDA (A.6)	RT-PX-PDA (A.9)
1	0.48	0.38	0.49
2	0.27	0.14	0.26
3	0.17	0.05	0.17
4	0.11	0.02	0.12
5	0.09	<0.01	0.09
10	0.05	<0.01	0.05

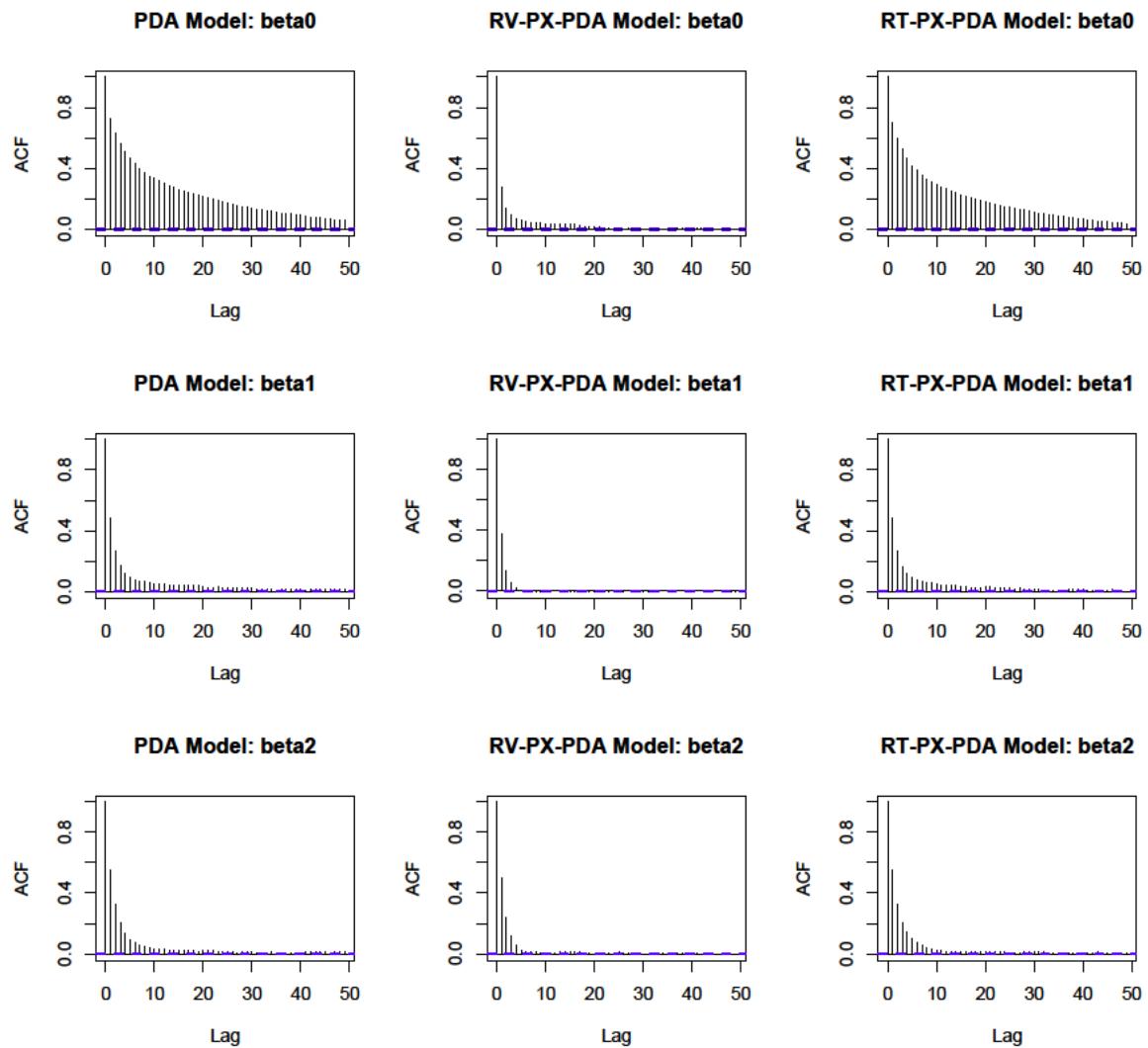


Figure 2.4: Autocorrelation plots for the coefficient vector,  $\beta$ . Again, RV-PX-PDA reports less autocorrelation within the Markov chain than both RT-PX-PDA and PDA for  $\beta_0$ ,  $\beta_1$ , and  $\beta_2$ .

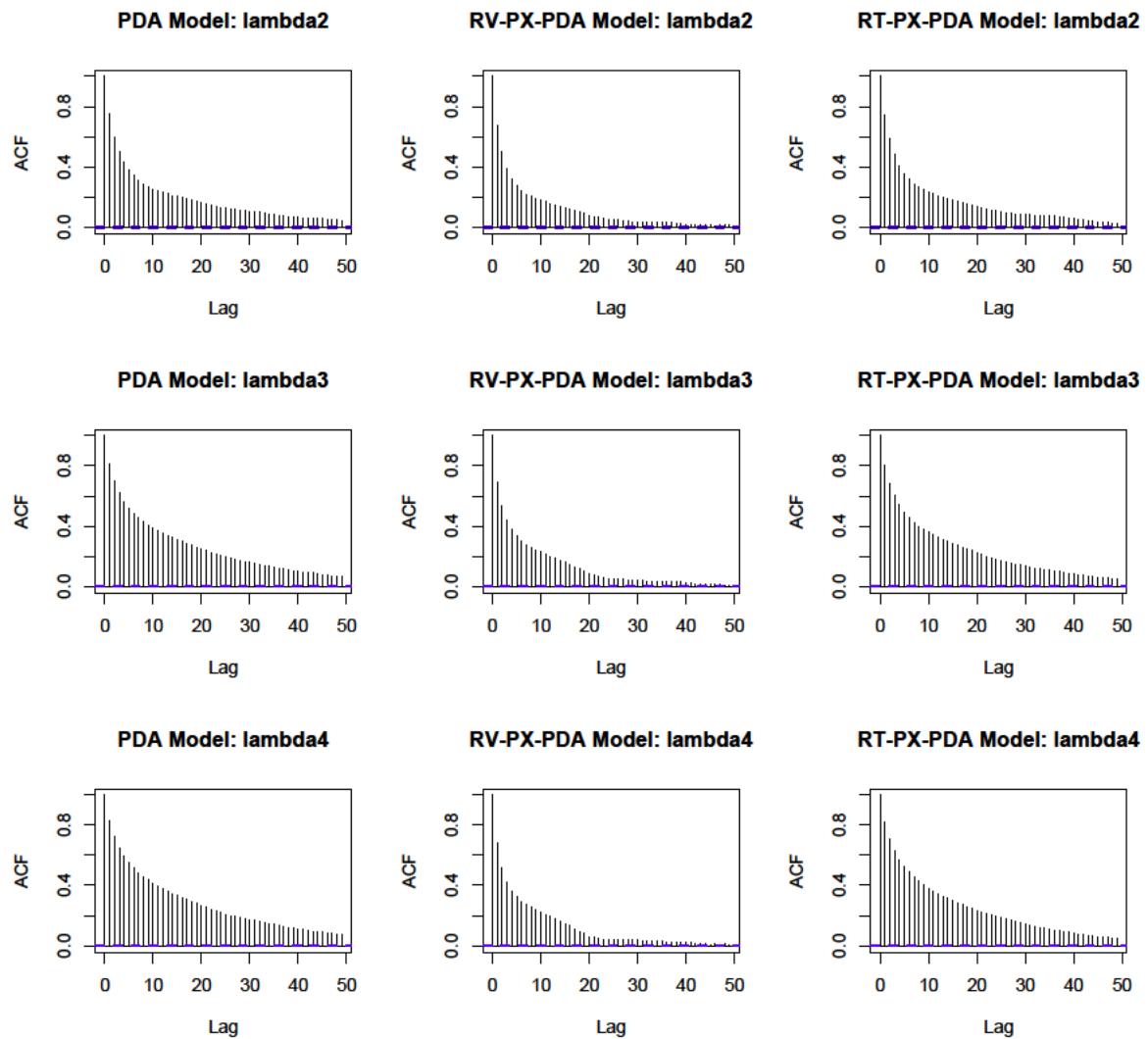


Figure 2.5: Autocorrelation plots for the identified threshold vector,  $\lambda$ . The autocorrelation using RV-PX-PDA is much lower than the other two algorithms for all three identifiable thresholds.

## 2.5 Data augmentation and parameter expansion for spatial probit models

We extend data augmentation and parameter expansion strategies for binary and ordinal response data by allowing for spatial correlation between locations within the domain of interest. Berrett and Calder (2012) developed data augmentation and parameter expansion algorithms for first-stage spatial probit models where the spatial correlation was assumed for the latent response,  $\mathbf{Z}(\mathbf{s})$ . The first-stage model assumes

$$\mathbf{Z}(\mathbf{s}) \sim N(\mathbf{X}(\mathbf{s})\boldsymbol{\beta}, \tau \mathbf{R}) \quad (18)$$

where  $\tau$  is fixed to 1 and  $\mathbf{R}$  is a valid correlation matrix. We develop data augmentation and parameter-expanded data augmentation algorithms for second-stage spatial models for both binary and ordinal response data. The second-stage spatial model (19) fits within the PLMM framework and is arguably more intuitive than the first-stage spatial model (18). Using the same notation as the first-stage spatial model where  $\mathbf{Z}(\mathbf{s})$  is the augmented data, let

$$\mathbf{Z}(\mathbf{s}) \sim N(\mathbf{X}(\mathbf{s})\boldsymbol{\beta} + \mathbf{W}(\mathbf{s}), \sigma^2 \mathbf{I}) \quad (19)$$

where  $\mathbf{W}(\mathbf{s})$  is a spatial random effect such that  $\mathbf{W}(\mathbf{s}) \sim N(0, \Sigma_W)$  and  $\Sigma_W = \tau \mathbf{R}$  and  $\sigma^2$  is fixed to 1. The second-stage model allows for spatial correlation in the mean of the latent variable  $\mathbf{Z}(\mathbf{s})$ . A further comparison between the first and second-stage models appears in Chapter (4). We propose two new parameter-expanded data augmentation algorithms for spatial PLMMs. The algorithms are compared through simulation in terms of Markov chain mixing and parameter estimation.

We begin by assigning prior distributions to the parameters of the spatial covariance of  $\mathbf{W}(\mathbf{s})$ . We assume an isotropic geostatistical spatial model with exponential covariance

function. That is,

$$Cov(W(\mathbf{s}_i), W(\mathbf{s}_j)) = \tau \exp^{-\frac{1}{\phi}d_{ij}}$$

where  $d_{ij}$  is the distance between locations  $\mathbf{s}_i$  and  $\mathbf{s}_j$ ,  $\tau$  is the partial sill parameter and  $\phi$  is the range parameter Let  $R(\phi, d_{ij})$  denote the correlation between locations  $\mathbf{s}_i$  and  $\mathbf{s}_j$  where

$$R(\phi, d_{ij}) = Cor(W(\mathbf{s}_i), W(\mathbf{s}_j)) = \exp^{-\frac{1}{\phi}d_{ij}} \quad (20)$$

We assume a conjugate prior for  $\tau$ , where  $p(\tau) \sim \text{Inv. Gamma}(\alpha_\tau, \beta_\tau)$ . The range parameter,  $\phi$ , requires a Metropolis-Hastings step within MCMC and is assigned a gamma prior with  $\phi \sim \text{Gamma}(\alpha_\phi, \beta_\phi)$ . For convenience, we define the fixed threshold vector for binary data as  $\boldsymbol{\lambda} = (\lambda_0, \lambda_1, \lambda_2) = (-\infty, 0, \infty)$ .

### 2.5.1 Algorithms for spatial PLMM for binary data

Algorithm 10 presents the DA-PLMM algorithm for binary data under the spatial PLMM. This is a modification of the two-step Gibbs sampler (A.1) where the additional steps are for the spatial parameters,  $\tau$  and  $\phi$ . For notational convenience, we drop the dependence on  $\mathbf{s}$  and also write  $\mathbf{R}(\phi, \mathbf{d})$  as  $\mathbf{R}$ .

Parameter-expanded data augmentation can be utilized for spatial PLMMs by modeling the unidentified latent variable as  $\tilde{\mathbf{Z}} \sim N(\mathbf{X}\tilde{\boldsymbol{\beta}} + \tilde{\mathbf{W}}, \sigma^2\mathbf{I})$ . Similar to the previous PX-DA algorithms, we let

$$\mathbf{Z} = \frac{1}{\sigma}\tilde{\mathbf{Z}}, \quad \boldsymbol{\beta} = \frac{1}{\sigma}\tilde{\boldsymbol{\beta}}, \quad \text{and} \quad \mathbf{W} = \frac{1}{\sigma}\tilde{\mathbf{W}}.$$

We do not extend the random threshold parameter-expanded data augmentation algorithms since they proved less optimal than the random variance algorithms in the non-spatial setting. Algorithm 11 gives the PX-DA-PLMM sampling scheme for binary data under the spatial PLMM.

**Algorithm 10.** DA-PLMM algorithm for spatially correlated binary data:

1. Draw  $(\boldsymbol{\beta}^t, \mathbf{W}^t, \tau^t, \phi^t)$  from  $p(\boldsymbol{\beta}, \mathbf{W}, \tau, \phi | \mathbf{y})$ .
  - (a) Draw  $\mathbf{Z}^t$  from  $p(\mathbf{Z} | \mathbf{y}, \boldsymbol{\beta}^{t-1}, \mathbf{W}^{t-1}, \tau^{t-1}, \phi^{t-1}) \sim TN(\mathbf{X}\boldsymbol{\beta}^{t-1} + \mathbf{W}^{t-1}, \mathbf{I}, \boldsymbol{\lambda}_{\mathbf{y}}^t, \boldsymbol{\lambda}_{\mathbf{y}+1}^t)$ .
  - (b) Draw  $\boldsymbol{\beta}^t$  from  $p(\boldsymbol{\beta} | \mathbf{y}, \mathbf{Z}^t, \mathbf{W}^{t-1}, \tau^{t-1}, \phi^{t-1}) \sim N(\mathbf{b}_p, \mathbf{B}_p)$  where  $\mathbf{b}_p = (\mathbf{X}^t \mathbf{X} + \Sigma_{\beta}^{-1})^{-1} \mathbf{X}^t (\mathbf{Z}^t - \mathbf{W}^{t-1})$  and  $\mathbf{B}_p = (\mathbf{X}^t \mathbf{X} + \Sigma_{\beta}^{-1})^{-1}$ .
  - (c) Draw  $\mathbf{W}^t$  from  $p(\mathbf{W} | \mathbf{y}, \mathbf{Z}^t, \boldsymbol{\beta}^t, \tau^{t-1}, \phi^{t-1}) \sim N((\mathbf{I} + \Sigma_W^{-1})^{-1} (\mathbf{Z}^t - \mathbf{X}\boldsymbol{\beta}^t), (\mathbf{I} + \Sigma_W^{-1})^{-1})$ .
  - (d) Draw  $\tau^t$  from  $p(\tau | \mathbf{y}, \mathbf{Z}^t, \boldsymbol{\beta}^t, \mathbf{W}^t, \phi^{t-1}) \sim Inv. Gamma(\alpha_{\tau} + \frac{n}{2}, \beta_{\tau} + (\mathbf{W}^t)'(\mathbf{R}^t)^{-1}\mathbf{W}^t)$  where  $\mathbf{R}$  has exponential correlation function (20).
  - (e) Draw  $\phi^t$  from  $p(\phi | \mathbf{y}, \mathbf{Z}^t, \boldsymbol{\beta}^t, \mathbf{W}^t, \tau^t) \propto p(\phi | \mathbf{W}^t, \tau^t)$ .

We propose a third algorithm for the spatial PLMMs that encompasses further parameter expansion. Recall that the advantage of parameter expansion is that it increases variation between sequential draws of the Markov chain. In the PX-DA algorithm,  $\sigma^2$  is drawn with the latent variable  $\tilde{\mathbf{Z}}$  and again with the remaining model parameters. Our extended parameter-expanded algorithms separates the “remaining model parameters” into two groups: fixed effect parameters and random effect parameters. Then a separate  $\sigma^2$  is sampled with each group and integrated over to obtain the identified fixed effect and random effect parameters, respectively. We refer to the algorithm in the binary setting as PX<sup>2</sup>-DA-PLMM, where PX<sup>2</sup> signifies that  $\sigma^2$  is expanded over twice. As in the previous PX algorithms, we denote the preliminary draws of  $\sigma^2$  as  $(\sigma^2)^*$ , and the iteration  $t$  draw as  $(\sigma^2)^t$ .

To compare DA-PLMM (A.10), PX-DA-PLMM (A.11), and PX<sup>2</sup>-DA-PLMM (A.12), we simulate 300 locations within the unit square from a spatial PLMM. The mean of the latent variable,  $\mathbf{Z}$ , contains an intercept term and one covariate, as well as a spatial random effect,  $\mathbf{W}$ . We aim to simulate data similarly to as was done in the previous data augmentation work of Imai and Van Dyk (2005) and Berrett and Calder (2012). The covariate,  $\mathbf{X}_1$  is drawn

**Algorithm 11.** *PX-DA-PLMM algorithm for spatially correlated binary data:*

1. Draw  $(\boldsymbol{\beta}^t, \mathbf{W}^t, \tau^t, \phi^t)$  from  $p(\boldsymbol{\beta}, \mathbf{W}, \tau, \phi | \mathbf{y})$ .
  - (a) Draw  $(\tilde{\mathbf{Z}}^t, (\sigma^2)^*)$  from  $p(\tilde{\mathbf{Z}}, \sigma^2 | \mathbf{y}, \tilde{\boldsymbol{\beta}}^{t-1}, \tilde{\mathbf{W}}^{t-1}, \tau^{t-1}, \phi^{t-1})$ .
    - i. Draw  $(\sigma^2)^*$  from  $p(\sigma^2 | \mathbf{y}, \tilde{\boldsymbol{\beta}}^{t-1}, \tilde{\mathbf{W}}^{t-1}, \tau^{t-1}, \phi^{t-1}) \sim p(\sigma^2)$ .
    - ii. Draw  $\tilde{\mathbf{Z}}^t$  from  $p(\tilde{\mathbf{Z}} | \mathbf{y}, \tilde{\boldsymbol{\beta}}^{t-1}, \tilde{\mathbf{W}}^{t-1}, \tau^{t-1}, \phi^{t-1}, (\sigma^2)^*)$   
 $\sim TN(\sigma^* \mathbf{X} \boldsymbol{\beta}^{t-1} + \sigma^* \mathbf{W}^{t-1}, (\sigma^*)^2 \mathbf{I}, \sigma^* \boldsymbol{\lambda}_{\mathbf{y}}^t, \sigma^* \boldsymbol{\lambda}_{\mathbf{y}+1}^t)$ .  
Set  $\mathbf{Z}^t = \frac{1}{\sigma^*} \tilde{\mathbf{Z}}^t$ .
  - (b) Draw  $(\tilde{\boldsymbol{\beta}}^t, \tilde{\mathbf{W}}^t, \tau^t, \phi^t, \sigma^2)$  from  $p(\tilde{\boldsymbol{\beta}}, \tilde{\mathbf{W}}, \tau, \phi, \sigma^2 | \mathbf{y}, \tilde{\mathbf{Z}}^t)$ .
    - i. Draw  $(\sigma^2)^t$  from  $p(\sigma^2 | \mathbf{y}, \tilde{\mathbf{Z}}^t)$ . See Appendix A.2.1.
    - ii. Draw  $\tilde{\boldsymbol{\beta}}^t$  from  $p(\tilde{\boldsymbol{\beta}} | \mathbf{y}, \tilde{\mathbf{Z}}^t, \tilde{\mathbf{W}}^{t-1}, \tau^{t-1}, \phi^{t-1}, (\sigma^2)^t) \sim N(\mathbf{b}_p, \mathbf{B}_p)$  where  
 $\mathbf{b}_p = (\mathbf{X}^t \mathbf{X} + \Sigma_{\boldsymbol{\beta}}^{-1})^{-1} \mathbf{X}^t (\mathbf{Z}^t - \mathbf{W}^{t-1})$  and  $\mathbf{B}_p = (\sigma^2)^t (\mathbf{X}^t \mathbf{X} + \Sigma_{\boldsymbol{\beta}}^{-1})^{-1}$ .  
Set  $\boldsymbol{\beta}^t = \frac{1}{\sigma^t} \tilde{\boldsymbol{\beta}}^t$ .
    - iii. Draw  $\tilde{\mathbf{W}}^t$  from  $p(\tilde{\mathbf{W}} | \mathbf{y}, \tilde{\mathbf{Z}}^t, \tilde{\boldsymbol{\beta}}^t, \tau^{t-1}, \phi^{t-1}, (\sigma^2)^t)$   
 $\sim N((\mathbf{I} + \Sigma_W^{-1})^{-1} (\mathbf{Z}^t - \mathbf{X} \boldsymbol{\beta}^t), (\sigma^2)^t (\mathbf{I} + \Sigma_W^{-1})^{-1})$ .  
Set  $\mathbf{W}^t = \frac{1}{\sigma^t} \tilde{\mathbf{W}}^t$ .
    - iv. Draw  $\tau^t$  from  $p(\tau | \mathbf{y}, \tilde{\mathbf{Z}}^t, \tilde{\boldsymbol{\beta}}^t, \tilde{\mathbf{W}}^t, \phi^{t-1}, (\sigma^2)^t)$   
 $\sim Inv. Gamma(\alpha_\tau + \frac{n}{2}, \beta_\tau + (\mathbf{W}^t)' (\mathbf{R}^t)^{-1} \mathbf{W}^t)$   
where  $\mathbf{R}$  has exponential correlation function (20).
    - v. Draw  $\phi^t$  from  $p(\phi | \mathbf{y}, \tilde{\mathbf{Z}}^t, \tilde{\boldsymbol{\beta}}^t, \tilde{\mathbf{W}}^t, \tau^t, (\sigma^2)^t) \propto p(\phi | \mathbf{W}^t, \tau^t)$ .

**Algorithm 12.** *PX<sup>2</sup>-DA-PLMM algorithm for spatially correlated binary data:*

1. Draw  $(\boldsymbol{\beta}^t, \mathbf{W}^t, \tau^t, \phi^t)$  from  $p(\boldsymbol{\beta}, \mathbf{W}, \tau, \phi | \mathbf{y})$ .
  - (a) Draw  $(\tilde{\mathbf{Z}}^t, (\sigma^2)^*)$  from  $p(\tilde{\mathbf{Z}}, \sigma^2 | \mathbf{y}, \tilde{\boldsymbol{\beta}}^{t-1}, \tilde{\mathbf{W}}^{t-1}, \tau^{t-1}, \phi^{t-1})$ .
    - i. Draw  $(\sigma^2)^*$  from  $p(\sigma^2 | \mathbf{y}, \tilde{\boldsymbol{\beta}}^{t-1}, \tilde{\mathbf{W}}^{t-1}, \tau^{t-1}, \phi^{t-1}) \sim p(\sigma^2)$ .
    - ii. Draw  $\tilde{\mathbf{Z}}^t$  from  $p(\tilde{\mathbf{Z}} | \mathbf{y}, \tilde{\boldsymbol{\beta}}^{t-1}, \tilde{\mathbf{W}}^{t-1}, \tau^{t-1}, \phi^{t-1}, (\sigma^2)^*)$   
 $\sim TN(\sigma^* \mathbf{X} \boldsymbol{\beta}^{t-1} + \sigma^* \mathbf{W}^{t-1}, (\sigma^*)^2 \mathbf{I}, \sigma^* \boldsymbol{\lambda}_{\mathbf{y}}^t, \sigma^* \boldsymbol{\lambda}_{\mathbf{y}+1}^t)$ .  
Set  $\mathbf{Z}^t = \frac{1}{\sigma^*} \tilde{\mathbf{Z}}^t$ .
  - (b) Draw  $(\tilde{\boldsymbol{\beta}}^t, (\sigma^2)^*)$  from  $p(\tilde{\boldsymbol{\beta}}, \sigma^2 | \mathbf{y}, \tilde{\mathbf{Z}}^t, \tilde{\mathbf{W}}^{t-1}, \tau^{t-1}, \phi^{t-1})$ .
    - i. Draw  $(\sigma^2)^*$  from  $p(\sigma^2 | \mathbf{y}, \tilde{\mathbf{Z}}^t, \tilde{\mathbf{W}}^{t-1}, \tau^{t-1}, \phi^{t-1})$ . See Appendix A.2.1.
    - ii. Draw  $\tilde{\boldsymbol{\beta}}^t$  from  $p(\tilde{\boldsymbol{\beta}} | \mathbf{y}, \tilde{\mathbf{Z}}^t, \tilde{\mathbf{W}}^{t-1}, \tau^{t-1}, \phi^{t-1}, (\sigma^2)^*) \sim N(\mathbf{b}_p, \mathbf{B}_p)$  where  
 $\mathbf{b}_p = (\mathbf{X}^t \mathbf{X} + \Sigma_{\boldsymbol{\beta}}^{-1})^{-1} \mathbf{X}^t (\mathbf{Z}^t - \mathbf{W}^{t-1})$  and  $\mathbf{B}_p = (\sigma^2)^* (\mathbf{X}^t \mathbf{X} + \Sigma_{\boldsymbol{\beta}}^{-1})^{-1}$ .  
Set  $\boldsymbol{\beta}^t = \frac{1}{\sigma^*} \tilde{\boldsymbol{\beta}}^t$ .
  - (c) Draw  $(\tilde{\mathbf{W}}^t, (\sigma^2)^t)$  from  $p(\tilde{\mathbf{W}}, \sigma^2 | \mathbf{y}, \tilde{\mathbf{Z}}^t, \tilde{\boldsymbol{\beta}}^t, \tau^{t-1}, \phi^{t-1})$ .
    - i. Draw  $(\sigma^2)^t$  from  $p(\sigma^2 | \mathbf{y}, \tilde{\mathbf{Z}}^t, \tilde{\boldsymbol{\beta}}^t, \tau^{t-1}, \phi^{t-1})$ . See Appendix A.2.1.
    - ii. Draw  $\tilde{\mathbf{W}}^t$  from  $p(\tilde{\mathbf{W}} | \mathbf{y}, \tilde{\mathbf{Z}}^t, \tilde{\boldsymbol{\beta}}^t, \tau^{t-1}, \phi^{t-1}, (\sigma^2)^t)$   
 $\sim N((\mathbf{I} + \Sigma_W^{-1})^{-1} (\mathbf{Z}^t - \mathbf{X} \boldsymbol{\beta}^t), (\sigma^2)^t (\mathbf{I} + \Sigma_W^{-1})^{-1})$ .  
Set  $\mathbf{W}^t = \frac{1}{\sigma^t} \tilde{\mathbf{W}}^t$ .
  - (d) Draw  $\tau^t$  from  $p(\tau | \mathbf{y}, \tilde{\mathbf{Z}}^t, \tilde{\boldsymbol{\beta}}^t, \tilde{\mathbf{W}}^t, \phi^{t-1}, (\sigma^2)^t)$   
 $\sim Inv. Gamma(\alpha_\tau + \frac{n}{2}, \beta_\tau + (\mathbf{W}^t)' (\mathbf{R}^t)^{-1} \mathbf{W}^t)$   
where  $\mathbf{R}$  has exponential correlation function (20).
  - (e) Draw  $\phi^t$  from  $p(\phi | \mathbf{y}, \tilde{\mathbf{Z}}^t, \tilde{\boldsymbol{\beta}}^t, \tilde{\mathbf{W}}^t, \tau^t, (\sigma^2)^t) \propto p(\phi | \mathbf{W}^t, \tau^t)$ .

from a Uniform(-0.5, 0.5) and the coefficient vector is  $(\beta_0, \beta_1) = (-0.5, -\sqrt{2})$ . The partial sill and range parameter of the spatial covariance function are set to  $\tau = 1.0$  and  $\phi = 0.2$ . We assign prior distributions to the spatial parameters such that  $\tau \sim \text{Inv. Gamma}(2, 3)$  and  $\phi \sim \text{Gamma}(1, 2)$ . The prior distribution of  $\tau$  has mean 3 and infinite variance (similar to that assigned by Gelfand et al. (2000)). The prior distribution of  $\phi$  has mean 0.5 and is assigned such that the effective range is reasonable for the unit square. The parameter expansion parameter,  $\sigma^2$ , is also assigned a conjugate prior, with  $\sigma^2 \sim \text{Inv. Gamma}(4, 3)$ . This prior distribution required some tuning because a large variance hindered convergence of the partial sill parameter,  $\tau$ , in the PX<sup>2</sup>-DA algorithm.

Each algorithm was run for 100,000 iterations, the first 10,000 disregarded as burn-in. Table 2.7 gives the posterior medians and 95% credible intervals for the identifiable parameters from each of the algorithms as well as the true values from the simulation. The true values are captured by the 95% credible intervals using all three algorithms, however  $\tau$  is overestimated and  $\phi$  is underestimated. The estimability of spatial parameters from binary response data is discussed further in Chapter 3. We compare the convergence of the algorithms by looking at plots of parameter autocorrelation as well as computing the effective sample size for the identifiable parameters. Figures 2.6 and 2.7 display the autocorrelation for each model parameter across the three algorithms. While the ESS values for all cases are not large, PX<sup>2</sup>-DA-PLMM greatly outperforms the other two algorithms in terms of independent samples for the fixed effect coefficients,  $\beta_0$  and  $\beta_1$ , as well as the partial sill parameter of the spatial covariance function,  $\tau$  (Table 2.8). This indicates more variation within the chain for these three parameters as well as faster convergence for the PX<sup>2</sup>-DA-PLMM algorithm. There is not much difference, however, between the algorithms in terms of the spatial range parameter,  $\phi$ . The small overall ESS for all algorithms reflects the challenges of estimating spatial fields based on binary response data.

Table 2.7: Posterior medians and 95% credible intervals of identifiable parameters for each of the spatial PLMM algorithms for binary data.

Parameter	True Value	Algorithms		
		DA-PLMM (A.10)	PX-DA-PLMM (A.11)	PX <sup>2</sup> -DA-PLMM (A.12)
$\beta_0$	-0.5	-0.86 (-1.65, 0.48)	-0.84 (-3.49, -0.06)	-0.60 (-1.24, 0.65)
$\beta_1$	$-\sqrt{2}$	-1.56 (-3.61, -0.71)	-1.66 (-8.06, -0.67)	-1.22 (-2.24, -0.37)
$\tau$	1	1.65 (0.58, 10.04)	1.86 (0.56, 60.3)	1.40 (0.51, 10.08)
$\phi$	0.2	0.04 (0.01, 0.94)	0.03 (0.14, 0.76)	0.05 (0.02, 1.40)

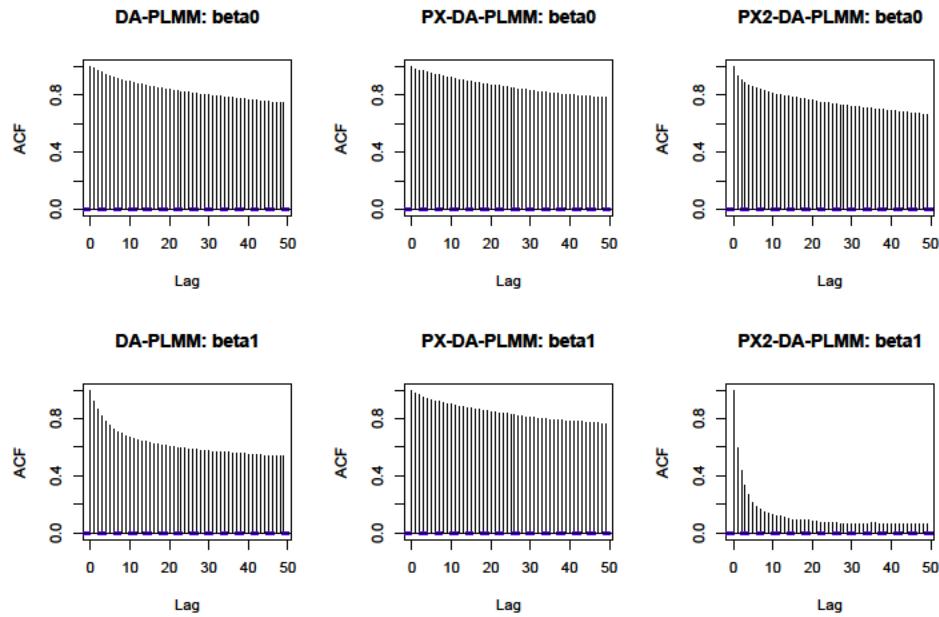


Figure 2.6: Autocorrelation plots for the coefficient vector,  $\beta$ , in the spatial PLMM for binary data. PX<sup>2</sup>-DA algorithm reports less autocorrelation within the Markov chain than the other two PLMM algorithms, PDA and PX-PDA, especially for  $\beta_1$ .

Table 2.8: Effective sample size estimates for the spatial PLMM algorithms for binary data for each of the identifiable parameters for 90,000 MCMC iterations.

Parameter	Algorithms		
	DA-PLMM (A.10)	PX-DA-PLMM (A.11)	PX <sup>2</sup> -DA-PLMM (A.12)
$\beta_0$	202	173	261
$\beta_1$	380	168	5,805
$\tau$	123	203	447
$\phi$	58	94	94

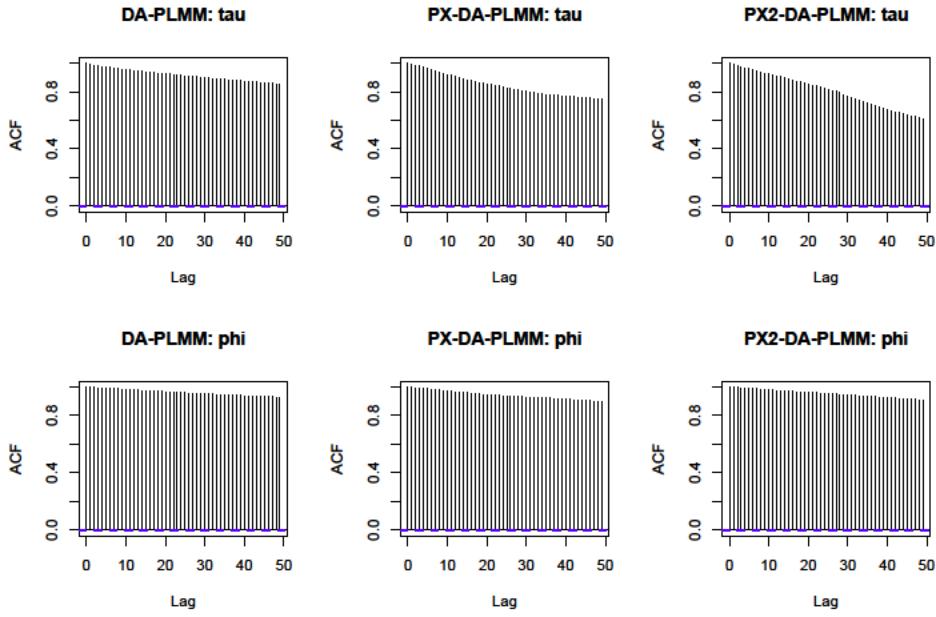


Figure 2.7: Autocorrelation plots for spatial parameters,  $\tau$  and  $\phi$ , in the spatial PLMM for binary data.

### 2.5.2 Algorithms for spatial PLMM for ordinal data

We extend each of the algorithms developed for spatially correlated binary data to allow for ordinal response data with 3 or more categories. Let PDA-PLMM (A.13) be the partial data-augmented algorithm, and PX-PDA-PLMM (A.14) and PX<sup>2</sup>-PDA-PLMM (A.15) be the two parameter-expanded partial data-augmented algorithms for the spatial PLMM for ordinal response data.

We use the same  $n = 300$  locations simulated in the binary case for fitting the ordinal response model using each of the algorithms. The observed response variable,  $y$ , is ordinal with  $K = 5$  categories. The mean of the latent variable,  $Z$ , once again contains an intercept term and one covariate, as well as the spatial random effect,  $W$ . The covariate data is the same as in the binary case, although the intercept coefficient is now set to  $\beta_0 = 0.5$  as opposed to  $\beta_0 = -0.5$  in the binary case. This was to ensure there was a sufficient number of response values in each of the five ordinal categories. Each algorithm was run for 100,000 iterations, the first 10,000 disregarded as burn-in. The true values of the parameters as well

**Algorithm 13.** *PDA-PLMM algorithm for spatially correlated ordinal data:*

1. Draw  $\boldsymbol{\lambda}^t$  from  $p(\boldsymbol{\lambda}|\mathbf{y}, \boldsymbol{\beta}^{t-1}\mathbf{W}^t, \tau^t, \phi^t)$ .
2. Draw  $(\boldsymbol{\beta}^t, \mathbf{W}^t, \tau^t, \phi^t)$  from  $p(\boldsymbol{\beta}, \mathbf{W}, \tau, \phi|\mathbf{y}, \boldsymbol{\lambda}^t)$ .
  - (a) Draw  $\mathbf{Z}^t$  from  $p(\mathbf{Z}|\mathbf{y}, \boldsymbol{\beta}^{t-1}, \mathbf{W}^{t-1}, \tau^{t-1}, \phi^{t-1}, \boldsymbol{\lambda}^t)$   
 $\sim TN(\mathbf{X}\boldsymbol{\beta}^{t-1} + \mathbf{W}^{t-1}, \mathbf{I}, \boldsymbol{\lambda}_{\mathbf{y}-1}^t, \boldsymbol{\lambda}_{\mathbf{y}}^t)$ .
  - (b) Draw  $\boldsymbol{\beta}^t$  from  $p(\boldsymbol{\beta}|\mathbf{y}, \mathbf{Z}^t, \mathbf{W}^{t-1}, \tau^{t-1}, \phi^{t-1}, \boldsymbol{\lambda}^t)$   
 $\sim N(\mathbf{b}_p, \mathbf{B}_p)$  where  
 $\mathbf{b}_p = (\mathbf{X}^t \mathbf{X} + \Sigma_{\beta}^{-1})^{-1} \mathbf{X}^t (\mathbf{Z}^t - \mathbf{W}^{t-1})$  and  $\mathbf{B}_p = (\mathbf{X}^t \mathbf{X} + \Sigma_{\beta}^{-1})^{-1}$ .
  - (c) Draw  $\mathbf{W}^t$  from  $p(\mathbf{W}|\mathbf{y}, \mathbf{Z}^t, \boldsymbol{\beta}^t, \tau^{t-1}, \phi^{t-1}, \boldsymbol{\lambda}^t)$   
 $\sim N((\mathbf{I} + \Sigma_W^{-1})^{-1} (\mathbf{Z}^t - \mathbf{X}\boldsymbol{\beta}^t), (\mathbf{I} + \Sigma_W^{-1})^{-1})$ .
  - (d) Draw  $\tau^t$  from  $p(\tau|\mathbf{y}, \mathbf{Z}^t, \boldsymbol{\beta}^t, \mathbf{W}^t, \phi^{t-1}, \boldsymbol{\lambda}^t)$   
 $\sim Inv. Gamma(\alpha_{\tau} + \frac{n}{2}, \beta_{\tau} + (\mathbf{W}^t)'(\mathbf{R}^t)^{-1}\mathbf{W}^t)$   
where  $\mathbf{R}$  has exponential correlation function (20).
  - (e) Draw  $\phi^t$  from  $p(\phi|\mathbf{y}, \mathbf{Z}^t, \boldsymbol{\beta}^t, \mathbf{W}^t, \tau^t, \boldsymbol{\lambda}^t) \propto p(\phi|\mathbf{W}^t, \tau^t)$ .

as the posterior medians and 95% credible intervals for each of the three algorithms are given in Table 2.9. The results are very similar across the algorithms with each credible interval capturing the true value.

Table 2.10 gives the effective sample size estimates for each parameter for each of the algorithms. The coefficients,  $\beta_0$  and  $\beta_1$ , and the threshold parameters,  $\lambda_2$ ,  $\lambda_3$ , and  $\lambda_4$ , all have much higher effective sample sizes using PX<sup>2</sup>-PDA-PLMM than in PX-PDA-PLMM and PDA-PLMM. Both parameter-expanded algorithms outperform the PDA-PLMM algorithm for all model parameters whereas PX<sup>2</sup>-PDA-PLMM outperforms PX-PDA-PLMM for all parameters except  $\tau$ , where the effective sample sizes are close at 560 and 599, respectively. The autocorrelation plots are shown in Figures 2.8, 2.9, and 2.10.

**Algorithm 14.** *PX-PDA-PLMM algorithm for spatially correlated ordinal data:*

1. Draw  $\boldsymbol{\lambda}^t$  from  $p(\boldsymbol{\lambda}|\mathbf{y}, \boldsymbol{\beta}^{t-1}\mathbf{W}^t, \tau^t, \phi^t)$ .
2. Draw  $(\boldsymbol{\beta}^t, \mathbf{W}^t, \tau^t, \phi^t)$  from  $p(\boldsymbol{\beta}, \mathbf{W}, \tau, \phi|\mathbf{y}, \boldsymbol{\lambda}^t)$ .
  - (a) Draw  $(\tilde{\mathbf{Z}}^t, (\sigma^2)^*)$  from  $p(\tilde{\mathbf{Z}}, \sigma^2|\mathbf{y}, \tilde{\boldsymbol{\beta}}^{t-1}, \tilde{\mathbf{W}}^{t-1}, \tau^{t-1}, \phi^{t-1}, \boldsymbol{\lambda}^t)$ .
    - i. Draw  $(\sigma^2)^*$  from  $p(\sigma^2|\mathbf{y}, \tilde{\boldsymbol{\beta}}^{t-1}, \tilde{\mathbf{W}}^{t-1}, \tau^{t-1}, \phi^{t-1}, \boldsymbol{\lambda}^t) \sim p(\sigma^2)$ .
    - ii. Draw  $\tilde{\mathbf{Z}}^t$  from  $p(\tilde{\mathbf{Z}}|\mathbf{y}, \tilde{\boldsymbol{\beta}}^{t-1}, \tilde{\mathbf{W}}^{t-1}, \tau^{t-1}, \phi^{t-1}, (\sigma^2)^*, \boldsymbol{\lambda}^t) \sim TN(\sigma^* \mathbf{X} \boldsymbol{\beta}^{t-1} + \sigma^* \mathbf{W}^{t-1}, (\sigma^*)^2 \mathbf{I}, \sigma^* \boldsymbol{\lambda}_{\mathbf{y}-1}^t, \sigma^* \boldsymbol{\lambda}_{\mathbf{y}}^t)$ .  
Set  $\mathbf{Z}^t = \frac{1}{\sigma^*} \tilde{\mathbf{Z}}^t$ .
  - (b) Draw  $(\tilde{\boldsymbol{\beta}}^t, \tilde{\mathbf{W}}^t, \tau^t, \phi^t, \sigma^2)$  from  $p(\tilde{\boldsymbol{\beta}}, \tilde{\mathbf{W}}, \tau, \phi, \sigma^2|\mathbf{y}, \tilde{\mathbf{Z}}^t, \boldsymbol{\lambda}^t)$ .
    - i. Draw  $(\sigma^2)^t$  from  $p(\sigma^2|\mathbf{y}, \tilde{\mathbf{Z}}^t, \boldsymbol{\lambda}^t)$ . See Appendix A.2.1.
    - ii. Draw  $\tilde{\boldsymbol{\beta}}^t$  from  $p(\tilde{\boldsymbol{\beta}}|\mathbf{y}, \tilde{\mathbf{Z}}^t, \tilde{\mathbf{W}}^{t-1}, \tau^{t-1}, \phi^{t-1}, (\sigma^2)^t, \boldsymbol{\lambda}^t) \sim N(\mathbf{b}_p, \mathbf{B}_p)$  where  
 $\mathbf{b}_p = (\mathbf{X}^t \mathbf{X} + \Sigma_{\boldsymbol{\beta}}^{-1})^{-1} \mathbf{X}^t (\mathbf{Z}^t - \mathbf{W}^{t-1})$  and  $\mathbf{B}_p = (\sigma^2)^t (\mathbf{X}^t \mathbf{X} + \Sigma_{\boldsymbol{\beta}}^{-1})^{-1}$ .  
Set  $\boldsymbol{\beta}^t = \frac{1}{\sigma^t} \tilde{\boldsymbol{\beta}}^t$ .
    - iii. Draw  $\tilde{\mathbf{W}}^t$  from  $p(\tilde{\mathbf{W}}|\mathbf{y}, \tilde{\mathbf{Z}}^t, \tilde{\boldsymbol{\beta}}^t, \tau^{t-1}, \phi^{t-1}, (\sigma^2)^t, \boldsymbol{\lambda}^t) \sim N((\mathbf{I} + \Sigma_W^{-1})^{-1} (\mathbf{Z}^t - \mathbf{X} \boldsymbol{\beta}^t), (\sigma^2)^t (\mathbf{I} + \Sigma_W^{-1})^{-1})$ .  
Set  $\mathbf{W}^t = \frac{1}{\sigma^t} \tilde{\mathbf{W}}^t$ .
    - iv. Draw  $\tau^t$  from  $p(\tau|\mathbf{y}, \tilde{\mathbf{Z}}^t, \tilde{\boldsymbol{\beta}}^t, \tilde{\mathbf{W}}^t, \phi^{t-1}, (\sigma^2)^t, \boldsymbol{\lambda}^t) \sim Inv. Gamma(\alpha_\tau + \frac{n}{2}, \beta_\tau + (\mathbf{W}^t)'(\mathbf{R}^t)^{-1} \mathbf{W}^t)$  where  $\mathbf{R}$  has exponential correlation function (20).
    - v. Draw  $\phi^t$  from  $p(\phi|\mathbf{y}, \tilde{\mathbf{Z}}^t, \tilde{\boldsymbol{\beta}}^t, \tilde{\mathbf{W}}^t, \tau^t, (\sigma^2)^t, \boldsymbol{\lambda}^t) \propto p(\phi|\mathbf{W}^t, \tau^t)$ .

**Algorithm 15.** *PX<sup>2</sup>-PDA-PLMM algorithm for spatially correlated ordinal data:*

1. Draw  $\boldsymbol{\lambda}^t$  from  $p(\boldsymbol{\lambda}|\mathbf{y}, \boldsymbol{\beta}^{t-1}\mathbf{W}^t, \tau^t, \phi^t)$ .
2. Draw  $(\boldsymbol{\beta}^t, \mathbf{W}^t, \tau^t, \phi^t)$  from  $p(\boldsymbol{\beta}, \mathbf{W}, \tau, \phi|\mathbf{y}, \boldsymbol{\lambda}^t)$ .
  - (a) Draw  $(\tilde{\mathbf{Z}}^t, (\sigma^2)^*)$  from  $p(\tilde{\mathbf{Z}}, \sigma^2|\mathbf{y}, \tilde{\boldsymbol{\beta}}^{t-1}, \tilde{\mathbf{W}}^{t-1}, \tau^{t-1}, \phi^{t-1}, \boldsymbol{\lambda}^t)$ .
    - i. Draw  $(\sigma^2)^*$  from  $p(\sigma^2|\mathbf{y}, \tilde{\boldsymbol{\beta}}^{t-1}, \tilde{\mathbf{W}}^{t-1}, \tau^{t-1}, \phi^{t-1}, \boldsymbol{\lambda}^t) \sim p(\sigma^2)$ .
    - ii. Draw  $\tilde{\mathbf{Z}}^t$  from  $p(\tilde{\mathbf{Z}}|\mathbf{y}, \tilde{\boldsymbol{\beta}}^{t-1}, \tilde{\mathbf{W}}^{t-1}, \tau^{t-1}, \phi^{t-1}, (\sigma^2)^*, \boldsymbol{\lambda}^t)$   
 $\sim TN(\sigma^* \mathbf{X} \boldsymbol{\beta}^{t-1} + \sigma^* \mathbf{W}^{t-1}, (\sigma^*)^2 \mathbf{I}, \sigma^* \boldsymbol{\lambda}_{\mathbf{y}-1}^t, \sigma^* \boldsymbol{\lambda}_{\mathbf{y}}^t)$ . Set  $\mathbf{Z}^t = \frac{1}{\sigma^*} \tilde{\mathbf{Z}}^t$ .
  - (b) Draw  $(\tilde{\boldsymbol{\beta}}^t, (\sigma^2)^*)$  from  $p(\tilde{\boldsymbol{\beta}}, \sigma^2|\mathbf{y}, \tilde{\mathbf{Z}}^t, \tilde{\mathbf{W}}^{t-1}, \tau^{t-1}, \phi^{t-1}, \boldsymbol{\lambda}^t)$ .
    - i. Draw  $(\sigma^2)^*$  from  $p(\sigma^2|\mathbf{y}, \tilde{\mathbf{Z}}^t, \tilde{\mathbf{W}}^{t-1}, \tau^{t-1}, \phi^{t-1}, \boldsymbol{\lambda})$ .  
 See Appendix A.2.1.
    - ii. Draw  $\tilde{\boldsymbol{\beta}}^t$  from  $p(\tilde{\boldsymbol{\beta}}|\mathbf{y}, \tilde{\mathbf{Z}}^t, \tilde{\mathbf{W}}^{t-1}, \tau^{t-1}, \phi^{t-1}, (\sigma^2)^*, \boldsymbol{\lambda}^t) \sim N(\mathbf{b}_p, \mathbf{B}_p)$   
 where  
 $\mathbf{b}_p = (\mathbf{X}^t \mathbf{X} + \Sigma_{\beta}^{-1})^{-1} \mathbf{X}^t (\mathbf{Z}^t - \mathbf{W}^{t-1})$  and  $\mathbf{B}_p = (\sigma^2)^* (\mathbf{X}^t \mathbf{X} + \Sigma_{\beta}^{-1})^{-1}$ .  
 Set  $\boldsymbol{\beta}^t = \frac{1}{\sigma^*} \tilde{\boldsymbol{\beta}}^t$ .
  - (c) Draw  $(\tilde{\mathbf{W}}^t, (\sigma^2)^t)$  from  $p(\tilde{\mathbf{W}}, \sigma^2|\mathbf{y}, \tilde{\mathbf{Z}}^t, \tilde{\boldsymbol{\beta}}^t, \tau^{t-1}, \phi^{t-1}, \boldsymbol{\lambda}^t)$ .
    - i. Draw  $(\sigma^2)^t$  from  $p(\sigma^2|\mathbf{y}, \tilde{\mathbf{Z}}^t, \tilde{\boldsymbol{\beta}}^{t-1}, \tau^{t-1}, \phi^{t-1}, \boldsymbol{\lambda}^t)$ .  
 See Appendix A.2.1.
    - ii. Draw  $\tilde{\mathbf{W}}^t$  from  $p(\tilde{\mathbf{W}}|\mathbf{y}, \tilde{\mathbf{Z}}^t, \tilde{\boldsymbol{\beta}}^t, \tau^{t-1}, \phi^{t-1}, (\sigma^2)^t, \boldsymbol{\lambda}^t)$   
 $\sim N((\mathbf{I} + \Sigma_W^{-1})^{-1} (\mathbf{Z}^t - \mathbf{X} \boldsymbol{\beta}^t), (\sigma^2)^t (\mathbf{I} + \Sigma_W^{-1})^{-1})$ .  
 Set  $\mathbf{W}^t = \frac{1}{\sigma^t} \tilde{\mathbf{W}}^t$ .
  - (d) Draw  $\tau^t$  from  $p(\tau|\mathbf{y}, \tilde{\mathbf{Z}}^t, \tilde{\boldsymbol{\beta}}^t, \tilde{\mathbf{W}}^t, \phi^{t-1}, (\sigma^2)^t, \boldsymbol{\lambda}^t)$   
 $\sim Inv. Gamma(\alpha_{\tau} + \frac{n}{2}, \beta_{\tau} + (\mathbf{W}^t)'(\mathbf{R}^t)^{-1} \mathbf{W})$   
 where  $\mathbf{R}$  has exponential correlation function (20).
  - (e) Draw  $\phi^t$  from  $p(\phi|\mathbf{y}, \tilde{\mathbf{Z}}^t, \tilde{\boldsymbol{\beta}}^t, \tilde{\mathbf{W}}^t, \tau^t, (\sigma^2)^t, \boldsymbol{\lambda}^t) \propto p(\phi|\mathbf{W}^t, \tau^t)$ .

Table 2.9: Posterior medians and 95% credible intervals of spatial PLMM parameters for ordinal data.

Parameter	True Value	Algorithms		
		PDA-PLMM (A.13)	PX-PDA-PLMM (A.14)	PX <sup>2</sup> -PDA-PLMM (A.15)
$\beta_0$	0.5	0.49 (-0.08, 0.98)	0.49 (-0.23, 1.12)	0.44 (-0.10, 0.99)
$\beta_1$	$-\sqrt{2}$	-1.77 (-2.67, -1.14)	-1.73 (-2.87, -1.06)	-1.67 (-2.63, -0.72)
$\tau$	1	1.25 (0.60, 3.25)	1.25 (0.57, 4.07)	1.22 (0.58, 3.69)
$\phi$	0.2	0.06 (0.03, 0.46)	0.07 (0.03, 0.48)	0.07 (0.03, 0.43)
$\lambda_2$	0.6	0.72 (0.54, 1.72)	0.71 (0.53, 1.08)	0.69 (0.50, 0.97)
$\lambda_3$	1.2	1.43 (1.15, 2.12)	1.43 (1.13, 2.14)	1.40 (1.09, 1.92)
$\lambda_4$	1.8	2.09 (1.72, 2.91)	2.09 (1.68, 3.12)	2.06 (1.65, 2.81)

Table 2.10: Effective sample size estimates for the parameters of the spatial PLMM for ordinal data for 90,000 MCMC iterations.

Parameter	Algorithms		
	PDA-PLMM (A.13)	PX-PDA-PLMM (A.14)	PX <sup>2</sup> -PDA-PLMM (A.15)
$\beta_0$	834	1010	1096
$\beta_1$	828	927	3677
$\tau$	390	599	560
$\phi$	107	183	196
$\lambda_2$	401	557	916
$\lambda_3$	288	405	594
$\lambda_4$	265	379	513

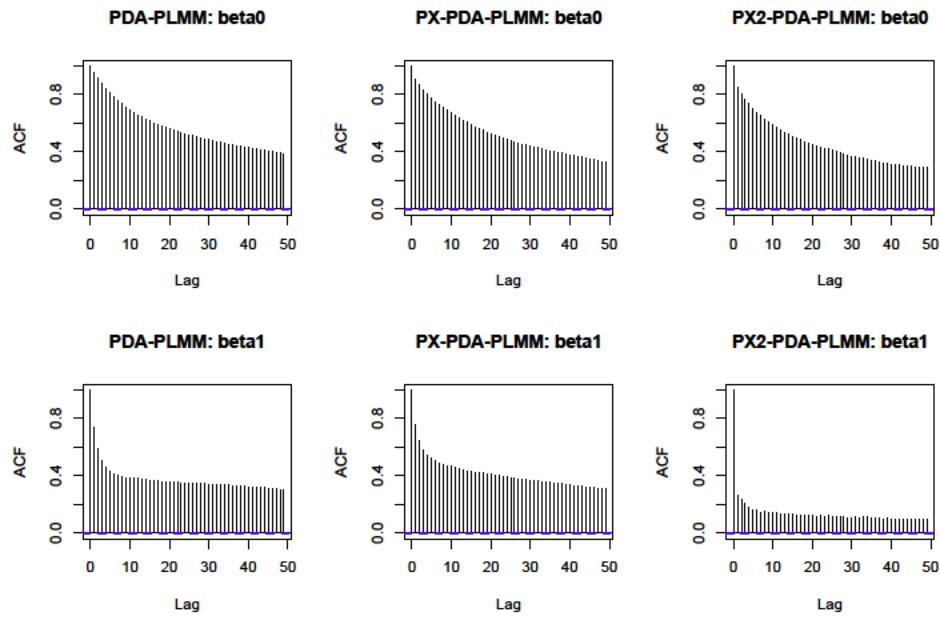


Figure 2.8: Autocorrelation plots for the coefficient vector,  $\beta$ , in the spatial PLMM for ordinal data. PX<sup>2</sup>-PDA algorithm reports less autocorrelation within the Markov chain than the other two PLMM algorithms, PDA and PX-PDA. This is especially true for coefficient  $\beta_1$ .

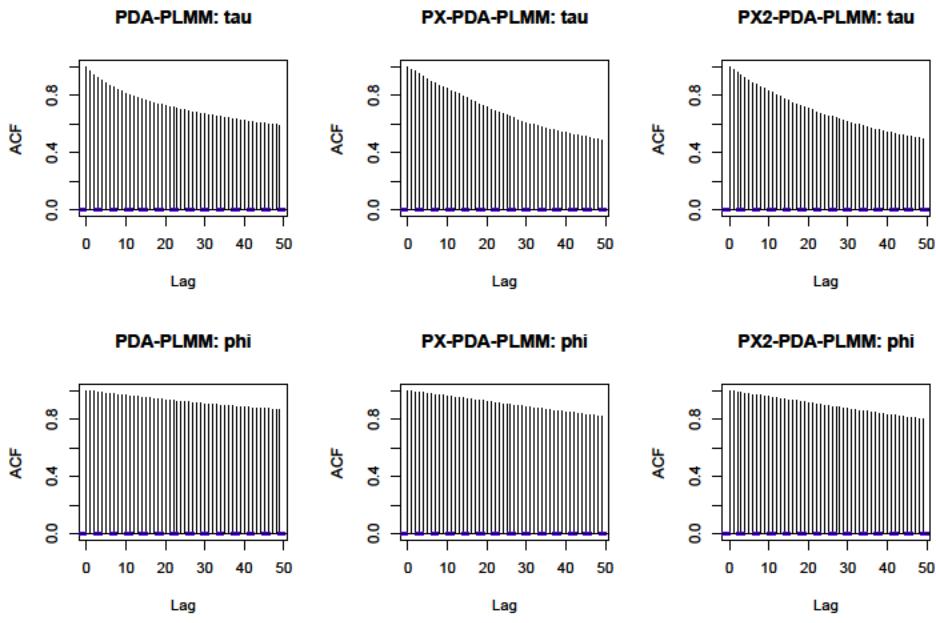


Figure 2.9: Autocorrelation plots for spatial parameters,  $\tau$  and  $\phi$ , in the spatial PLMM for ordinal data.

## 2.6 Discussion

When utilizing data augmentation and parameter-expanded data augmentation strategies, it is important that model parameters be identifiable. It is easy to see that fitting the parameter-expanded probit regression model naively without integrating over the variance parameter would result in an unidentifiable model, and thus, a divergent Markov chain. Unfortunately, this issue isn't completely resolved even when the model is likelihood identifiable. For example, the low effective sample size estimates from the spatial PLMMs indicate issues with near nonidentifiability of the model (Table 2.10). When fitting spatial models in practice, it is important to consider the trade-off between near nonidentifiability and under-dispersion. In the extreme case when  $\phi = 0$  such that there is no spatial correlation,  $\Sigma_W = \tau \mathbf{R} = \tau \mathbf{I}$  in model (19) and the model is nonidentifiable because the marginal variance of the latent variable,  $Z$ , is  $(1 + \tau)\mathbf{I}$ . When  $\phi$  is close to 0, the model is weakly identifiable

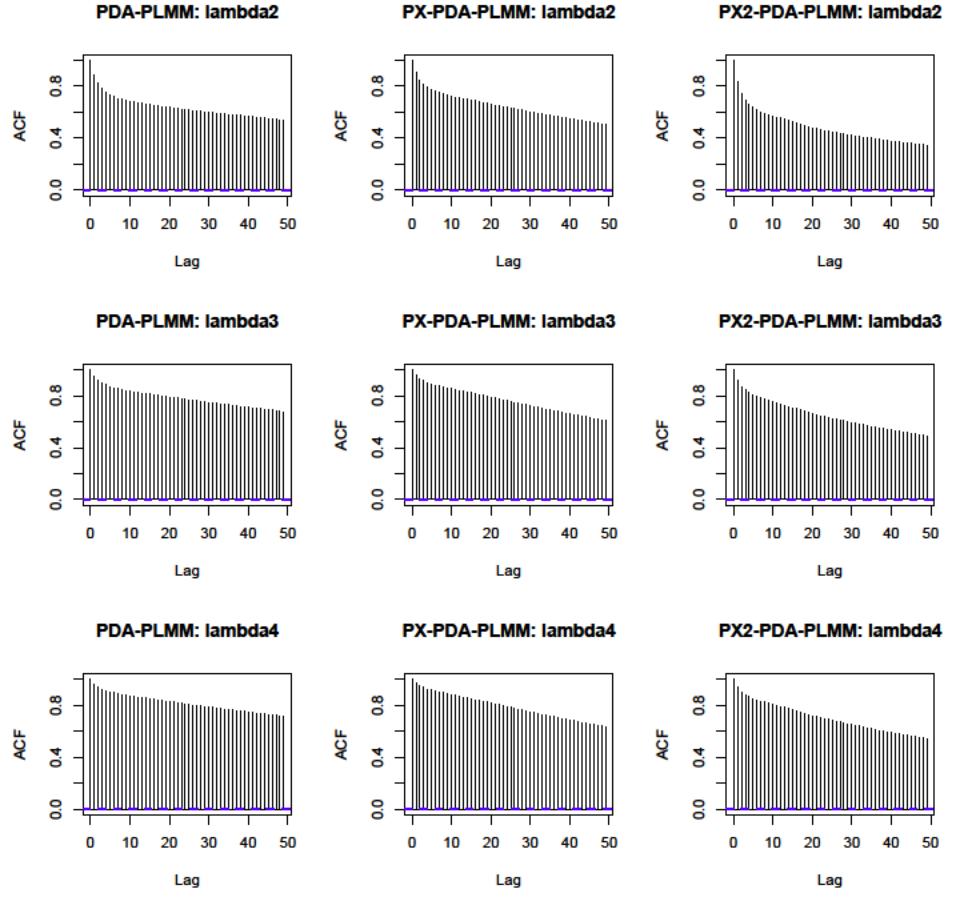


Figure 2.10: Autocorrelation plots for the coefficient vector,  $\lambda$ , in the spatial PLMM for ordinal data.

or near-nonidentifiable. When  $\phi >> 0$ , the model is identifiable but the spatial field can be under-dispersed. In the Bayesian framework, assigning informative priors to the parameters of the spatial field in the model can help mediate between the two extremes. This issue will be discussed further in Chapter 3.

Data augmentation strategies and, more recently, parameter-expanded data augmentation strategies have been shown to increase the ease of sampling from posterior distributions while also increasing the rate of convergence and mixing when using MCMC. In situations where posterior distributions are not known in closed form, as is the case in modeling ordinal data, or as datasets get larger, which can occur with richer spatial data, convergence rates become extremely important. For this reason alone, alternative methods for drawing inference are being introduced. One method in particular that has been shown to be fast and flexible for latent Gaussian models is the integrated nested Laplace approximation approach (INLA), which can also extend to spatial and spatiotemporal models (Rue et al., 2009). For example, INLA can be used to fit spatial GLMMs in seconds where it can take hours to run MCMC. Unfortunately, INLA also has its limitations as it is not available for ordinal response data, including PLMMs, or mixture models, for example, and it is difficult to adapt the INLA method to new model structures. Therefore, MCMC is still an obvious choice for Bayesian model inference and increasing the efficiency of MCMC is relevant and necessary.

## CHAPTER 3

### IDENTIFIABILITY

Informally, nonidentifiability is the inability of the data to distinguish between parameter values. In simple models, it can be obvious which parameters are well-identified and which are not, but in complex models, the distinction can be subtle. Advances in technology and efficient computational algorithms have led to an even greater increase in complexity in statistical models causing identifiability issues across many areas of statistical modeling. Over the past half-century, there have been distinct schools of thought regarding identifiability and its affects on frequentist and Bayesian inference. Lindley (1972) stated “In passing, it might be noted that unidentifiability causes no real difficulty in the Bayesian approach.” Kadane (1974) argued that “identification is a property of the likelihood function, and thus is the same for both frequentists and Bayesian approaches.”

In a sense, both of these statements are correct. Markov chain Monte Carlo (MCMC) algorithms have been shown by many authors (e.g. Besag et al., 1995) to be justified for fitting models that are nonidentifiable given that the samples are used only to summarize the components of the identifiable, proper posterior. This theory is grounded in the fact that in the Bayesian framework, inference is drawn based on both the likelihood function and the prior distributions of the unknown parameters as opposed to only the likelihood in the classical framework. Therefore, prior distributions can be used to assist in parameter estimation by giving additional information to the model that is missing in the data. Furthermore, proper prior distributions lead to proper posterior distributions meaning every parameter can be well-estimated. Conversely, many others have also shown nonidentifiability to be a major roadblock for Bayesians using MCMC methods (e.g. Gelfand and Smith, 1990; Gelfand and Sahu, 1999; Eberly and Carlin, 2000). The issue is magnified when Bayesians employ vague

and/or improper prior distributions on nonidentifiable parameters. A practical issue is that nonidentifiability can cause high correlation between parameters in the posterior distribution causing poor mixing in the Markov chain and slow convergence. Improperity in the posterior that results from a nonidentifiable parameter having an improper prior will present itself as convergence failure of MCMC (Eberly and Carlin, 2000). Second, nonidentifiability can result in the inability for even large sample sizes to overcome prior distributions. Philosophically, when one assigns a noninformative prior distribution to a nonidentifiable parameter, there is not enough information in the data or the prior distribution to appropriately draw inference from the model (Swartz et al., 2004). These issues only add to the fact that sometimes the nonidentifiable parameters are values of interest so only using the samples to summarize the identifiable, proper posterior is of little use.

In this chapter, we begin by discussing identifiability in a general context as it applies to frequentists and Bayesians alike. This includes defining and comparing the terms *likelihood identifiability* and *Bayesian identifiability*. In Section 3.2 we introduce nonidentifiability in latent variable models, with focus on linear models and linear mixed models. Section 3.2.1 gives general methods for investigating parameter identifiability. In Section 3.2.2 we apply a mapping approach for checking identifiability in the specific case of latent variable Gaussian process models for ordinal response data. We investigate identifiability in latent variable models for spatially correlated binary and ordinal response data in Section 3.3. In Section 3.4 we propose a new parameterization for fitting an exponential covariogram using MCMC and evaluate it against the current parameterization via simulation. The chapter concludes with a discussion in Section 3.5.

### 3.1 Identifiability, as it applies to frequentists and Bayesians

We begin by stating a general definition of identifiability first given in Basu (1983).

**Definition 1.** Let  $Y$  be an observable random variable with distribution function  $p_{\theta}$  indexed by a parameter  $\theta$  where  $\theta \in \Omega$ . Here,  $\theta$  could be scalar or vector-valued. We say that  $\theta$  is nonidentifiable for  $Y$  if there is at least one pair  $(\theta, \theta')$ ,  $\theta \neq \theta'$ , where  $\theta$  and  $\theta'$  both belong to  $\Omega$  such that  $p_{\theta}(y) = p_{\theta'}(y)$  for all  $y$ . In the contrary case we shall say  $\theta$  is identifiable.

Therefore, a parametric statistical model is said to be identified if the distribution  $p_{\theta}(y)$  is generated by a unique value,  $\theta$ . When a model is not identified, the parameters of interest are not unique and therefore cannot be used for drawing inference. As models become more complex with higher-dimensional parameter spaces, identifiability becomes much more difficult to ascertain (see Section 3.2.1 for an example).

Another way of illustrating identifiability is in the context of conditional independence (Dawid, 1979). Casting identifiability in the light of conditional independence closely relates to the Bayesian paradigm of posterior distributions. Letting  $Y$  be from a distribution defined by the pair of parameters  $(\theta_1, \theta_2)$ , we interpret  $(Y \perp\!\!\!\perp \theta_2) | \theta_1$  as the distribution of  $Y$  being determined solely by  $\theta_1$ , where  $\theta_2$  is redundant given  $\theta_1$ . Here,  $\theta_1$  can be referred to as a sufficient parameter, which is similar in context to sufficient statistics (Barankin, 1960). This implies that once we have learned about  $\theta_1$  from the data, there is nothing more we can learn about  $\theta_2$  beyond what we already know. Therefore,  $\theta_2$  is not identified by the data.

In the Bayesian framework we have the likelihood function,  $L(\theta_1, \theta_2; \mathbf{y})$ , where  $\mathbf{y}$  is the observed data and assign prior distributions to  $\theta_1$  and  $\theta_2$ . The posterior of  $\theta_2$  is written as

$$p(\theta_2 | \mathbf{y}, \theta_1) \propto L(\theta_1, \theta_2; \mathbf{y}) p(\theta_2 | \theta_1) p(\theta_1). \quad (21)$$

We say  $\theta_2$  is not *likelihood identifiable* if and only if  $L(\theta_1, \theta_2; \mathbf{y})$  can be written free of  $\theta_2$ . As is the case when  $\theta_1$  is a sufficient parameter such that  $(Y \amalg \theta_2) | \theta_1$ , the posterior of  $\theta_2$  is

$$p(\theta_2 | \mathbf{y}, \theta_1) \propto p(\theta_2 | \theta_1). \quad (22)$$

This implies that the conditional distribution of the redundant parameter,  $\theta_2$ , given the sufficient parameter,  $\theta_1$ , is the same in the posterior distribution as in the prior.

There is a subtle difference when we switch from *likelihood identifiability* to *Bayesian identifiability*. If there exists a one-to-one transformation from  $(\theta_1, \theta_2)$  to  $(\phi_1, \phi_2)$  such that  $\phi_2$  is not identifiable in that (22) holds when we replace  $\theta_1$  and  $\theta_2$  by  $\phi_1$  and  $\phi_2$ , then we say the model is not *Bayesian identifiable* (Gelfand and Sahu, 1999). Whereas models that are not likelihood identifiable can be more obvious to detect, Bayesian nonidentifiability signifies that it is not a sufficient condition for parameter identifiability for both  $\theta_1$  and  $\theta_2$  to be in the likelihood,  $L(\theta_1, \theta_2; \mathbf{y})$ . At the risk of oversimplifying, the distinction between likelihood nonidentifiability and Bayesian nonidentifiability is that  $\theta_2$  is likelihood nonidentifiable if the likelihood function is free of  $\theta_2$  and Bayesian nonidentifiable if the posterior distribution of  $\theta_2$  is free of the data,  $\mathbf{y}$ .

Bayesian hierarchical models (BHMs) are a specific collection of models that can suffer from Bayesian nonidentifiability since they are often written as

$$p(\phi_1, \phi_2 | \mathbf{y}) \propto L(\phi_1; \mathbf{y}) p(\phi_1 | \phi_2) p(\phi_2).$$

Therefore, the posterior of  $\phi_2$  follows (22) in that

$$p(\phi_2 | \phi_1, \mathbf{y}) \propto p(\phi_2 | \phi_1).$$

Swartz et al. (2004) say “nonidentifiability according to Basu (1983) exists for any hierarchical model.” Gelfand and Sahu (1999) consider models that are not Bayesian identifiable to

be contained in a class of “weakly identifiable” models. Extending the formal definition of identifiability given in (1), we say that  $\boldsymbol{\theta}$  is *near nonidentifiable*, or *weakly identifiable*, if  $p_{\boldsymbol{\theta}}(\mathbf{y}) \approx p_{\boldsymbol{\theta}'}(\mathbf{y})$  for all  $\mathbf{y}$  and some suitable metric (Swartz et al., 2004). Therefore, near nonidentifiability occurs when the likelihood surface is flat such that changes in the parameter  $\boldsymbol{\theta}$  result in insignificant changes in the likelihood. The vagueness of this terminology suggests an area for future research. The implications of nonidentifiability discussed above extend to near nonidentifiability and can therefore impede accurate parameter inference. Near nonidentifiability is common in models with random effects when the model has a high-dimensional parameter space.

Weakly identifiable models make up an interesting class of models because they still have the potential to provide useful statistical inference when fitted well. Seeing that parameter identifiability is not an artifact of prior distributions, weakly identifiable parameters can be assigned informative prior distributions that maximize *Bayesian learning*. For example, Gelfand and Sahu (1999) suggest vague prior distributions when fitting a partially-identified Gaussian linear model with design matrix less than full rank. In this case, convergence rates greatly improve as the variance of the prior distribution of the coefficient,  $\boldsymbol{\beta}$ , goes to infinity. Informally, Bayesian learning is the ability to gain information about a parameter from the data beyond that which stems from the prior distribution. It results from comparing the parameter’s posterior distribution to its prior distribution. An important thing to understand, however, is that nonidentifiability, and thus, weak identifiability, does not preclude Bayesian learning (e.g. Xie and Carlin, 2006; Gelfand and Sahu, 1999). Whereas nonidentifiability is such that  $p(\theta_2|\mathbf{y}, \theta_1) = p(\theta_2|\theta_1)$ , this is a weaker condition than  $p(\theta_2|\mathbf{y}) = p(\theta_2)$ , which implies no Bayesian learning. Therefore, in a Bayesian framework one can still fit models that contain nonidentifiable parameters when Bayesian learning of the nonidentifiable parameters exists. Xie and Carlin (2006) propose methods for quantifying the amount of potential information one can obtain about unidentifiable or weakly identifiable parameters in the context of Gaussian hierarchical linear models.

In order to resolve nonidentifiability, one must incorporate non-data-based information into the model. Whereas frequentists do this by instilling constraints on the parameter space, Bayesians have the additional option of placing proper prior distributions on the unidentified parameters. Prior distributions created specifically for a particular parameter, e.g., spatial range parameter, can lead to suitable inference of models fitted with Bayesian nonidentifiable parameters. Further, prior distributions also influence the level of Bayesian learning. Unfortunately, there are cases when even well-assigned priors are not able to assist in fitting nonidentifiable models. This results in the posterior and prior distributions being identical meaning that all information about the parameter comes from the prior distribution as opposed to learning from the data.

## 3.2 Identifiability of non-spatial latent variable models

Latent variable models, and specifically, latent Gaussian models, have become a popular class of models. They are very flexible models that can be fitted within a GLM framework for various types of data and can extend to GLMMs when the model contains additional random effects. Similarly, latent variable models can be used when fitting probit linear models (PLMs) and probit linear mixed models (PLMMs). In spatial statistics, for example, latent Gaussian models can be used to model a Gaussian random field. Model identifiability comes into question when fitting latent variable models since latent variables are not observed and only inferred by the model. The issue compounds when fitting multivariate and multilevel latent variable models (see Chapter 5 for an example of such a model). We are interested in investigating identifiability in a subset of latent Gaussian models used for modeling multivariate ordinal response data. In particular, we assume the multivariate processes have spatial correlation which we would like to capture within the latent variable model.

To first explore identifiability in latent variable models, we consider a common factor model or one factor model for ordinal response data. This is a simplified version of the model we propose in Chapter 5. Suppose the response variable,  $\mathbf{Y}(\mathbf{s})$  at location  $\mathbf{s}$  is multivariate

ordinal and assumed to be generated by an underlying continuous latent variable,  $\mathbf{Z}(\mathbf{s})$ . We define the  $j$ th continuous latent variable for location  $\mathbf{s}_i$  as  $Z_j(\mathbf{s}_i)$ .

We model the continuous latent variable as

$$Z_j(\mathbf{s}_i) = \theta_j + \omega_j W(\mathbf{s}_i) + \epsilon_j(\mathbf{s}_i) \quad (23)$$

where  $W(\mathbf{s}_i) \stackrel{iid}{\sim} N(\eta, \phi)$  is the common factor with variable-specific factor loadings  $\omega_j$ , and  $\epsilon_j(\mathbf{s}_i) \stackrel{iid}{\sim} N(0, \sigma_j^2)$ .

Dropping the dependence on  $\mathbf{s}$  for notational ease, we denote  $Z_j(\mathbf{s}_i)_j$  as  $Z_{ij}$ ,  $W(\mathbf{s}_i)$  as  $W_i$ , and  $\epsilon_j(\mathbf{s}_i)$  as  $\epsilon_{ij}$ . Notice that this model is simplified by assuming the mean of  $W_i$  is constant for all locations and the variance is set to a scalar parameter,  $\phi$ , as opposed to a valid spatial covariance function. To show that this model is not identifiable, let  $W_i^* = aW_i + b$ . Then, write

$$\begin{aligned} Z_{ij} &= \theta_j + \omega_j W_i + \epsilon_{ij} \\ &= \theta_j + \omega_j \frac{W_i^* - b}{a} + \epsilon_{ij} \\ &= \left( \theta_j - \frac{\omega_j b}{a} \right) + \frac{\omega_j}{a} W_i^* + \epsilon_{ij} \\ &= \theta_j^* + \omega_j^* W_i + \epsilon_{ij} \end{aligned}$$

where the parameters of interest are now written as  $\theta_j^* = \theta_j - \frac{\omega_j b}{a}$  and  $\omega_j^* = \frac{\omega_j}{a}$ . Therefore, the model is not identifiable because different parameter value sets,  $\{\theta_j, \omega_j\}$  and  $\{\theta_j^*, \omega_j^*\}$ , generate the same reduced-form distribution of  $Z_{ij}$ .

### 3.2.1 Approaches for showing identifiability

As conveyed in the motivating example with the common factor model (23), showing a model is nonidentifiable can often be done in just a few lines. Showing that a model is identifiable, on the other hand, is generally much more complicated. To our knowledge,

there does not exist a general framework for showing a model to be identifiable. The main reason being that identifiability is very model-specific. Nevertheless, there do exists some basic guidelines that modelers can extend or modify to check identifiability in their model.

Rothenberg (1971) derive several identifiability criteria in the case of general parametric models. His method is based on the information matrix of classical mathematical statistics. Here, the information matrix is a measure of the amount of information about the unknown parameters that is available in the sample data. The lack of sufficient information in the observed data directly results in a lack of identifiability in the model. The information matrix method is arguably not feasible in practice, however, since most complex models are usually analytically intractable.

Wald (1950) propose a different method for checking parameter identifiability that focuses on a mapping between structural and reduced-form parameters. Structural parameters are the parameters that are of scientific interest and defined when writing down the model. Reduced-form parameters are the parameters that can be estimated from the data. Reduced-form parameters can often be defined as functions of structural parameters. For example, in Section 3.1, we defined a simple model where  $\mathbf{Y} \sim N(\theta_1 + \theta_2, \sigma^2)$ . The structural parameters are those defined in the model,  $\theta_1$ ,  $\theta_2$ , and  $\sigma^2$ . However, we know that only the sum,  $\theta_1 + \theta_2$ , and the variance,  $\sigma^2$ , are identifiable. Therefore, our reduced-form parameters are  $\phi$  and  $\sigma^2$  where  $\phi = \theta_1 + \theta_2$ . Since the two sets of parameters differ, we need to be conscious of the ability of the model in drawing inference on the structural model we impose. When a model is identifiable, we can draw inference on the scientific (structural) parameters of interest through the reduced-form parameters. Thus, the observed data are able to inform inference on the structural parameters.

Dupacová and Wold (1982) apply the mapping approach to structural equation models with latent variables. This method was developed and generally used for models where the observable random variable is Gaussian because they can be defined in entirety by their 1<sup>st</sup> and 2<sup>nd</sup> moments. The reduced-form parameters are the mean and variance-covariance

parameters of the reduced-form distribution. The mapping approach is extended to models with dichotomous and ordinal response models generated by Gaussian latent variables by Skrondal and Rabe-Hesketh (2004, Chapter 5). We begin by first outlining their method and then apply it to our multivariate multilevel latent variable model from Chapter 5.

Define the fundamental parameter vector of length  $T$  as  $\boldsymbol{\theta} \in A$  and the reduced-form parameter vector of length  $S$  as  $\boldsymbol{\xi} \in A'$ . We then define the mapping functions between the fundamental and reduced parameter sets as

$$\xi_s = h_s(\boldsymbol{\theta}) \text{ for } 1 \leq s \leq S$$

where the mapping functions,  $h_s(\cdot)$ , are continuously differentiable known functions. Lemma 1 gives a necessary but not sufficient condition for parameter identifiability (Skrondal and Rabe-Hesketh, 2004, Chapter 5).

**Lemma 1.** *The number of reduced-form parameters  $S$  be greater than or equal to the number of unknown fundamental parameters in the model.*

The probability distribution of estimable variables depends on the  $T$ -dimensional fundamental parameter vector,  $\boldsymbol{\theta}$ , only through the  $S$ -dimensional reduced-form parameter vector,  $\boldsymbol{\xi}$ . For all  $\boldsymbol{\theta} \in A$ , the probability distribution is such that

$$f(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) = f^*(\mathbf{y}|\mathbf{X}, h_1(\boldsymbol{\theta}), h_2(\boldsymbol{\theta}), \dots, h_S(\boldsymbol{\theta})) = f^*(\mathbf{y}|\mathbf{X}, \boldsymbol{\xi}).$$

We study the identifiability of  $\boldsymbol{\theta}$  through the characteristics of the mapping of  $\boldsymbol{\theta} \rightarrow \boldsymbol{\xi}$ . For some fundamental parameter vector  $\boldsymbol{\theta}^o$  that generates reduced parameter vector  $\boldsymbol{\xi}^o$ , we can define  $\xi_s^o = h_s(\boldsymbol{\theta}^o)$  for  $1 \leq s \leq S$ . Then we say that  $\boldsymbol{\theta}^o$  is identifiable if and only if  $\boldsymbol{\theta}^o$  is the unique solution of the equations  $\xi_s^o = h_s(\boldsymbol{\theta})$  for  $1 \leq s \leq S$ . Therefore, the uniqueness of the solutions to the systems of equations  $h_s(\boldsymbol{\theta})$ ,  $1 \leq s \leq S$ , determines identifiability of the fundamental parameters.

We investigate the mapping between  $\boldsymbol{\theta} \rightarrow \boldsymbol{\xi}$  by computing the Jacobian matrix,  $[J(\boldsymbol{\theta})]$ , defined as

$$[J(\boldsymbol{\theta})] = \left[ \frac{dh_s}{d\theta_t}, 1 \leq s \leq S, 1 \leq t \leq T \right].$$

The Jacobian matrix contains the derivative of each mapping function  $h_s(\cdot)$  with respect to each of the  $T$  fundamental parameters of  $\boldsymbol{\theta}$ . Therefore, the dimension of the Jacobian is  $S \times T$ . As stated above, a necessary condition for parameter identifiability is that  $S \geq T$ . When the number of columns of  $[J(\boldsymbol{\theta})]$  is greater than the number of rows, the model is not identifiable. Shown by Skrondal and Rabe-Hesketh (2004, p. 128), a unique solution  $\boldsymbol{\theta}^o$  exists if and only if the rank $[J(\boldsymbol{\theta})] = T$ .

### 3.2.2 Applying mapping approach to multivariate multilevel latent variable model

To apply the mapping method to our multivariate multilevel latent variable model for ordinal response data, we first need to determine the fundamental parameters. Then, we must deduce the reduced-form parameters of the model and derive the mapping functions between the two. One of our model assumptions is that there exists an underlying multivariate Gaussian latent variable that is being thresholded into the multivariate ordinal response variables. Therefore, the reduced-form model parameters are the means and variance-covariance parameters that can be computed from the observed data through the assumed underlying multivariate Gaussian distributions.

The observed multivariate ordinal response is  $\mathbf{y} = [\mathbf{y}_1, \dots, \mathbf{y}_J]$ , where the rows of  $\mathbf{y}$  are of the form  $\mathbf{y}_i = [y_{i1}, \dots, y_{iJ}]$  and the columns are of the form  $\mathbf{y}_j = [y_{1j}, \dots, y_{nj}]'$ . The subscript  $i$  denotes locations, where we dropped writing  $\mathbf{s}_i$  for notational convenience, and subscript  $j$  denotes metric. We let  $n$  be the number of observed locations,  $J$  be the number of metrics in the response, and  $K$  be the number of ordinal categories such that  $y_{ij} \in \{1, \dots, K\}$ . We

will derive model identification for the case where  $J = 3$  and  $K = 5$ , however the methods can be generalized to other values of  $J$  and  $K$ .

### 3.2.2.1 Discerning the fundamental and reduced-form parameters

The deterministic relationship between the observable ordinal response variable  $Y_{ij}$  and the latent continuous random variable  $Z_{ij}$ , for  $i = 1, \dots, n$  and  $j = 1, \dots, 3$  is

$$Y_{ij} = \begin{cases} 1 & -\infty < Z_{ij} \leq \lambda_1 \\ 2 & 0 < Z_{ij} \leq \lambda_2 \\ 3 & \lambda_2 < Z_{ij} \leq \lambda_3 \\ 4 & \lambda_3 < Z_{ij} \leq \lambda_4 \\ 5 & \lambda_4 < Z_{ij} < \infty \end{cases}$$

We assume  $Z_{ij}$  is the underlying continuous latent Gaussian variable that is generating the ordinal response,  $Y_{ij}$ . We let  $Z_{ij} = \theta_j + \omega_j W_i + \epsilon_{ij}$  where  $W_i \stackrel{iid}{\sim} N(\eta, \phi)$  and  $\epsilon_{ij} \stackrel{iid}{\sim} N(0, \sigma_j^2)$ . The threshold parameter vector,  $\boldsymbol{\lambda}$ , is constrained such that  $-\infty = \lambda_0 \leq \lambda_1 \leq \dots \lambda_K = \infty$ . Albert and Chib (1993) define this binning approach for ordered categorical data and require one restriction on bin boundaries to ensure identifiability. This constraint can be shown to be necessary by letting  $\lambda_0 = -\infty$  and  $\lambda_K = \infty$ , where  $K$  is the number of categories of the ordinal response. If the remaining threshold parameters,  $\lambda_k$ , for  $k = 1, \dots, (K - 1)$  are to be estimated, we are unable to identify the  $(J \times 1)$ -dimension vector of metric-specific random effects,  $\boldsymbol{\theta}$ . By writing out the marginal probability

$$P(Y_{ij} = 1) = P(Z_{ij} < \lambda_1) = \Phi \left( \frac{\lambda_1 - (\theta_j + \omega_j \eta)}{\sqrt{\omega_j^2 \phi + \sigma_j^2}} \right).$$

and letting  $\lambda_1^* = \lambda_1 + c$  for  $c \neq 0$  and  $\theta_j^* = \theta_j - c$ , then

$$\Phi\left(\frac{\lambda_1^* - (\theta_j^* + \omega_j\eta)}{\sqrt{\omega_j^2\phi + \sigma_j^2}}\right) = \Phi\left(\frac{\lambda_1 - (\theta_j + \omega_j\eta)}{\sqrt{\omega_j^2\phi + \sigma_j^2}}\right).$$

This result holds for all  $j = 1, \dots, J$  since we assume the threshold vector is not metric-specific. Therefore, without loss of generality, we set  $\lambda_1 = 0$ . There are  $T = 14$  remaining parameters to be estimated which define the fundamental parameter vector,  $\boldsymbol{\theta}$ . That is,

$$\boldsymbol{\theta} = \{\lambda_2, \lambda_3, \lambda_4, \theta_1, \theta_2, \theta_3, \omega_1, \omega_2, \omega_3, \sigma_1^2, \sigma_2^2, \sigma_3^2, \eta, \phi\}.$$

Our reduced-form parameter vector,  $\boldsymbol{\xi}$ , contains the means and variance-covariance values of our latent continuous random variables  $Z_{ij}$ . Since the observable response variables,  $\mathbf{Y}$ , are ordinal and can be defined using the latent continuous variables,  $\mathbf{Z}$ , and threshold vector,  $\boldsymbol{\lambda}$ , we begin by computing the following:

$$\begin{aligned} P(Y_{ij} = 1) &= P(Z_{ij} \leq 0) = P\left(\frac{Z_{ij} - (\theta_j + \omega_j\eta)}{\sqrt{\omega_j^2\phi + \sigma_j^2}} \leq \frac{-(\theta_j + \omega_j\eta)}{\sqrt{\omega_j^2\phi + \sigma_j^2}}\right) \\ &= P\left(Z_{ij}^* \leq \frac{-(\theta_j + \omega_j\eta)}{\sqrt{\omega_j^2\phi + \sigma_j^2}}\right) \end{aligned}$$

where  $Z_{ij}^*$  is standard normal and  $j = 1, \dots, 3$ . Therefore, we can show

$$P(Y_{ij} = 1) = \Phi\left(\frac{-(\theta_j + \omega_j\eta)}{\sqrt{\omega_j^2\phi + \sigma_j^2}}\right) \quad (24)$$

where  $\Phi$  is the cumulative distribution function of a standard normal random variable. Likewise, we can write out

$$P(Y_{ij} = 2) = P(Z_{ij} \leq \lambda_2) - P(Z_{ij} \leq 0) = \Phi\left(\frac{\lambda_2 - (\theta_j + \omega_j\eta)}{\sqrt{\omega_j^2\phi + \sigma_j^2}}\right) - \Phi\left(\frac{-(\theta_j + \omega_j\eta)}{\sqrt{\omega_j^2\phi + \sigma_j^2}}\right). \quad (25)$$

These same calculations can be made for  $P(Y_{ij} = 3)$  and  $P(Y_{ij} = 4)$ . Lastly,

$$P(Y_{ij} = 5) = 1 - \Phi\left(\frac{\lambda_4 - (\theta_j + \omega_j\eta)}{\sqrt{\omega_j^2\phi + \sigma_j^2}}\right). \quad (26)$$

We define the mean and threshold reduced-form parameters for  $j = 1, \dots, J$  using (24), (25), and (26). From (24), the mean for each metric  $j$  is given by

$$\mu_j = \frac{\theta_j + \omega_j\eta}{\sqrt{\omega_j^2\phi + \sigma_j^2}}.$$

We discern the thresholds from (25) and (26) as

$$\tau_{2j} = \frac{\lambda_2}{\sqrt{\omega_j^2\phi + \sigma_j^2}}$$

$$\tau_{3j} = \frac{\lambda_3}{\sqrt{\omega_j^2\phi + \sigma_j^2}}$$

$$\tau_{4j} = \frac{\lambda_4}{\sqrt{\omega_j^2\phi + \sigma_j^2}}$$

The correlation matrix of the latent response variables is

$$\mathbf{R}(\boldsymbol{\theta}) = \begin{bmatrix} 1 & & \\ c_{21} & 1 & \\ c_{31} & c_{32} & 1 \end{bmatrix} = \begin{bmatrix} 1 & & \\ \frac{\omega_2\omega_1\phi}{\sqrt{\omega_2^2\phi+\sigma_2^2}\sqrt{\omega_1^2\phi+\sigma_1^2}} & 1 \\ \frac{\omega_3\omega_1\phi}{\sqrt{\omega_3^2\phi+\sigma_3^2}\sqrt{\omega_1^2\phi+\sigma_1^2}} & \frac{\omega_3\omega_2\phi}{\sqrt{\omega_3^2\phi+\sigma_3^2}\sqrt{\omega_2^2\phi+\sigma_2^2}} & 1 \end{bmatrix}$$

where  $c_{21} = \text{corr}(\mathbf{Z}_2, \mathbf{Z}_1)$ , for example. In this model, there is only correlation between metric since the latent variables are assumed independent within metric. Our reduced-form parameter vector is given by

$$\boldsymbol{\xi} = \{\mu_1, \mu_2, \mu_3, \tau_{21}, \tau_{22}, \tau_{23}, \tau_{31}, \tau_{32}, \tau_{33}, \tau_{41}, \tau_{42}, \tau_{43}, c_{21}, c_{31}, c_{32}\}$$

Notice that the length of  $\boldsymbol{\xi}$  is  $S = 15$  and the number of fundamental parameters is  $T = 14$ . While it is a necessary condition that for the model to be identifiable,  $S \geq T$ , this is not a sufficient condition. We know that a simpler version of this model is not identifiable from the common factor model example at the beginning of Section 3.2.

### 3.2.2.2 Imposing parameter constraints

To achieve parameter identifiability, Skrondal and Rabe-Hesketh (2004) propose setting the mean and variance of the common factor,  $W_i$ , to 0 and 1, respectively. Whereas our initial goal is to establish parameter identifiability for the general model with  $W_i \sim N(\eta, \phi)$ , our main objective is to show identifiability in a complex common factor model containing covariates and a correlated error term. Therefore, we do not wish to fix the parameters of the distribution of  $W_i$  in order for identifiability in the general model to be naturally extended to the complex model. We begin to establish identifiability of the model parameters through a series of steps. The first step is to set  $\omega_1 = 1$  to establish a benchmark relationship between the latent response,  $Z_{ij}$ , and the latent variable, or common factor,  $W_i$ . We will show that this is not sufficient to ensure identifiability. Our second step is to fix one of the variance parameters. Without loss of generality, we set  $\sigma_1^2 = 1$ . We argue that while these parameter

restrictions do not make the general model identifiable, they aid in establishing identifiability in the complex model.

The factor-loading parameter vector,  $\boldsymbol{\omega}$ , contains metric-specific values that relate the latent variable,  $\mathbf{W}$ , to each of the response variables. Since  $\boldsymbol{\omega}$  is only necessary for multivariate response models, the model is able to identify only  $J - 1$  variables of the  $(J \times 1)$ -dimension vector,  $\boldsymbol{\omega}$ . This can be seen by letting  $\omega_j^* = c\omega_j$  and  $\mathbf{W}^* = \frac{1}{c}\mathbf{W}$  for  $c \neq 0$ . Then  $\omega_j\mathbf{W} = \omega_j^*\mathbf{W}^*$  and the model does not change. By fixing one of the factor loadings, the model establishes a base relationship between one of the response variable and the latent variable  $\mathbf{W}$ . This is similar to what is done in analysis of variance (ANOVA) where for a model with  $g$  groups, only  $g - 1$  group effects are identifiable when there exists an intercept or grand mean term in the model. Without loss of generality, we set  $\omega_1 = 1$ . The remaining factor loadings,  $\omega_j$ , for  $j = 2, \dots, J$  can adjust to estimate the metric-specific relationships between the latent variable  $\mathbf{Z}_j$  and  $\mathbf{W}$ .

Because  $Z_{ij}$  is latent, it can be fixed to have any scale. That is, letting  $\sigma_j^* = c\sigma_j$ ,  $\lambda_4^* = c\lambda_4$ ,  $\theta_j^* = c\theta_j$  and  $\eta^* = c\eta$ ,

$$\Phi\left(\frac{\lambda_4^* - (\theta_j^* + \omega_j\eta^*)}{\sigma_j^*}\right) = \Phi\left(\frac{\lambda_4 - (\theta_j + \omega_j\eta)}{\sigma_j}\right).$$

For a univariate response variable, we would fix the scale parameter  $\sigma^2 = 1$ . Since our model is multivariate, by fixing one of the scale parameters, namely  $\sigma_1^2 = 1$ , the conditional variance of  $Z_{ij}$ ,  $\sigma_j^2$ , can be uniquely estimated for  $j = 2, \dots, J$ .

After fixing both  $\omega_1 = 1$  and  $\sigma_1^2 = 1$ , we compute the Jacobian matrix

$$[J(\boldsymbol{\theta})] = \left[ \frac{dh_s}{d\theta_t}, 1 \leq s \leq 15, 1 \leq t \leq 12 \right].$$

Let  $v_1 = \phi + 1$ ,  $v_2 = \omega_2^2\phi + \sigma_2^2$ , and  $v_3 = \omega_3^2\phi + \sigma_3^2$ . The Jacobian matrix is

$$\begin{bmatrix}
0 & 0 & 0 & \frac{1}{\sqrt{v_1}} & 0 & 0 & 0 & 0 & 0 & \frac{1}{\sqrt{v_1}} & -\frac{1}{2}(\theta_1 + \eta) \\
0 & 0 & 0 & 0 & \frac{1}{\sqrt{v_2}} & 0 & \frac{\eta\sigma_2^2 - \theta_2\omega_2\phi}{(v_2)^{\frac{3}{2}}} & 0 & -\frac{1}{2}(\theta_2 + \omega_2\eta) & 0 & \frac{(\phi+1)^{\frac{3}{2}}}{\frac{1}{2}\omega_2^2(\theta_2 + \omega_2\eta)} \\
0 & 0 & 0 & 0 & 0 & \frac{1}{\sqrt{3}} & 0 & \frac{\eta\sigma_3^2 - \theta_3\omega_3\phi}{(v_3)^{\frac{3}{2}}} & 0 & -\frac{1}{2}(\theta_3 + \omega_3\eta) & \frac{(\v_2)^{\frac{3}{2}}}{-\frac{1}{2}\omega_3^2(\theta_3 + \omega_3\eta)} \\
\frac{1}{\sqrt{v_1}} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{(\v_1)^{\frac{3}{2}}}{-\frac{1}{2}\lambda_2} \\
\frac{1}{\sqrt{v_2}} & 0 & 0 & 0 & 0 & 0 & \frac{-\lambda_2\omega_2\phi}{(v_2)^{\frac{3}{2}}} & 0 & -\frac{1}{2}\lambda_2 & 0 & \frac{(\v_2)^{\frac{3}{2}}}{-\frac{1}{2}\omega_2^2\lambda_2} \\
\frac{1}{\sqrt{v_3}} & 0 & 0 & 0 & 0 & 0 & 0 & \frac{-\lambda_2\omega_3\phi}{(v_3)^{\frac{3}{2}}} & 0 & -\frac{1}{2}\lambda_2 & \frac{(\v_3)^{\frac{3}{2}}}{-\frac{1}{2}\omega_3^2\lambda_2} \\
0 & \frac{1}{\sqrt{\phi+1}} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{(\phi+1)^{\frac{3}{2}}}{-\frac{1}{2}\lambda_3} \\
0 & \frac{1}{\sqrt{v_2}} & 0 & 0 & 0 & 0 & \frac{-\lambda_3\omega_2\phi}{(v_2)^{\frac{3}{2}}} & 0 & -\frac{1}{2}\lambda_3 & 0 & \frac{(\v_2)^{\frac{3}{2}}}{-\frac{1}{2}\omega_2^2\lambda_3} \\
0 & \frac{1}{\sqrt{v_3}} & 0 & 0 & 0 & 0 & 0 & \frac{-\lambda_3\omega_3\phi}{(v_3)^{\frac{3}{2}}} & 0 & -\frac{1}{2}\lambda_3 & \frac{(\v_3)^{\frac{3}{2}}}{-\frac{1}{2}\omega_3^2\lambda_3} \\
0 & 0 & \frac{1}{\sqrt{\phi+1}} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{(\phi+1)^{\frac{3}{2}}}{-\frac{1}{2}\lambda_4} \\
0 & 0 & \frac{1}{\sqrt{v_2}} & 0 & 0 & 0 & \frac{-\lambda_4\omega_2\phi}{(v_2)^{\frac{3}{2}}} & 0 & -\frac{1}{2}\lambda_4 & 0 & \frac{(\v_2)^{\frac{3}{2}}}{-\frac{1}{2}\omega_2^2\lambda_4} \\
0 & 0 & \frac{1}{\sqrt{v_3}} & 0 & 0 & 0 & 0 & \frac{-\lambda_4\omega_3\phi}{(v_3)^{\frac{3}{2}}} & 0 & -\frac{1}{2}\lambda_4 & \frac{(\v_3)^{\frac{3}{2}}}{-\frac{1}{2}\omega_3^2\lambda_4} \\
0 & 0 & 0 & 0 & 0 & 0 & \frac{\sigma_2^2\phi}{(v_2)^{\frac{3}{2}}(v_1)^{\frac{1}{2}}} & 0 & -\frac{1}{2}\omega_2\phi & 0 & \frac{(\v_1)^{\frac{3}{2}}(v_2)^{\frac{3}{2}}}{\omega_2(\sigma_2^2 + \frac{1}{2}\phi\sigma_2^2 + \frac{1}{2}\omega_2^2\phi)} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{\sigma_3^2\phi}{(v_3)^{\frac{3}{2}}(v_1)^{\frac{1}{2}}} & 0 & -\frac{1}{2}\omega_3\phi & 0 & \frac{(\v_1)^{\frac{3}{2}}(v_3)^{\frac{3}{2}}}{\omega_3(\sigma_3^2 + \frac{1}{2}\phi\sigma_3^2 + \frac{1}{2}\omega_3^2\phi)} \\
0 & 0 & 0 & 0 & 0 & 0 & \frac{\omega_3\phi\sigma_2^2}{(v_3)^{\frac{3}{2}}(v_2)^{\frac{3}{2}}} & \frac{\omega_2\phi\sigma_3^2}{(v_3)^{\frac{3}{2}}(v_2)^{\frac{3}{2}}} & -\frac{1}{2}\omega_3\omega_2\phi & 0 & \frac{(\v_1)^{\frac{3}{2}}(v_3)^{\frac{3}{2}}}{\omega_3\omega_2(\sigma_2^2\sigma_3^2 + \frac{1}{2}\omega_2^2\phi\sigma_2^2 + \frac{1}{2}\omega_3^2\phi\sigma_3^2)}
\end{bmatrix} \quad [15 \times 12] \quad (27)$$

$$\begin{bmatrix} \frac{d\mu_1}{d\theta_1} & \frac{d\mu_1}{d\theta_2} & \frac{d\mu_1}{d\theta_3} & \frac{d\mu_1}{d\eta} \\ \frac{d\mu_2}{d\theta_1} & \frac{d\mu_2}{d\theta_2} & \frac{d\mu_2}{d\theta_3} & \frac{d\mu_2}{d\eta} \\ \frac{d\mu_3}{d\theta_1} & \frac{d\mu_3}{d\theta_2} & \frac{d\mu_3}{d\theta_3} & \frac{d\mu_3}{d\eta} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{dc_{32}}{d\theta_1} & \frac{dc_{32}}{d\theta_2} & \frac{dc_{32}}{d\theta_3} & \frac{dc_{32}}{d\eta} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{\phi+1}} & 0 & 0 & \frac{1}{\sqrt{\phi+1}} \\ 0 & \frac{1}{\sqrt{\omega_2^2\phi+\sigma_2^2}} & 0 & \frac{\omega_2}{\sqrt{\omega_2^2\phi+\sigma_2^2}} \\ 0 & 0 & \frac{1}{\sqrt{\omega_3^2\phi+\sigma_3^2}} & \frac{\omega_3}{\sqrt{\omega_3^2\phi+\sigma_3^2}} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 \end{bmatrix}_{[15 \times 4]} \quad (28)$$

It can be seen that the Jacobian matrix in (27) is not full rank. The matrix in (28) focuses on four of the columns of the full Jacobian matrix (27), showing that a linear combination of the first 3 columns is equal to the fourth column. Therefore,  $\text{rank}[\mathbf{J}(\boldsymbol{\theta})] = 11$ , which is less than the number of fundamental parameters in the model.

We can understand the lack of parameter identifiability by examining the marginal distribution of  $Z_{ij}$ . The expected value of the marginal distribution is  $E(Z_{ij}) = \theta_j + \omega_j\eta$ . For metric  $j = 1$ ,  $E(Z_{i1}) = \theta_1 + \eta$ . Letting  $\theta_1^* = \theta_1 + c$  and  $\eta^* = \eta - c$  for  $c \neq 0$ ,  $E(Z_{i1}) = \theta_1 + \eta = \theta_1^* + \eta^*$  and the model stays the same. Therefore we would need to fix either one of the  $J$  parameters in the vector  $\boldsymbol{\theta}$  or the mean of the common factor,  $\eta$ , to a known constant to preserve parameter identifiability for the general model where  $W_i \stackrel{iid}{\sim} N(\eta, \phi)$ . For ease of interpretation and metric comparison, we would fix  $\eta = 0$ .

### 3.2.2.3 Extension to the proposed model for wetland health

The model proposed in (57) of Chapter 5 aims to model the common factor  $\mathbf{W}$  as a latent Gaussian process. We assume the latent Gaussian process has spatial correlation as well as location-specific covariates in the mean. That is,  $\mathbf{W}(\mathbf{s}) \sim GP(\mathbf{X}(\mathbf{s})\boldsymbol{\beta}, \Sigma_W)$  where  $\Sigma_W$  is a spatial covariance matrix. Unfortunately, the mapping approach does not easily apply to this model because the mean, threshold, and correlation parameters defined in Section 3.2.2.1 would be location-metric-specific. However, we can deduce identifiability of the coefficient vector,  $\boldsymbol{\beta}$ , by writing out a non-spatial form of the model in matrix form. We

postpone investigating identifiability of the latent spatial process parameters until Section 3.3.

Assume a simpler, non-spatial model where  $\mathbf{W}(\mathbf{s}) \sim GP(\mathbf{X}(\mathbf{s})\beta, \phi\mathbf{I})$  and  $\phi$  is the variance parameter. The identifiability issue between parameter vector  $\boldsymbol{\theta}$  and  $\eta$  shown in Section 3.2.2.2 is avoided by eliminating the 1-vector in  $\mathbf{X}(\mathbf{s})$ . This can be shown through the following example. Define  $Z_j(\mathbf{s}_i) = \theta_j + \omega_j W(\mathbf{s}_i) + \epsilon_{ij}$  where  $i = 1, \dots, 4$  and  $j = 1, 2$ . Assume further that  $\epsilon_{ij} \stackrel{iid}{\sim} N(0, \sigma_j^2)$  where  $\sigma_1^2 = 1$  and  $W(\mathbf{s}_i) \sim N(X(\mathbf{s}_i)' \beta, \phi)$ . Fixing  $\omega_1 = 1$ , we write the conditional distribution of  $\mathbf{Z}(\mathbf{s})$  in matrix notation as

$$\begin{bmatrix} Z_1(\mathbf{s}_1) \\ Z_1(\mathbf{s}_2) \\ Z_1(\mathbf{s}_3) \\ Z_1(\mathbf{s}_4) \\ Z_2(\mathbf{s}_1) \\ Z_2(\mathbf{s}_2) \\ Z_2(\mathbf{s}_3) \\ Z_2(\mathbf{s}_4) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} + \begin{bmatrix} W(\mathbf{s}_1) & 0 \\ W(\mathbf{s}_2) & 0 \\ W(\mathbf{s}_3) & 0 \\ W(\mathbf{s}_4) & 0 \\ 0 & W(\mathbf{s}_1) \\ 0 & W(\mathbf{s}_2) \\ 0 & W(\mathbf{s}_3) \\ 0 & W(\mathbf{s}_4) \end{bmatrix} + \begin{bmatrix} \epsilon_{11} \\ \epsilon_{21} \\ \epsilon_{31} \\ \epsilon_{41} \\ \epsilon_{12} \\ \epsilon_{22} \\ \epsilon_{32} \\ \epsilon_{42} \end{bmatrix}$$

The marginal distribution of  $Z_j(\mathbf{s}_i)$  is such that

$$Z_j(\mathbf{s}_i) \sim N(\theta_j + \omega_j X(\mathbf{s}_i)' \beta, \sigma_j^2 + \omega_j^2 \phi) \quad \text{for } i = 1, \dots, 4, \text{ and } j = 1, 2.$$

Since the lack of identifiability shown in Section 3.2.2.2 was due to  $\boldsymbol{\theta}$  and  $\eta$ , both parameters of the mean of  $\mathbf{Z}(\mathbf{s})$ , we will focus on the expected value of the marginal distribution of  $\mathbf{Z}(\mathbf{s})$ . We write the expected value of  $\mathbf{Z}(\mathbf{s})$  as

$$E \begin{pmatrix} Z_1(\mathbf{s}_1) \\ Z_1(\mathbf{s}_2) \\ Z_1(\mathbf{s}_3) \\ Z_1(\mathbf{s}_4) \\ Z_2(\mathbf{s}_1) \\ Z_2(\mathbf{s}_2) \\ Z_2(\mathbf{s}_3) \\ Z_2(\mathbf{s}_4) \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} + \begin{bmatrix} X(\mathbf{s}_1) & 0 \\ X(\mathbf{s}_2) & 0 \\ X(\mathbf{s}_3) & 0 \\ X(\mathbf{s}_4) & 0 \\ 0 & X(\mathbf{s}_1) \\ 0 & X(\mathbf{s}_2) \\ 0 & X(\mathbf{s}_3) \\ 0 & X(\mathbf{s}_4) \end{bmatrix} \begin{bmatrix} \beta \\ \beta\omega_2 \end{bmatrix}$$

This can be rewritten to look like a mixed effects model or a multiple regression model with interaction terms. In this case, the model would be

$$E \begin{pmatrix} Z_1(\mathbf{s}_1) \\ Z_1(\mathbf{s}_2) \\ Z_1(\mathbf{s}_3) \\ Z_1(\mathbf{s}_4) \\ Z_2(\mathbf{s}_1) \\ Z_2(\mathbf{s}_2) \\ Z_2(\mathbf{s}_3) \\ Z_2(\mathbf{s}_4) \end{pmatrix} = \begin{pmatrix} 1 & 0 & X(\mathbf{s}_1) & 0 \\ 1 & 0 & X(\mathbf{s}_2) & 0 \\ 1 & 0 & X(\mathbf{s}_3) & 0 \\ 1 & 0 & X(\mathbf{s}_4) & 0 \\ 0 & 1 & 0 & X(\mathbf{s}_1) \\ 0 & 1 & 0 & X(\mathbf{s}_2) \\ 0 & 1 & 0 & X(\mathbf{s}_3) \\ 0 & 1 & 0 & X(\mathbf{s}_4) \end{pmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \\ \beta \\ \beta\omega_2 \end{bmatrix} \quad (29)$$

Written this way, it is easy to see that the columns of the design matrix are linearly independent. Therefore, no column of the design matrix can be written as a linear combination of the other columns. In the model where the  $W(\mathbf{s}_i) \stackrel{iid}{\sim} N(\eta, \phi)$ , for  $i = 1, \dots, 4$ , the expected value of the marginal distribution of  $\mathbf{Z}(\mathbf{s})$  would be written as

$$E \begin{pmatrix} Z(\mathbf{s}_1)_1 \\ Z(\mathbf{s}_2)_1 \\ Z(\mathbf{s}_3)_1 \\ Z(\mathbf{s}_4)_1 \\ Z(\mathbf{s}_1)_2 \\ Z(\mathbf{s}_2)_2 \\ Z(\mathbf{s}_3)_2 \\ Z(\mathbf{s}_4)_2 \end{pmatrix} = \begin{bmatrix} 1 & 0 & \eta & 0 \\ 0 & 1 & 0 & \eta \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \\ 1 \\ \omega_2 \end{bmatrix} \quad (30)$$

It is clear that not all  $\eta$ ,  $\theta_1$ ,  $\theta_2$ , and  $\omega_2$  are identifiable since the design matrix in (30) is not full rank. Therefore, (29) shows that the parameters  $\theta_1, \theta_2, \beta$ , and  $\omega_2$  are identifiable when  $\mathbf{W}(\mathbf{s}) \sim N(\mathbf{X}(\mathbf{s})\boldsymbol{\beta}, \phi\mathbf{I})$ . This easily extends to our model with multivariate ordinal response data where  $i = 1, \dots, n$  and  $j = 1, \dots, J$ .

### 3.3 Identifiability of spatial random effect models

In this section, we investigate parameter identifiability in spatial probit linear mixed models (PLMMs) for binary and ordinal response data. We focus on continuous spatial random effects of the form  $\mathbf{W}(\mathbf{s}) \sim GP(\mathbf{X}(\mathbf{s})\boldsymbol{\beta}, \Sigma_W)$  where  $\Sigma_W$  is a spatial covariance matrix. Our analysis is based on the asymptotic theory of parameter identifiability in GLMMs by Zhang (2004) and the empirical results of linear mixed models (LMMs) for continuous data with spatially correlated errors by Irvine et al. (2007). We compare the signal in the likelihood function of PLMMs for the spatial parameters by simulating binary, ordinal, and continuous response data for different parameter values for the spatial covariance function. We relate our results of identifiability in spatial PLMMs to spatial GLMMs.

### 3.3.1 Asymptotics for geostatistical models

There exists a long history and extensive literature of spatial modeling for lattice data using a Markov random field specification introduced by Besag (1974). Markov random fields are well suited for both likelihood and Bayesian inference, particularly simulation based inference such as Gibbs sampling. A disadvantage of these models is that the off-diagonal entries of the precision matrix measuring correlation between pairs is defined through a neighborhood structure and must be pre-specified. Also, Markov random fields have no implied spatial process on the domain (Gelfand et al., 2000). In contrast, geostatistics is a field of statistics that models spatial variation over a continuous spatial region. In some applications, continuous spatial processes may be more appropriate when the main objective is spatial interpolation. Continuous spatial fields have their disadvantages as well. Their biggest drawback is that both likelihood or Bayesian inference require matrix inversion for an  $n \times n$  matrix where  $n$  is the sample size to evaluate the density of the spatial random field. This inversion can be slow, especially as the the number of locations in the sample increases. Approximation and rank-reducing methods have drawn much attention recently as ways of avoiding matrix inversion. Integrated nested Laplacian approximations (INLA) is one approach for approximating Bayesian posteriors and Gaussian random fields (Rue et al., 2009). Other methods include fixed-rank kriging (Cressie and Johannesson, 2008), covariance tapering (Furrer et al., 2006), and Gaussian predictive process (Banerjee et al., 2008).

We begin by discussing relevant work in the area of parameter estimation of geostatistical spatial models. The majority of the literature in geostatistical spatial models pertains to continuous response data, such as Gaussian response data, or count data, such as that generated from a Poisson model. Zhang (2004) investigates the consistency of estimators in model-based geostatistics, focusing on spatial GLMMs. He defines the spatial GLMM as follows:

1. Let  $\{W(\mathbf{s}), \mathbf{s} \in \mathbb{R}^d\}$  be a second-order stationary Gaussian process with mean  $\mathbf{0}$ .
2. Conditional on  $\{W(\mathbf{s}), \mathbf{s} \in \mathbb{R}^d\}$ , the observable random variables  $\{Y(\mathbf{s}), \mathbf{s} \in \mathbb{R}^d\}$  are mutually independent.
3. Assuming  $Y(\mathbf{s})$  follows a GLM with distribution specified by the conditional mean  $\mu(\mathbf{s}) = E(Y(\mathbf{s})|W(\mathbf{s}))$ ,  $g(\mu(\mathbf{s})) = X(\mathbf{s})\boldsymbol{\beta} + W(\mathbf{s})$  for some link function  $g$ , covariate vector  $\mathbf{X}(\mathbf{s})$  and coefficient vector  $\boldsymbol{\beta}$ .

Zhang (2004) derives results for the Matérn class of covariance functions of which exponential is a special case. The exponential covariogram is defined as

$$\text{Cov}(W(\mathbf{s}_i), W(\mathbf{s}_j)) = \tau \exp^{-\frac{1}{\phi}d_{ij}} \quad (31)$$

where  $d_{ij}$  is the distance between locations  $\mathbf{s}_i$  and  $\mathbf{s}_j$ . Zhang (2004) uses properties of equivalence of probability measures to show that while neither parameter  $\tau$  or  $\phi$  in a spatial GLMM with exponential covariance is consistently estimable, the quantity  $\tau/\phi$  is consistently estimable. Stein (1990) showed that predictions obtained from an incorrect covariance function are asymptotically optimal if the incorrect covariance function is compatible with the correct covariance function. This means that the difference between predictions obtained from an incorrect exponential covariance function and predictions obtained using the correct exponential covariance function goes to 0 asymptotically for compatible covariance functions. Two covariance functions are compatible if the probability measures of the two processes have equal means functions and the covariance functions are mutually absolutely continuous (Stein, 1988). Further, stationary covariance functions are compatible if they behave similarly at the origin (Stein, 1988).

It is worth mentioning that spatial asymptotics are classified into two groups: increasing-domain asymptotics and fixed-domain, or infill, asymptotics. Increasing-domain is such that the domain of the spatial field increases as the amount of data increases. Fixed-domain

refers to spatial fields with fixed boundaries such that increases in data results in higher density sampling. Consistent estimators of  $\tau$  and  $\phi$  are only available under increasing-domain asymptotics (Mardia and Marshall, 1984). Using simulation, Zhang (2004) shows that predicted values and prediction variances at unobserved locations of a binary response variable modeled with an exponential covariance function are nearly identical when the ratio  $\tau/\phi$  is the same. Further, for different values of the ratio  $\tau/\phi$ , similar predicted values are produced but the prediction variances vary. This indicates that when interpolation is the objective, the ratio  $\tau/\phi$  matters more than the individual parameters.

Zhang's theoretical results assumed a GLMM with no covariates. Irvine et al. (2007) investigate empirical behavior of estimates of  $\tau$  and  $\phi$  for LMMs with covariates, a continuous response, and spatially correlated errors. They compare the strength of spatial correlation on maximum likelihood (ML) and restricted maximum likelihood (REML) estimates of  $\tau$  and  $\phi$  of the exponential spatial covariance model plus nugget for different sample sizes, sampling designs, and nugget-to-sill ratios. For all combinations of sample size ( $n = 144$  to  $361$ ), design, and ratio, they conclude that ML and REML give reasonable estimates when  $\phi = 1$  for a  $10 \times 10$  spatial domain. Further, they also show that the variance in the estimates decreases as the sample size increases. When  $\phi = 3$ , corresponding to larger effective range values, ML tends to underestimate the spatial autocorrelation function and the variability of the estimates is large. The variability, in this case, does not tend to decrease with larger sample sizes. They also note a positive association between the parameters  $\tau$  and  $\phi$ , which is not surprising given the consistency results of Zhang (2004). Irvine et al. (2007) and others have shown that cluster sample designs where the clusters are distributed across the region of interest are optimal for covariance parameter estimation under the exponential-with-nugget model.

### 3.3.2 Simulations for PLMMs

The theory of Zhang (2004) is for GLMMs with no covariates and the empirical results of Irvine et al. (2007) were for continuous response data. We investigate the behavior for spatial probit latent variable models for discrete response data, or PLMMs. Recall that for binary response data, a PLMM is a GLMM. However, for ordinal data, a PLMM is not a GLMM (see Chapter 1). We evaluate the estimability of the covariance parameters in the exponential covariogram (31) for spatial PLMMs with binary and ordinal response data through simulation. The spatial PLMM can be defined through a Gaussian latent variable  $\mathbf{Z}(\mathbf{s})$  such that

$$\mathbf{Z}(\mathbf{s}) \sim N(\mathbf{X}(\mathbf{s})\boldsymbol{\beta} + \mathbf{W}(\mathbf{s}), \sigma^2 \mathbf{I}) \quad (32)$$

where  $\mathbf{W}(\mathbf{s})$  is a spatial random effect. This will be called a second-stage spatial model as defined in Chapter 4.1. We simulate  $n = 300$  locations, 200 of which are simulated via Poisson cluster sampling and an additional 100 locations using a lattice design on the unit square (see Figure 3.1). We note that the results might vary for other spatial designs. For simplicity, we do not include any fixed effects in the model. We simulate  $\mathbf{W}(\mathbf{s})$  from a mean-zero Gaussian spatial process with exponential covariance in (31). Using the data augmentation strategies of Albert and Chib (1993), we draw

$$\mathbf{Z}(\mathbf{s}) \sim N(\mathbf{W}(\mathbf{s}), \sigma^2 \mathbf{I}). \quad (33)$$

For identifiability in the probit regression model,  $\sigma^2$  is fixed to 1. Given  $\mathbf{Z}(\mathbf{s})$ , the binary response  $\mathbf{Y}(\mathbf{s})$  is equal to 1 for values of  $\mathbf{Z}(\mathbf{s}) > 0$  and 0 for values of  $\mathbf{Z}(\mathbf{s}) \leq 0$ . For ordinal response data,  $\mathbf{Z}(\mathbf{s})$  is thresholded according to  $\boldsymbol{\lambda}$ , which we assume is fixed and known. Figure 3.2 shows a simulated response field,  $\mathbf{Y}(\mathbf{s})$ , for binary and ordinal data, where the ordinal response is assumed to have five categories. Both regions show spatial correlation in the response variable where locations at close proximity are similar.

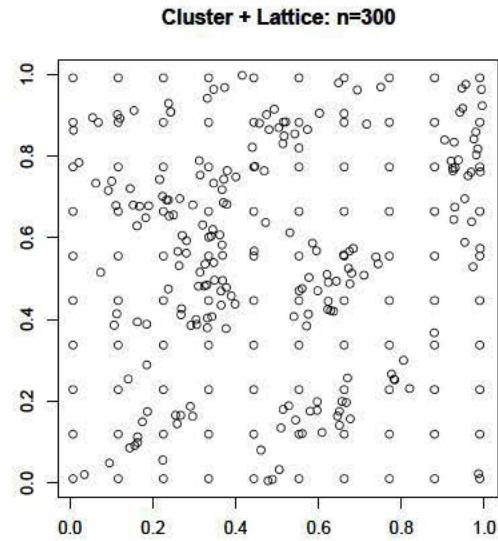


Figure 3.1: Simulated locations via Poisson cluster sampling and lattice design.

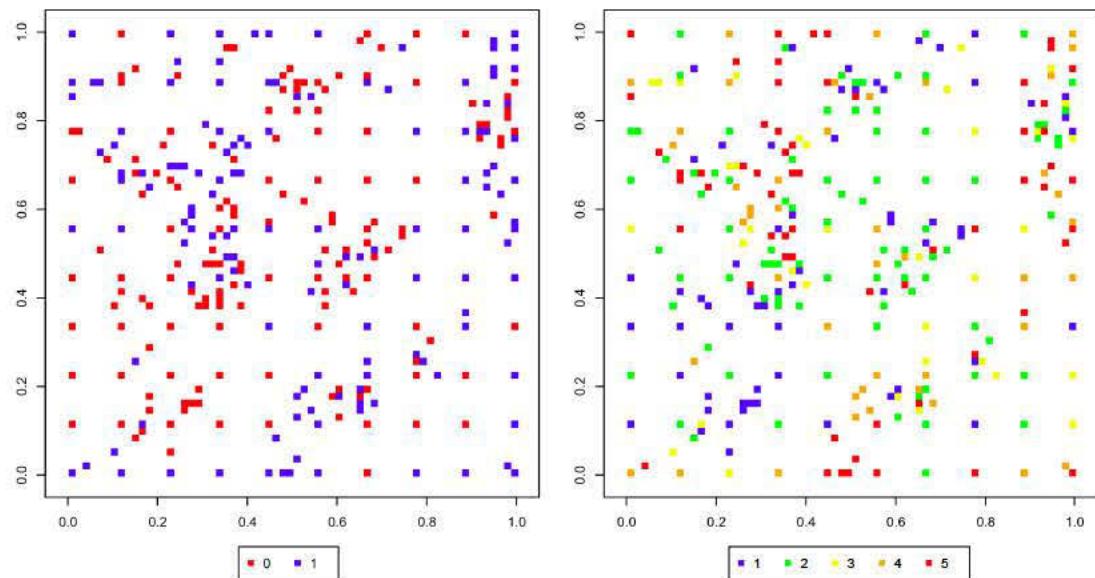


Figure 3.2: Simulated binary (left) and ordinal (right) spatial response fields of  $\mathbf{Y}(s)$  at  $n = 300$  simulated locations.

Table 3.1: Simulation values for exponential covariogram parameters,  $\tau$  and  $\phi$ , for comparison.

	$\tau = 0.25$	$\tau = 1$	$\tau = 2$	$\tau = 4$
$\phi = 0.1$	A A	A B A		A D
$\phi = 0.2$	B A	B B B	A	B D
$\phi = 0.4$	C A	C B C	B	C D A
$\phi = 0.8$			C	D B
$\phi = 1.6$				D C

The objective of our simulation study is to determine how well binary and ordinal data can estimate the parameters of the exponential covariogram. We simulate data for different values of  $\tau$  and  $\phi$  in (31), as well as different ratios of  $\tau/\phi$  (Table 3.1). We aim to address the effects of the spatial parameters on the likelihood for the following scenarios:

1. Changes in the partial-sill parameter,  $\tau$ , for a fixed range,  $\phi$  (red in Table 3.1).
2. Changes in the range parameter,  $\phi$ , for a fixed partial-sill parameter,  $\tau$  (green).
3. Changes in both partial-sill parameter,  $\tau$ , and range parameter,  $\phi$ , for a fixed ratio,  $\tau/\phi$  (blue).

Large values of  $\phi$  imply the field has long-range spatial correlation. The effective range is the distance beyond which the correlation between observations is less than or equal to 0.05. For the exponential covariance function without nugget, the effective range is approximately  $3\phi$ . The effective range for the exponential-with-nugget covariance function is

$$-\phi \log \left( 0.05 \frac{\tau + \sigma^2}{\sigma^2} \right).$$

The nugget is defined as the amount of variance that is not explained or modeled as spatial correlation. Therefore, the PLMM assumes a nugget of 1 by fixing  $\sigma^2 = 1$ . Figure (3.3) shows the true autocorrelogram for the exponential covariance function for parameter values of  $\tau$  and  $\phi$  of interest. Larger values of  $\tau$  correspond to smaller nugget-to-sill ratios and thus

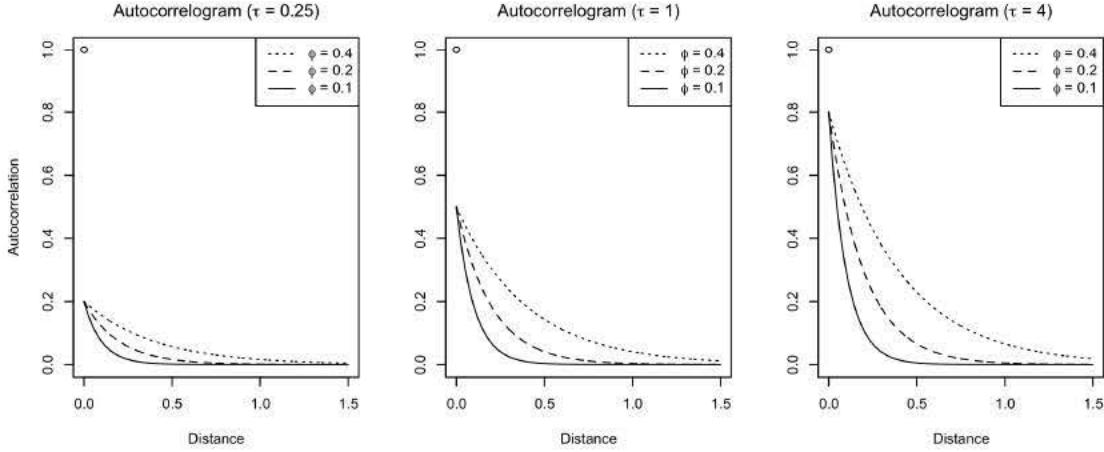


Figure 3.3: Autocorrelograms for the exponential covariance function for  $\tau = 0.25$  (left), 1 (middle), and 4 (right) and  $\phi = 0.1, 0.2$ , and  $0.4$ .

larger spatial signals. The rate of decay of spatial correlation decreases as the range parameter,  $\phi$ , increases. Figure (3.4) gives the empirical autocorrelogram for binary, ordinal, and continuous response data simulated when  $\tau = 1$  and  $\phi = 0.2$ . The empirical autocorrelation values suggest estimates of both  $\tau$  and  $\phi$  below their true values for all of the data types.

We plot the log-likelihood surfaces for the binary, ordinal, and continuous response models for different values of  $\tau$  and  $\phi$ . Note that the plot axis in Figures (3.5) - (3.14) differ. This is because the empirical likelihood is computed on a grid of parameter values for  $\tau$  and  $\phi$  and we wanted the resolution of the grid to scale with  $\tau$  and  $\phi$ . Figure 3.5 shows the log-likelihood surface for each data type when  $\tau = 0.25$  and  $\phi$  is 0.1 (top), 0.2 (middle), and 0.4 (bottom). The same plots are shown in Figures 3.6 and 3.7, where  $\tau = 1$  and  $\tau = 4$ , respectively. For all data types and values of  $\tau$  and  $\phi$ , the log-likelihood surface reaches its maximum at locations that underestimate the true value of  $\tau$  and  $\phi$  indicated by  $\blacktriangle$  on the plot. This agrees with the findings of Irvine et al. (2007) for continuous data. In general, the mode of the log-likelihood surface is elliptical with major axis having a positive slope. This is not surprising since Zhang (2004) showed that only the ratio  $\tau/\phi$  can be consistently estimated. It is less noticeable in Figure 3.5 where  $\tau = 0.25$  because the spatial signal is

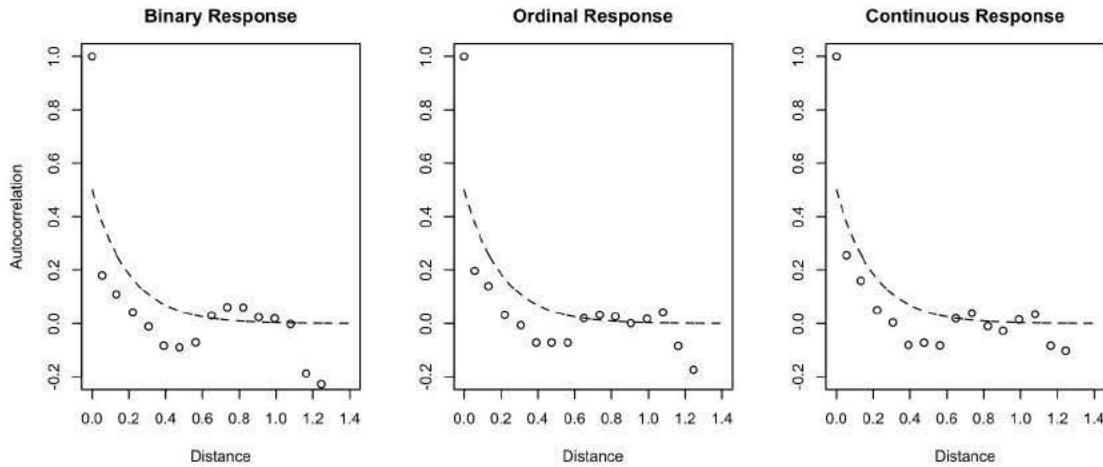


Figure 3.4: Empirical autocorrelograms for the exponential covariance function for  $\tau = 1$  and  $\phi = 0.2$  for binary, ordinal, and continuous response data. The dashed line on the plots indicate the true autocorrelation for the parameter values used in the simulation.

weak. Comparing across Figures 3.5, 3.6, and 3.7, we see that as  $\tau$  increases for fixed  $\phi$ , the slope of the ridge increases. Looking at each figure separately,  $\tau$  is fixed and we see that as  $\phi$  increases the slope of the ridge decreases. The modal region of the surface becomes more localized as the spatial signal increases and the spatial range decreases (i.e.,  $\tau$  increases and  $\phi$  decreases).

The log-likelihood values vary greatly between the different types of data. To compare across types, we create a log-likelihood ratio surface by dividing each value on the surface by its maximum log-likelihood value. Values of 1 on the log-likelihood ratio surface indicate the maximum log-likelihood. Larger values on the ratio surface indicate less optimal parameter values. Figure 3.8 gives the log-likelihood ratio surface for  $\tau = 0.25$ . Each data type underestimates both  $\tau$  and  $\phi$  while the underestimation is greater when  $\phi = 0.4$  than when  $\phi = 0.1$ . Define data richness as the amount of information in the data, where binary data is the least rich and continuous data is the most rich. The contour lines indicate that the continuous response ratio surface is steeper than both the ordinal and binary response surfaces at similar parameter values. This indicates that the signal in the response likelihood for  $\tau$  and  $\phi$  increases with data richness. When the spatial signal increases (i.e.,  $\tau$  increases

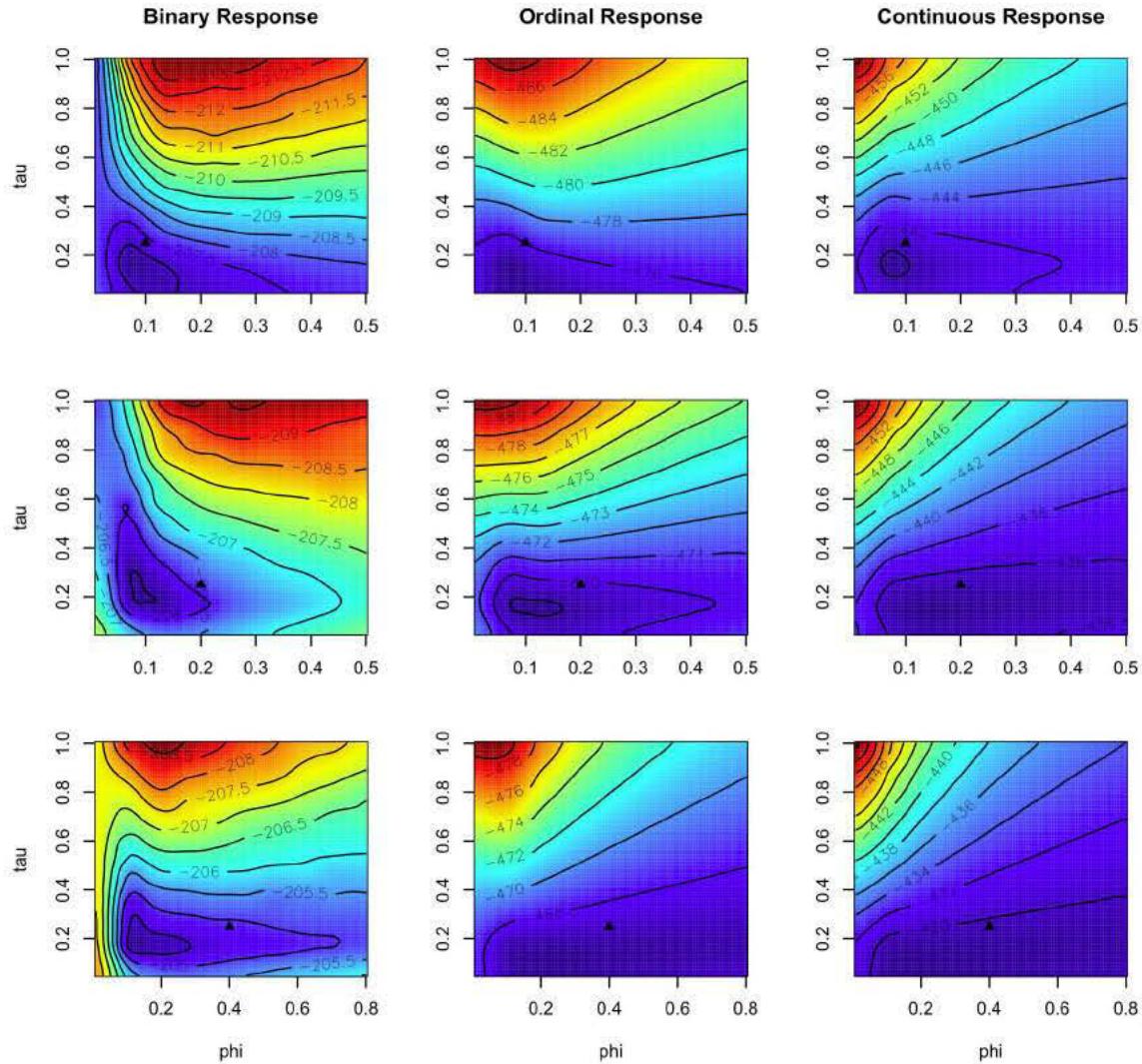


Figure 3.5: Log-likelihood surfaces for binary (left), ordinal (middle), continuous (right) response data for  $\tau = 0.25$  and  $\phi = 0.1$  (top), 0.2 (middle), and 0.4 (bottom) where  $\blacktriangle$  denotes the true values of  $\tau$  and  $\phi$ .

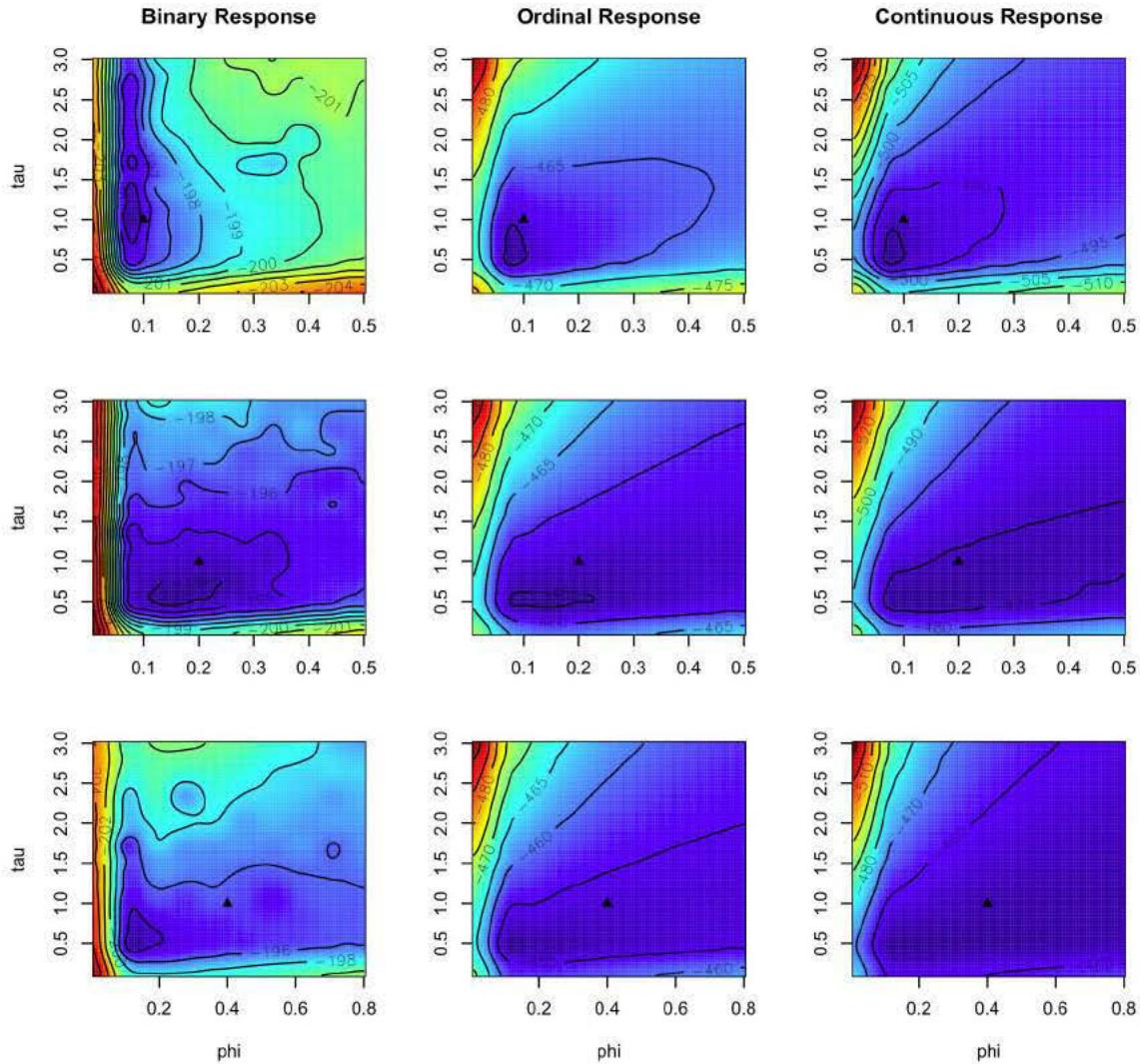


Figure 3.6: Log-likelihood surfaces for binary (left), ordinal (middle), continuous (right) response data for  $\tau = 1$  and  $\phi = 0.1$  (top),  $0.2$  (middle), and  $0.4$  (bottom) where  $\blacktriangle$  denotes the true values of  $\tau$  and  $\phi$ .

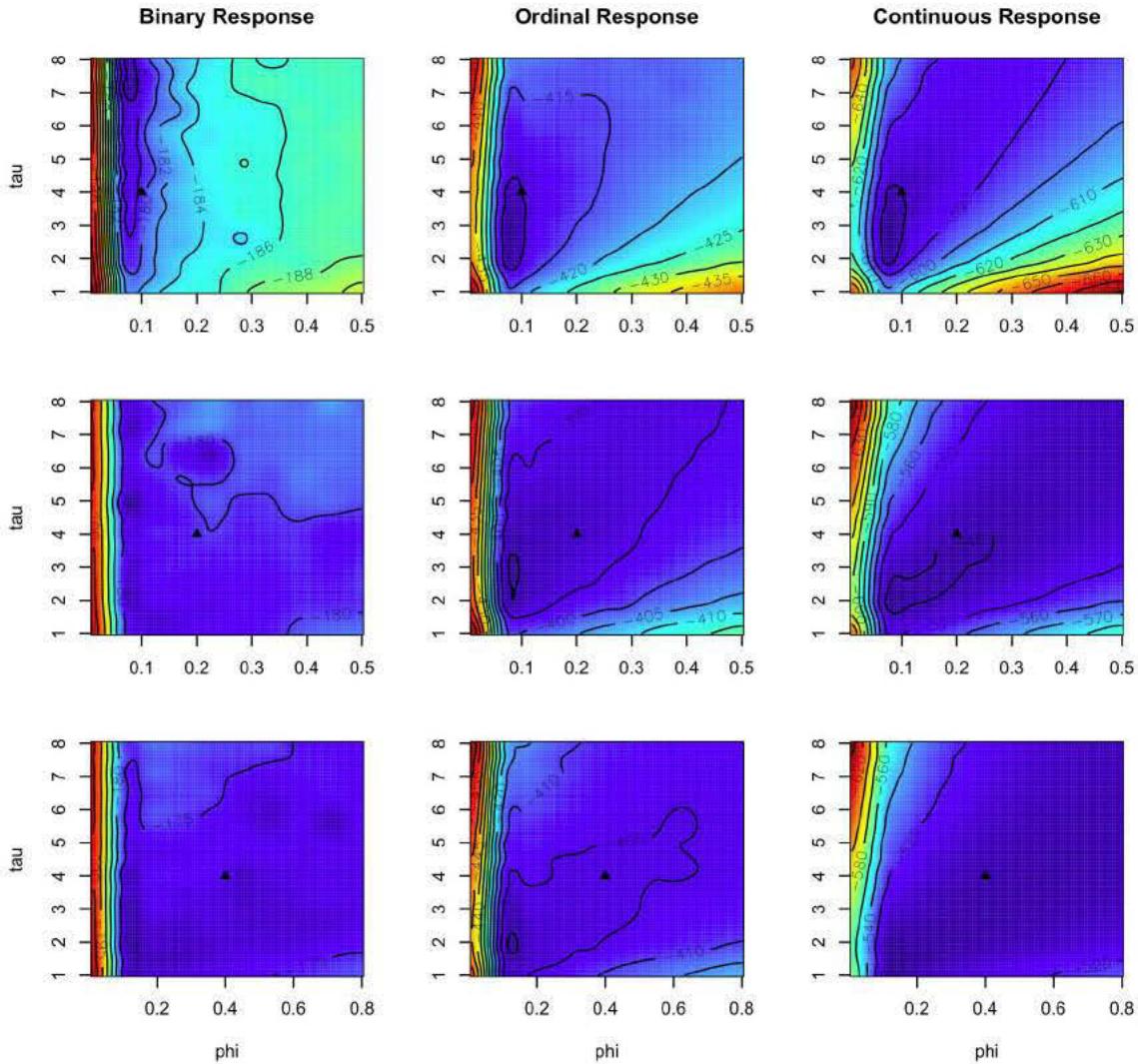


Figure 3.7: Log-likelihood surfaces for binary (left), ordinal (middle), continuous (right) response data for  $\tau = 4$  and  $\phi = 0.1$  (top), 0.2 (middle), and 0.4 (bottom) where  $\blacktriangle$  denotes the true values of  $\tau$  and  $\phi$ .

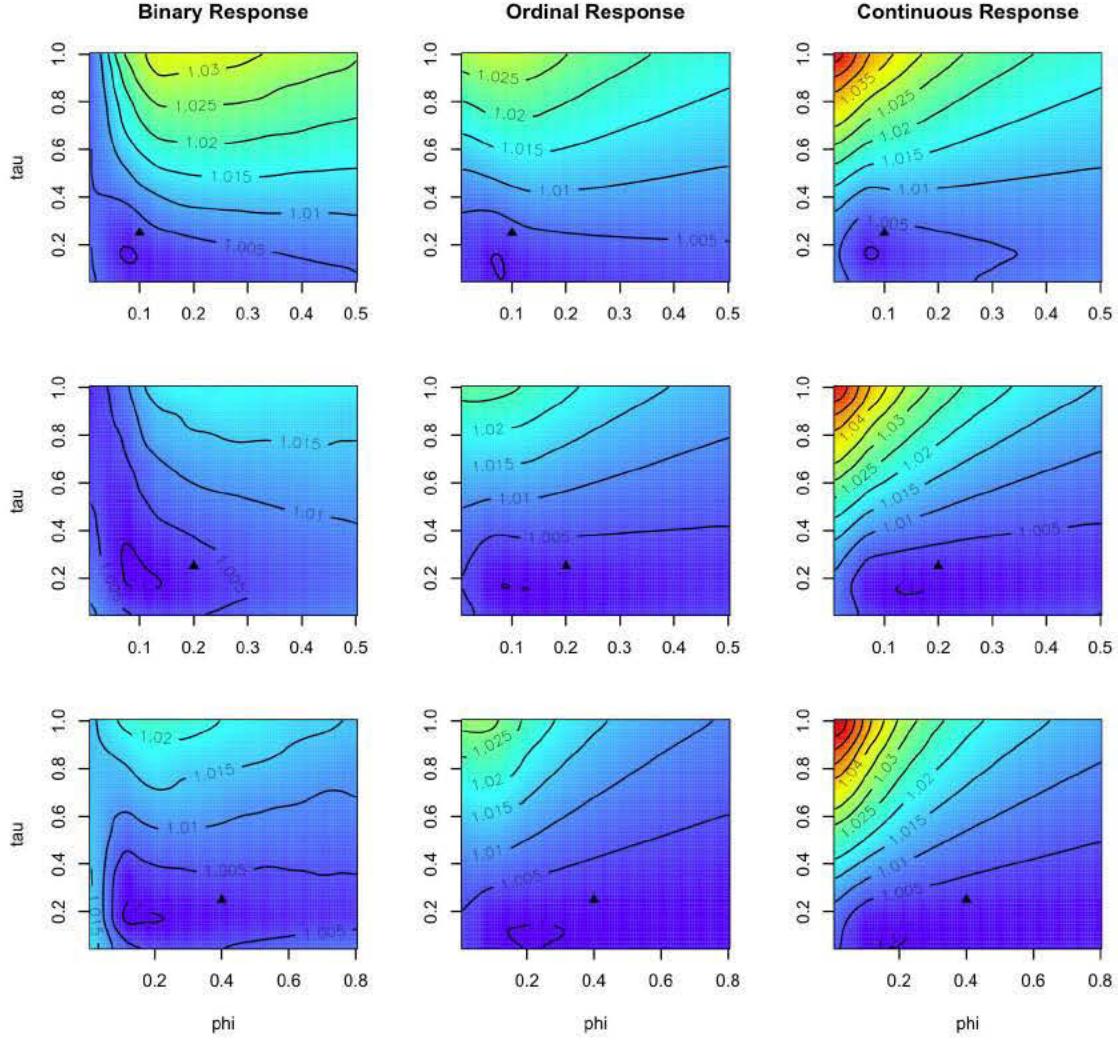


Figure 3.8: Log-likelihood ratio surfaces for binary (left), ordinal (middle), continuous (right) response data for  $\tau = 0.25$  and  $\phi = 0.1$  (top), 0.2 (middle), and 0.4 (bottom). Ratio computed as the log-likelihood divided by the maximum log-likelihood within each field.

for a fixed  $\phi$ ), the difference between response type becomes more clear. Figure 3.9 shows the log-likelihood ratio surface when  $\tau = 1$ . Again, the continuous response surface is the steepest. The ordinal response surface with  $k = 5$  categories is very similar to the continuous response surface. It is interesting to note that as the richness in the response data increases, the maximized log-likelihood ratio region becomes less localized, specifically for the range parameter  $\phi$ . Similar patterns are seen in Figure 3.10 when  $\tau = 4$ .

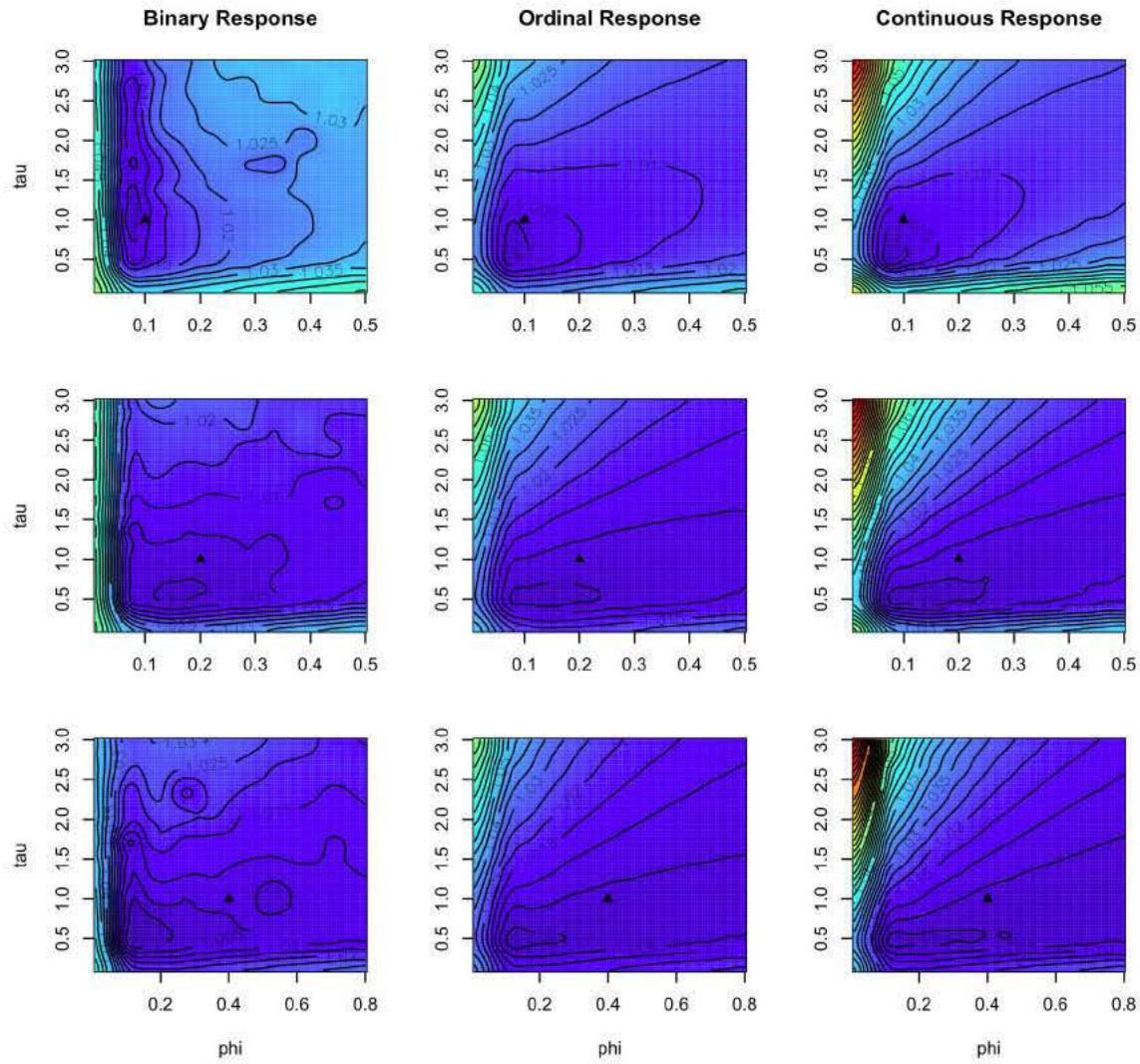


Figure 3.9: Log-likelihood ratio surfaces for binary (left), ordinal (middle), continuous (right) response data for  $\tau = 1$  and  $\phi = 0.1$  (top), 0.2 (middle), and 0.4 (bottom) where  $\blacktriangle$  denotes the true values of  $\tau$  and  $\phi$ . Ratio computed as the log-likelihood divided by the maximum log-likelihood within each field.

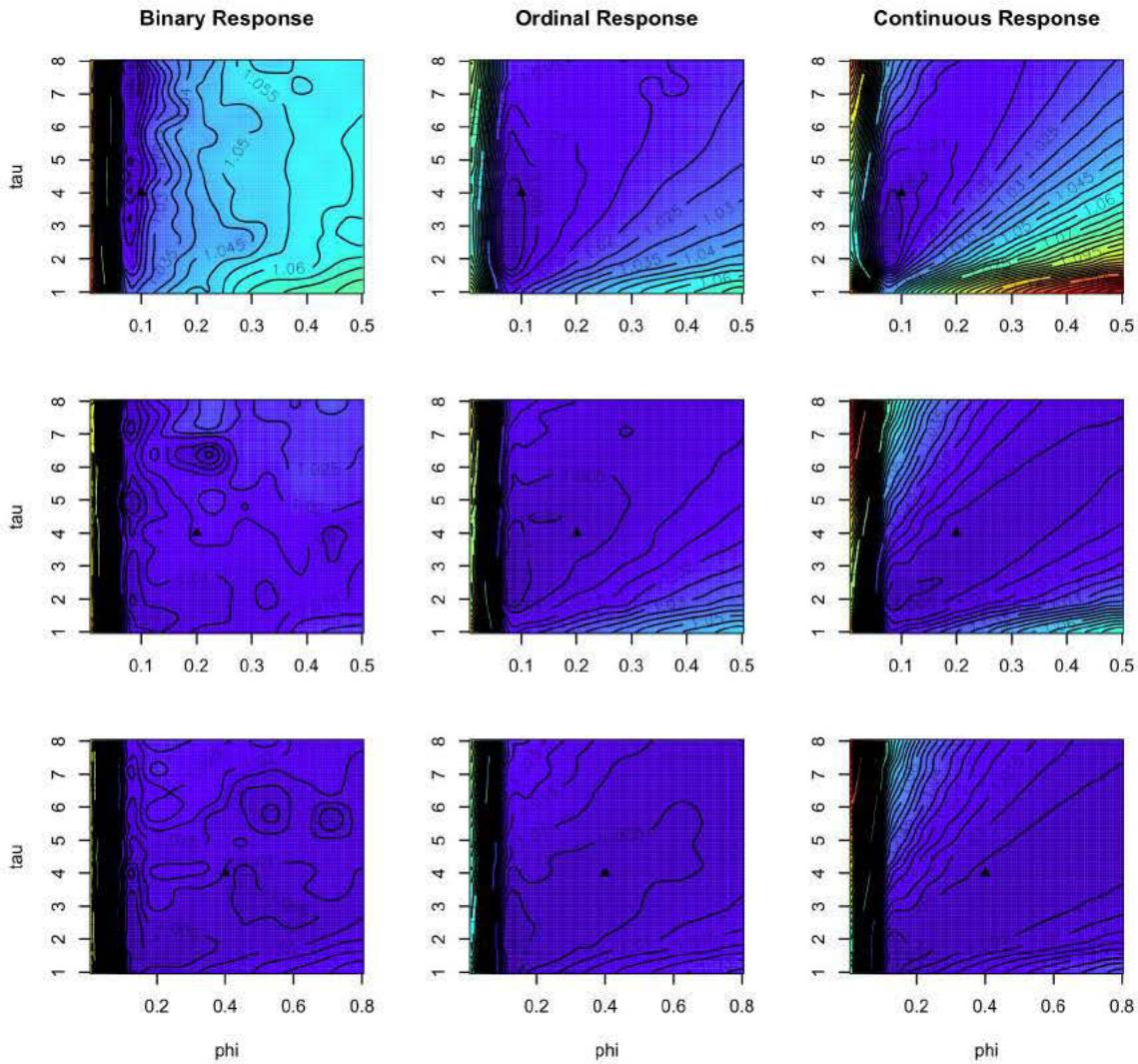


Figure 3.10: Log-likelihood ratio surfaces for binary (left), ordinal (middle), continuous (right) response data for  $\tau = 4$  and  $\phi = 0.1$  (top),  $0.2$  (middle), and  $0.4$  (bottom) where  $\blacktriangle$  denotes the true values of  $\tau$  and  $\phi$ . Ratio computed as the log-likelihood divided by the maximum log-likelihood within each field.

The theory of consistent estimation of the spatial parameters by Zhang (2004) indicated that similar patterns may be observed in the likelihood surfaces for different values of  $\tau$  and  $\phi$  when the ratio of  $\tau/\phi$  is the same. Therefore, we simulated data with  $\tau = 2$  and  $\phi = 0.2$ , 0.4, and 0.8 as well as  $\tau = 4$ , and  $\phi = 0.4$ , 0.8, and 1.6. The ratios of  $\tau/\phi$  are the same as those shown in Figures 3.6 and 3.9, where  $\tau/\phi = 10, 5$ , and 2.5. Figures 3.11 and 3.12 give the log-likelihood surfaces and Figures 3.13 and 3.14 give the log-likelihood ratio surfaces for the scaled-parameter simulations. Similar surfaces are produced when  $\tau$  and  $\phi$  are scaled such that  $\tau/\phi$  is the same. As the scaling factor increases, the size of the modal region of the log-likelihood surface also scales by extending further along the  $\tau/\phi$  ridge. Once again, similar patterns appear in the ordinal and continuous response surfaces. The localization of the modal region of the log-likelihood surfaces also appear to scale with  $\tau$  and  $\phi$ . The maximum regions are larger in Figure 3.13 and 3.14 than in Figure 3.9 for all data types. We believe this to predominantly be a feature of the large range values since the effective range of these surfaces is larger than the maximum distance between locations. Therefore, the large scale spatial correlation cannot be captured within a  $1 \times 1$  unit square for values of  $\phi \geq 0.4$ .

In summary, the signal of the spatial parameters,  $\tau$  and  $\phi$ , in the likelihood of PLMMs is similar to that found in the likelihood of GLMMs and LMMs. Both binary and ordinal data tend to underestimate the spatial parameters while preserving the ratio of the consistently estimable derived parameter,  $\tau/\phi$  (Figures 3.5 - 3.10). This is indicative of the positively correlated modal ridge in each log-likelihood surface plot. The log-likelihood surfaces tend to underestimate the parameters less at lower values of  $\phi$  for fixed values of  $\tau$  (comparison across green letters in Table 3.1). This is seen in each Figure (3.5 - 3.10) as you compare within data type. The surfaces also indicate that the signal in the spatial parameters is greater for larger values of  $\tau$  for fixed values of  $\phi$  (comparison across red letters in Table 3.1). The log-likelihood surfaces are very similar for different values of  $\tau$  and  $\phi$  when the ratio,  $\tau/\phi$  is preserved (comparison across blue letters in Table 3.1), matching the asymptotic results of

Zhang (2004) (Figures 3.9, 3.6, 3.11 - 3.14). As expected, the amount of signal in the response likelihood is positively correlated with the richness in the data. Ordinal response data with 5 categories is richer than binary response data. Further, the log-likelihood surfaces produced by ordinal response data are similar to those produced by continuous response data. Lastly, the difference in signal of the spatial parameters between the data types is more apparent when the spatial signal increases relative to the nugget (Figures 3.8 - 3.10).

### 3.4 Fitting spatial models using MCMC

The likelihood surfaces shown in Section 3.3 indicate that there is weak parameter identifiability for the spatial parameters in spatial PLMMs. Recall that weak identifiability means that  $p_{\boldsymbol{\theta}}(y) \approx p_{\boldsymbol{\theta}'}(y)$  for all  $y$  and  $\boldsymbol{\theta} \neq \boldsymbol{\theta}'$ . The PLMM for binary response data, for example, has likelihood function

$$L(\tau, \phi, \boldsymbol{\beta}; \mathbf{y}(\mathbf{s})) = \int_{\lambda_{y(\mathbf{s}_1)}}^{\lambda_{y(\mathbf{s}_1)}+1} \cdots \int_{\lambda_{y(\mathbf{s}_n)}}^{\lambda_{y(\mathbf{s}_n)}+1} \left( \frac{1}{2\pi} \right)^{n/2} |\Sigma_W + \mathbf{I}|^{-1/2} \times \exp \left\{ -\frac{1}{2} (\mathbf{Z}(\mathbf{s}) - \mathbf{X}(\mathbf{s})\boldsymbol{\beta})' (\Sigma_W + \mathbf{I})^{-1} (\mathbf{Z}(\mathbf{s}) - \mathbf{X}(\mathbf{s})\boldsymbol{\beta}) \right\} d\mathbf{Z}, \quad (34)$$

where  $\mathbf{y}(\mathbf{s})$  are the observable binary data,  $\boldsymbol{\lambda} = (\lambda_0, \lambda_1, \lambda_2) = (-\infty, 0, \infty)$ , and  $\Sigma_W$  is a spatial covariance matrix defined by (31). Consider the conditional posterior distributions of each spatial parameter,  $\tau$  and  $\phi$ , as well as the coefficient vector  $\boldsymbol{\beta}$ . For example,

$$\begin{aligned} p(\phi | \mathbf{y}(\mathbf{s}), \tau, \boldsymbol{\beta}) &\propto L(\mathbf{y}(\mathbf{s}); \tau, \phi, \boldsymbol{\beta}) p(\phi, \tau, \boldsymbol{\beta}) \\ &\propto L(\tau, \phi, \boldsymbol{\beta}; \mathbf{y}(\mathbf{s})) p(\phi), \end{aligned}$$

where the second line holds assuming  $\phi$ ,  $\tau$ , and  $\boldsymbol{\beta}$  are independent a priori. As seen in Section 3.3, there exists  $(\tau', \phi') \neq (\tau, \phi)$  such that

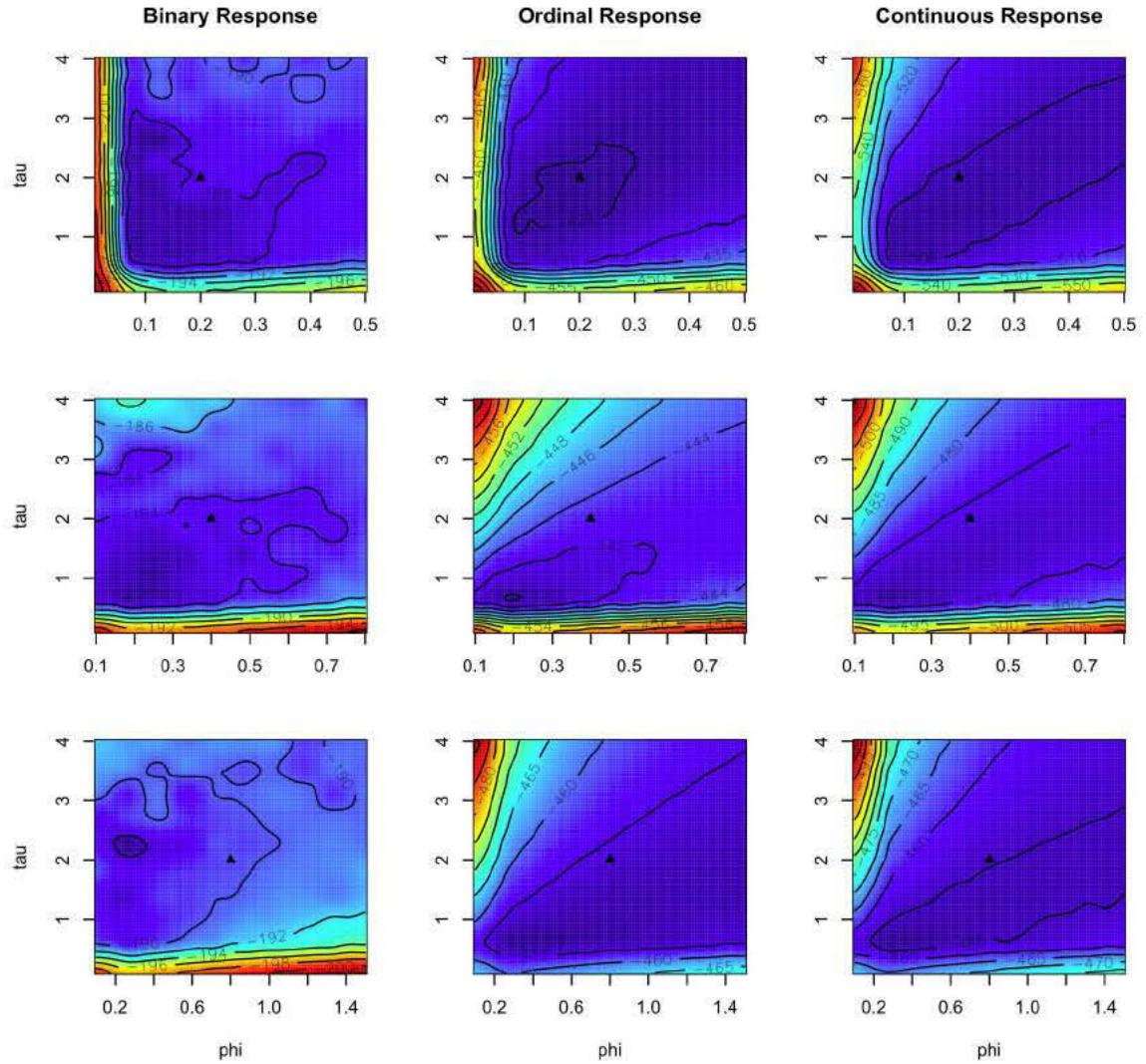


Figure 3.11: Log-likelihood surfaces for binary (left), ordinal (middle), continuous (right) response data for  $\tau = 2$  and the ratio  $\tau/\phi = 10$  (top), 5 (middle), and 2.5 (bottom) where  $\blacktriangle$  denotes the true values of  $\tau$  and  $\phi$ .

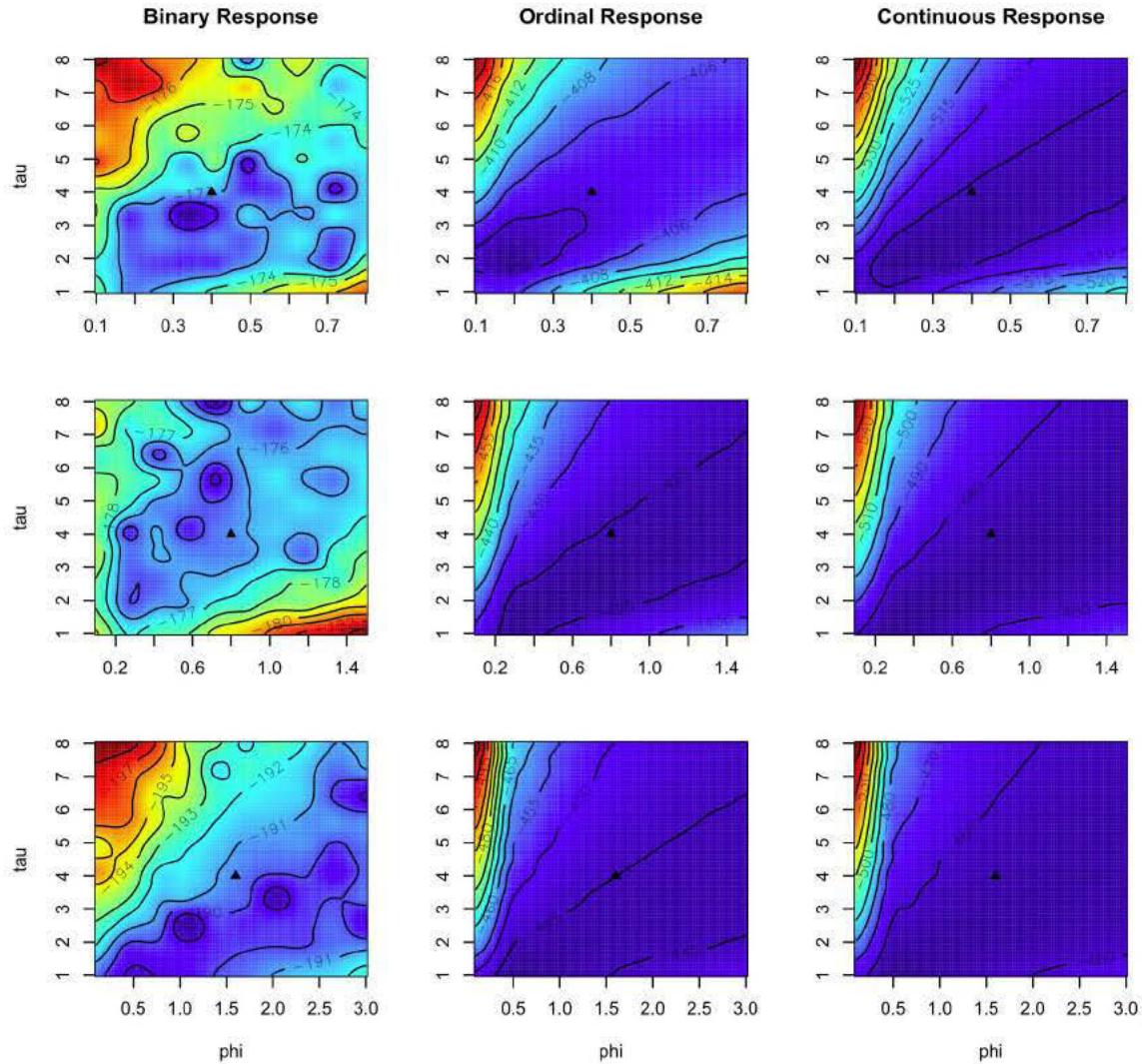


Figure 3.12: Log-likelihood surfaces for binary (left), ordinal (middle), continuous (right) response data for  $\tau = 4$  and the ratio  $\tau/\phi = 10$  (top), 5 (middle), and 2.5 (bottom) where  $\blacktriangle$  denotes the true values of  $\tau$  and  $\phi$ .

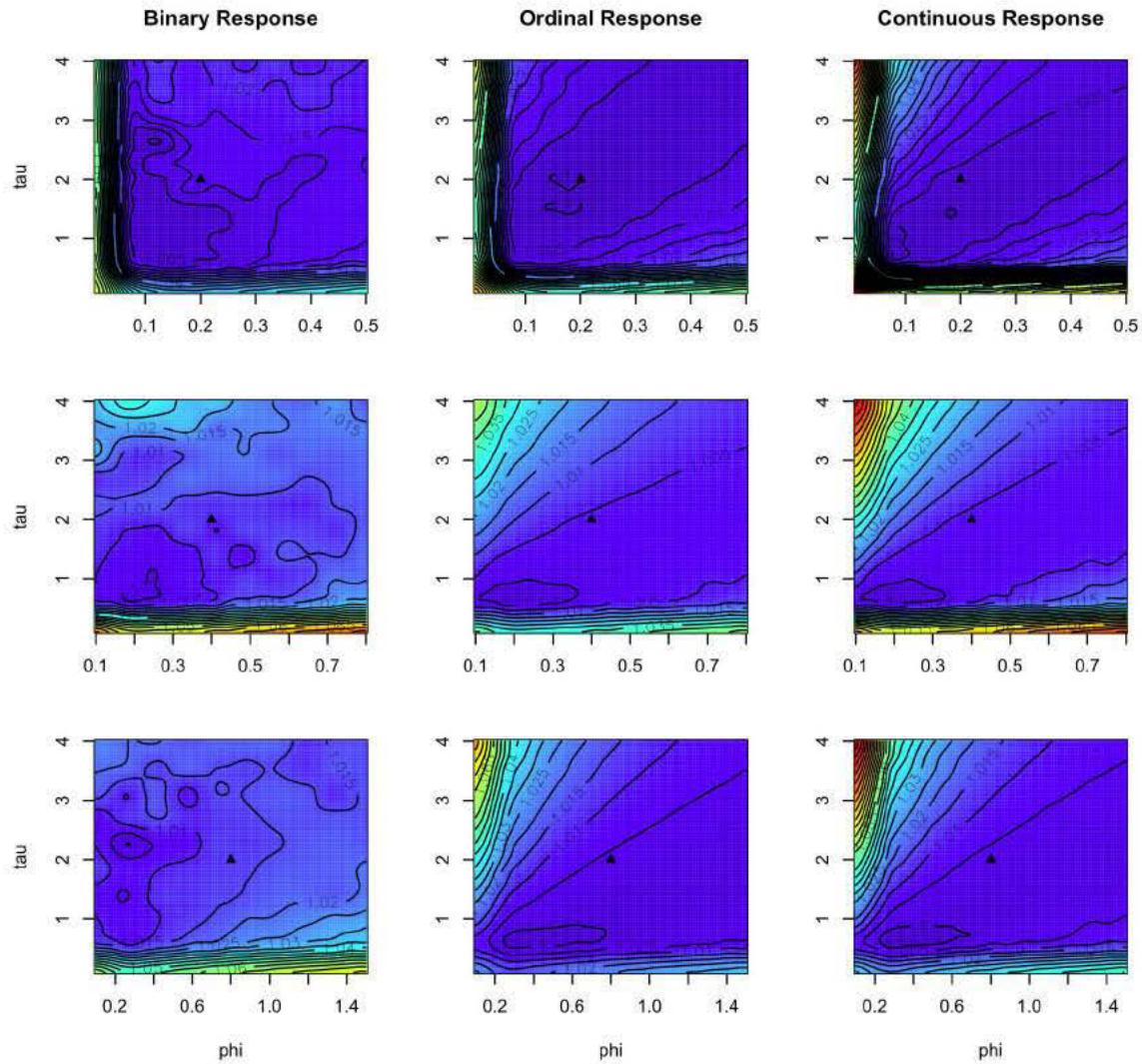


Figure 3.13: Log-likelihood ratio surfaces for binary (left), ordinal (middle), continuous (right) response data when  $\tau = 2$  and the ratio  $\tau/\phi = 10$  (top), 5 (middle), and 2.5 (bottom) where  $\blacktriangle$  denotes the true values of  $\tau$  and  $\phi$ . Ratio computed as the log-likelihood divided by the maximum log-likelihood within each field.

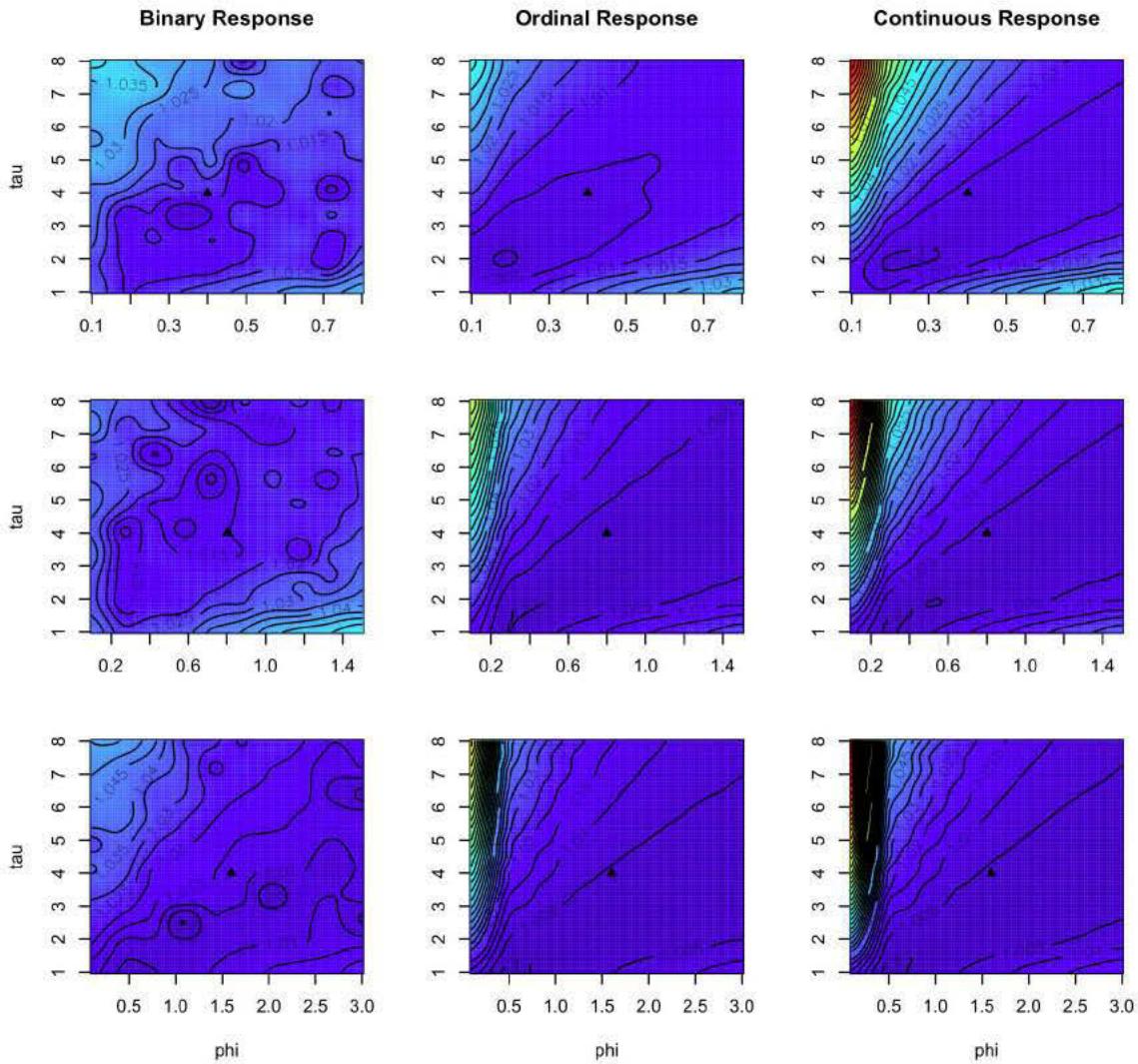


Figure 3.14: Log-likelihood ratio surfaces for binary (left), ordinal (middle), continuous (right) response data when  $\tau = 4$  and the ratio  $\tau/\phi = 10$  (top), 5 (middle), and 2.5 (bottom) where  $\blacktriangle$  denotes the true values of  $\tau$  and  $\phi$ . Ratio computed as the log-likelihood divided by the maximum log-likelihood within each field.

$$L(\tau', \phi', \boldsymbol{\beta}; \mathbf{y}(\mathbf{s})) \approx L(\tau, \phi, \boldsymbol{\beta}; \mathbf{y}(\mathbf{s})).$$

Therefore,  $\phi$  is a weakly identifiable parameter. The same holds true for the partial sill parameter,  $\tau$ .

We can also classify the weak identifiability of  $\tau$  and  $\phi$  as Bayesian nonidentifiability by examining the full conditional distribution. The full conditional distributions include conditioning on the spatial random effect,  $\mathbf{W}(\mathbf{s})$ . For comparison, the posterior of  $\phi$  conditional on the data, parameters, and spatial random effect  $\mathbf{W}(\mathbf{s})$ , can be written as

$$\begin{aligned} p(\phi | \mathbf{y}(\mathbf{s}), \mathbf{W}(\mathbf{s}), \tau, \boldsymbol{\beta}) &\propto p(\mathbf{y}(\mathbf{s}) | \mathbf{W}(\mathbf{s}), \tau, \phi, \boldsymbol{\beta}) p(\mathbf{W}(\mathbf{s}) \phi, \tau, \boldsymbol{\beta}) \\ &\propto p(\mathbf{y}(\mathbf{s}) | \mathbf{W}(\mathbf{s}), \boldsymbol{\beta}) p(\phi | \mathbf{W}(\mathbf{s}), \tau) \\ &\propto p(\phi | \mathbf{W}(\mathbf{s}), \tau). \end{aligned}$$

Therefore, the full conditional posterior distribution of  $\phi$  does not depend on the data,  $\mathbf{y}(\mathbf{s})$ . It is important to remember that even though the spatial parameters are Bayesian nonidentifiable, it does mean there is no Bayesian learning. This encourages exploring different prior distributions for  $\tau$  and  $\phi$  to maximize the potential for Bayesian learning of the parameters from the data.

When fitting an MCMC algorithm to estimate the parameters in a spatial PLMM, we want to assign suitable prior distributions to the weakly identifiable parameters,  $\tau$  and  $\phi$ . In the literature, “suitable” ranges the entire spectrum from informative to vague with infinite variance (e.g. Gelfand and Sahu, 1999; Eberly and Carlin, 2000; Gelfand et al., 2000; Berger et al., 2001; Banerjee et al., 2003; Schmidt et al., 2008) This can also include various parameterizations of the spatial covariogram. We begin by briefly discussing a few different choices of prior distributions for  $\tau$  and  $\phi$ . Then, we propose a new parameterization of the exponential covariogram and compare it to a current prior specifications via simulation.

The second-stage spatial model (32) assumes the spatial error,  $W(\mathbf{s}_i)$ , have a stationary, isotropic, mean-zero Gaussian process with exponential covariance function(31). That is,

$$\text{Cov}(W(\mathbf{s}_i), W(\mathbf{s}_j)) = \tau \exp^{-\frac{1}{\phi}d_{ij}}$$

where  $d_{ij}$  is the distance between locations  $\mathbf{s}_i$  and  $\mathbf{s}_j$ . To fit the model in the Bayesian framework, we need to assign prior distributions to both  $\tau$  and  $\phi$ . It is well-known that prior specifications strongly influence inference in spatial models (Berger et al., 2001). Therefore, we want to be aware of the implications of assigning certain priors.

It is common in the literature of geostatistics to assume that  $\tau$  and  $\phi$  are independent a priori. Therefore, we will first discuss priors where  $p(\tau, \phi) = p(\tau)p(\phi)$ . Banerjee et al. (2003) suggest assigning informative prior distributions since improper priors for the spatial covariogram can lead to improper posteriors. Many assign a conjugate inverse-Gamma prior distribution to the partial-sill parameter,  $\tau$ , and a Gamma prior distribution to the range parameter,  $\phi$ . The difficulty stems from choosing hyperpriors for these prior distributions. Gelfand et al. (2000), for example, suggests  $\tau \sim \text{Inv. Gamma}(2, 1)$  such that prior has mean 1 and infinite variance. They suggest this vague prior is reasonable since pure heterogeneity (white noise or nugget) in the model is fixed at 1. The prior distribution assigned to  $\phi$  is specific to the size of the spatial domain. The hyperpriors can be specified using the effective range where the effective range of the exponential covariogram is approximately  $3\phi$ . A suitable prior for the effective range is  $3\phi \sim \text{Unif}(d_{max}/100, d_{max})$  where  $d_{max}$  is the maximum inter-location distance (Higgs and Hoeting, 2010). On the unit square, this translates to  $\phi \in [0.004, 0.7]$ . Schmidt and Gelfand (2003) suggest fitting the mean of the prior distribution for  $\phi$  to  $\frac{1}{6}d_{max}$ . Berger et al. (2001) suggest assigning a reference prior to  $p(\tau, \phi)$  such that  $p(\tau, \phi) \propto \frac{p(\phi)}{\tau^a}$  where  $p(\phi)$  is the prior distribution of  $\phi$ . In contrast to other common non-informative priors, the reference prior is advantageous as it is non-informative yet yields a proper posterior distribution.

Recall that the second-stage spatial probit model with no fixed effects is

$$\mathbf{Z}(\mathbf{s}) \sim N(\mathbf{W}(\mathbf{s}), \sigma^2 \mathbf{I})$$

where a deterministic relationship exists between  $\mathbf{Z}(\mathbf{s})$  and the binary or ordinal observable response,  $\mathbf{Y}(\mathbf{s})$ . Instead of fixing  $\sigma^2$  to 1 for identifiability, Heagerty and Lele (1998) suggest fixing  $\sigma^2 + \tau = 1$ . The spatial process,  $\mathbf{W}(\mathbf{s})$ , remains unchanged and is modeled

$$\mathbf{W}(\mathbf{s}) \sim N(\mathbf{0}, \tau \mathbf{R}(\mathbf{d}, \phi))$$

but the latent variable  $\mathbf{Z}(\mathbf{s})$  is now modeled

$$\mathbf{Z}(\mathbf{s}) \sim N(\mathbf{W}(\mathbf{s}), (1 - \tau) \mathbf{I}).$$

In this case,  $\tau$  represents the total variance attributable to the spatial variation.

Since neither  $\tau$  nor  $\phi$  is consistently estimable and only  $\tau/\phi$  is consistently estimable, it may suggest that  $\tau$  and  $\phi$  should be assigned a joint prior distribution. The two parameters could also be updated in block-form using a Metropolis-Hastings step. Further, it seems not unreasonable to fix one of the spatial parameters in lieu of estimating the other. For example, we could fix  $\phi$  to its estimate from an indicator variogram. Conversely, we could fix  $\tau$  to a value that gives a reasonable nugget-to-sill ratio for spatial variation.

Re-parameterization of the exponential covariogram is also a reasonable alternative when fitting the model within the Bayesian framework. Whereas re-parameterizing will not change the likelihood of the model, it will change the conditional posterior distributions. In Chapter 4.1, we discussed second-stage and nugget-plus-covariance spatial modeling. Second-stage modeling is a type of hierarchical centering and can lead to faster convergence of the MCMC algorithm. Even though a re-parameterization of the exponential covariogram (31) cannot make all parameters consistently estimable, it is possible to re-parameterize such that one

of the parameters is consistently estimable. Recall the equivalency of the the following covariograms:

$$\tau\phi_1 \exp^{-\frac{1}{\phi_1}d} = \tau\phi_2 \exp^{-\frac{1}{\phi_2}d},$$

where  $\phi_1 \neq \phi_2$  and  $\tau > 0$  is a constant. This results from the exponential covariogram being fully defined by its ratio  $\tau/\phi$  (Zhang, 2004, Theorem 2). Therefore, we define  $\eta$  to be the consistently estimable parameter such that  $\eta = \tau/\phi$  and re-write the exponential covariogram in (31) as

$$\text{Cov}(W(\mathbf{s}_i), W(\mathbf{s}_j)) = \eta\phi \exp^{-\frac{1}{\phi}d_{ij}}. \quad (35)$$

The motivation for the re-parameterization in (35) is two-fold: first, it is asymptotically advantageous since  $\eta$  is consistently estimable, and second, it enhances the performance of the MCMC algorithm. Having both the partial-sill and the range parameter be a function of  $\phi$  leads to better mixing properties of the MCMC by increasing the variability of the parameters between iterations. It should also lead to better mixing of the fixed effect parameter,  $\beta$ , since for small values of  $\phi$  the fixed and random effect parameters can be weakly identifiable, and thus, highly correlated in the MCMC. This is similar to implementing parameter expansion to the iterative sampling algorithms discussed in Chapter 2

Christensen et al. (2006) suggest a similar re-parameterization where  $\theta_1 = \log(\tau^{1/2})$  and  $\theta_2 = \log(\tau/\phi)$ . Their approach is based on the fact that geostatistical data is most informative about the covariance function near the origin. Therefore, they let  $\tau$  control the value at the origin where  $\tau/\phi$  is its derivative. Diggle and Ribeiro (2007, Chapter 5.4) offer a slightly different parameterization where they define  $\theta_1 = \log(\tau/\phi)$  and  $\theta_2 = \log(\phi)$ . Both of these logarithmic transformations result in each  $\theta_1$  and  $\theta_2$  requiring a Metropolis-Hastings step.

To fit the re-parameterized model (35), we need to assign prior distributions to  $\eta$  and  $\phi$ . We aim to assign similar prior distributions to make comparisons between the two model parameterizations. Therefore, the prior on  $\phi$  is the same Gamma prior used in the original

parameterization. The inverse-Gamma is still a conjugate prior for  $\eta$ , where we adjust the hyperpriors to compensate for  $\eta$  being scaled by  $\phi$ .

We simulate 300 locations on the unit square from a spatial PLMM. The mean of the latent variable  $\mathbf{Z}(\mathbf{s})$  contains an intercept term and one covariate, as well as the spatial random effect,  $\mathbf{W}(\mathbf{s})$ . The covariate,  $\mathbf{X}_1(\mathbf{s})$  is drawn independently from a Uniform(-0.5, 0.5). The coefficient vector is  $(\beta_0, \beta_1) = (-0.5, -\sqrt{2})$  for the binary response model and  $(\beta_0, \beta_1) = (0.5, -\sqrt{2})$  for the ordinal response model. The partial-sill and range parameter of the spatial covariance function are set to  $\tau = 1.0$  and  $\phi = 0.2$ , respectively, corresponding to a derived parameter value of  $\eta = 5$ . For the original parameterization, the priors assigned to the spatial parameters are  $\tau \sim \text{Inv. Gamma}(2, 3)$  and  $\phi \sim \text{Gamma}(1, 2)$ . The prior for  $\eta$  is adjusted such that  $\eta\phi$  spans a similar range as the prior of  $\tau$ . Thus,  $\eta \sim \text{Inv. Gamma}(2, 30)$ .

We run MCMC for 100,000 iterations for the binary and ordinal response data under both parameterizations and disregard the first 10,000 as burn-in. The posterior medians and 95% credible intervals for the parameters of the spatial PLMM for binary response data are given in Table 3.2. The estimates of the parameters for the two parameterizations are quite similar where they both underestimate the range parameter,  $\phi$ , and overestimate the partial-sill parameter,  $\tau$ . Estimates of effective sample size and autocorrelation plots for the spatial probit model for binary response data are given in Table 3.3 and Figures 3.15 and 3.16. All effective sample sizes using the proposed parameterization are higher than those resulting from the original parameterization. The differences are quite dramatic in the effective sample size for the fixed effect covariate,  $\beta_1$ . The autocorrelation plots reiterate that the proposed parameterization results in lower autocorrelation and faster mixing of the chain when fitting the spatial PLMM to binary data.

Table 3.4 gives the posterior medians and 95% credible intervals for the ordinal response model. The estimates of  $\beta_0$  and  $\beta_1$  are similar between the two parameterizations. The spatial parameter estimates, however, differ in that the original parameterization slightly overestimates  $\tau$  and underestimates  $\phi$ , whereas the proposed parameterization underestimates  $\eta$  and

Table 3.2: Posterior medians and 95% credible intervals of identifiable parameters of the spatial PLMM for binary response data under each parameterization for 90,000 iterations

Parameter	True value	Parameterization	
		Original	Proposed
$\beta_0$	-0.50	-0.77 (-1.65, 0.48)	-0.65 (-1.24, -0.24)
$\beta_1$	$-\sqrt{2}$	-1.56 (-3.61, -0.71)	-1.37 (-2.42, -0.66)
$\tau$	1	1.65 (0.53, 10.04)	*0.98 (0.35, 4.18)
$\eta$	5	*38.55 (1.40, 491.88)	17.87 (5.46, 119.73)
$\phi$	0.2	0.04 (0.01, .94)	0.05 (0.02, 0.19)

\* Denotes estimate is of a derived parameter

Table 3.3: Estimates of effective sample size of identifiable parameters of the spatial PLMM for binary response data under each parameterization for 90,000 iterations.

Parameter	Parameterization	
	Original	Proposed
$\beta_0$	201	285
$\beta_1$	381	1,373
$\tau$	123	
$\eta$		223
$\phi$	58	324

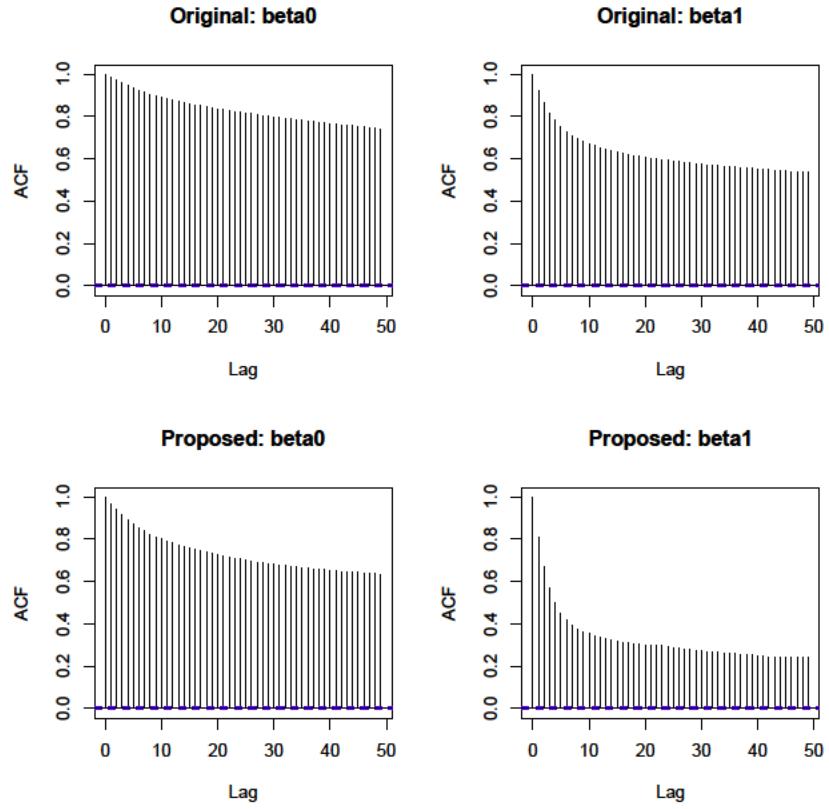


Figure 3.15: Autocorrelation plots for the coefficient vector,  $\beta$ , for the binary spatial PLMM under the original parameterization (top) and new proposed parameterization (bottom).

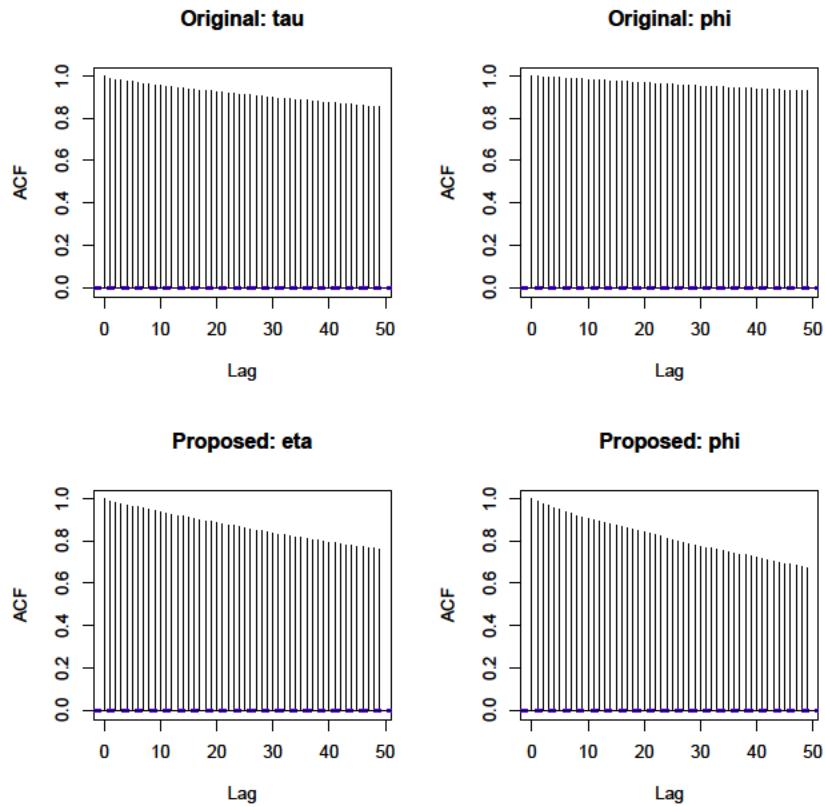


Figure 3.16: Autocorrelation plots for the spatial parameters,  $\tau$  and  $\phi$ , for the binary spatial PLMM under the original parameterization (top), and  $\eta$  and  $\phi$  under the proposed parameterization (bottom).

Table 3.4: Posterior medians and 95% credible intervals of identifiable parameters of the spatial PLMM for ordinal response data under each parameterization for 90,000 iterations.

Parameter	True value	Parameterization	
		Original	Proposed
$\beta_0$	0.5	0.49 (-0.08, -0.98)	0.41 (-0.53, 1.52)
$\beta_1$	$-\sqrt{2}$	-1.77 (-2.67, -1.14)	-1.45 (-1.92, -0.98)
$\tau$	1	1.25 (0.59, 3.25)	*0.51 (0.20, 2.86)
$\eta$	5	*18.93 (3.04, 86.98)	2.20 (1.78, 2.74)
$\phi$	0.2	0.06 (0.03, 0.46)	0.23 (0.09, 1.29)
$\lambda_2$	0.6	0.72 (0.54, 1.01)	0.56 (.046, 0.68)
$\lambda_3$	1.2	1.45 (1.15, 2.00)	1.14 (1.00, 1.30)
$\lambda_4$	1.8	2.12 (1.72, 2.91)	1.69 (1.51, 1.87)

\* Denotes estimate is of a derived parameter

estimates  $\phi$  well. This leads to an overall underestimate of the derived partial-sill parameter when fitting the proposed parameterized model. Weak identifiability is apparent in the estimates for both parameterizations. The original parameterization overestimates  $\tau$  and each value of the threshold vector,  $\boldsymbol{\lambda}$ , whereas the proposed parameterization underestimates  $\eta$  and  $\boldsymbol{\lambda}$ . Table 3.5 and Figures 3.17, 3.18, and 3.19 give the effective sample size estimates and autocorrelation plots for the spatial PLMM for ordinal response data. Mixing appears extremely fast for  $\beta_1$  and  $\eta$  where autocorrelation drops to near zero within the first few lags.

Our results suggest that weak identifiability of the spatial parameters of the spatial probit regression model for binary and ordinal response data greatly affect convergence of  $\boldsymbol{\beta}$  and  $\boldsymbol{\lambda}$  as well. Increasing the variability within the chain by defining the partial-sill of the spatial covariogram as a function of  $\phi$  greatly increases the mixing of the chain. This is a similar idea to parameter expansion strategies presented in Chapter ?? where similar results were obtained. Therefore, when fitting a spatial probit regression model, we strongly suggest using the proposed parameterization of the spatial covariogram.

Table 3.5: Estimates of effective sample size of identifiable parameters of the spatial PLMM for ordinal response data under each parameterization for 90,000 iterations.

Parameter	Parameterization	
	Original	Proposed
$\beta_0$	834	462
$\beta_1$	828	47,253
$\tau$	390	
$\eta$		87,907
$\phi$	107	1,147
$\lambda_2$	401	11,157
$\lambda_3$	288	8,602
$\lambda_4$	265	8,166

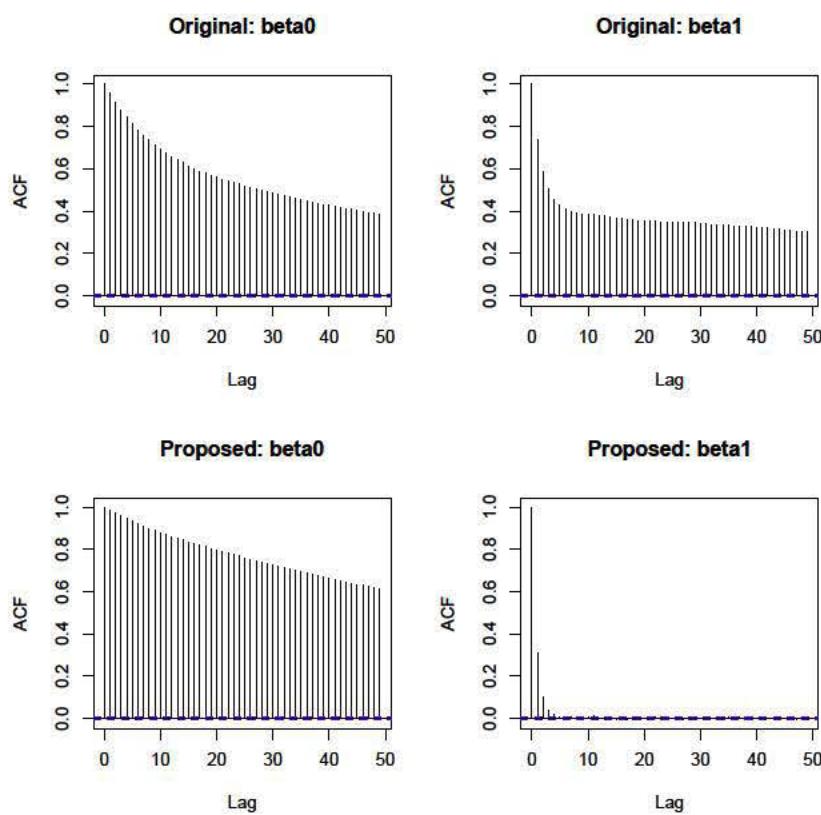


Figure 3.17: Autocorrelation plots for the coefficient vector,  $\beta$ , for the spatial PLMM for ordinal response data under the original parameterization (top) and new proposed parameterization (bottom).

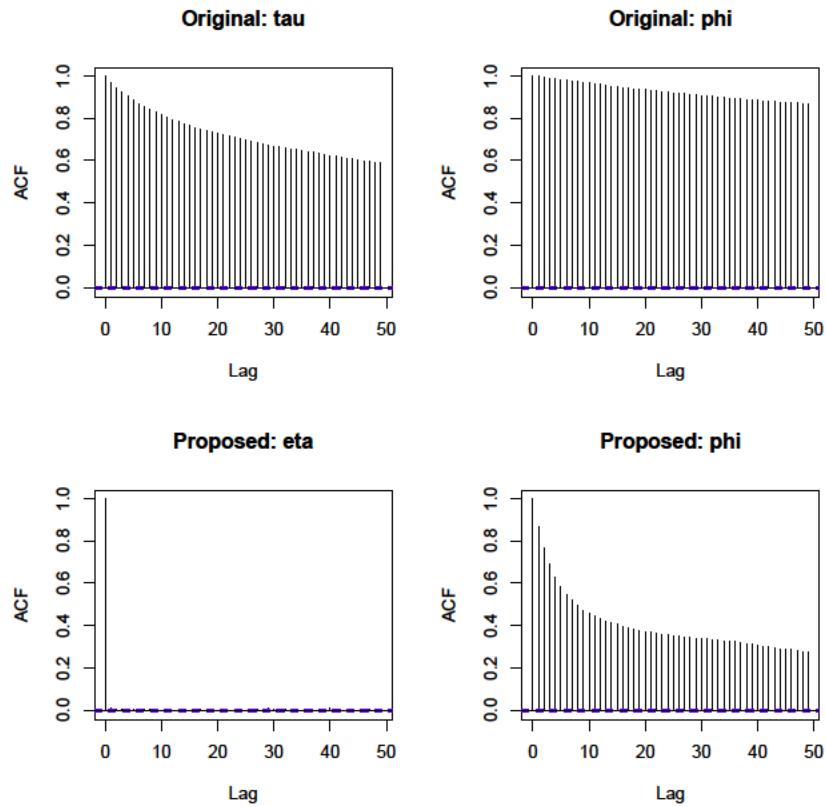


Figure 3.18: Autocorrelation plots for the spatial parameters,  $\tau$  and  $\phi$ , for the spatial PLMM for ordinal response data under the original parameterization (top), and  $\eta$  and  $\phi$  under the proposed parameterization (bottom).

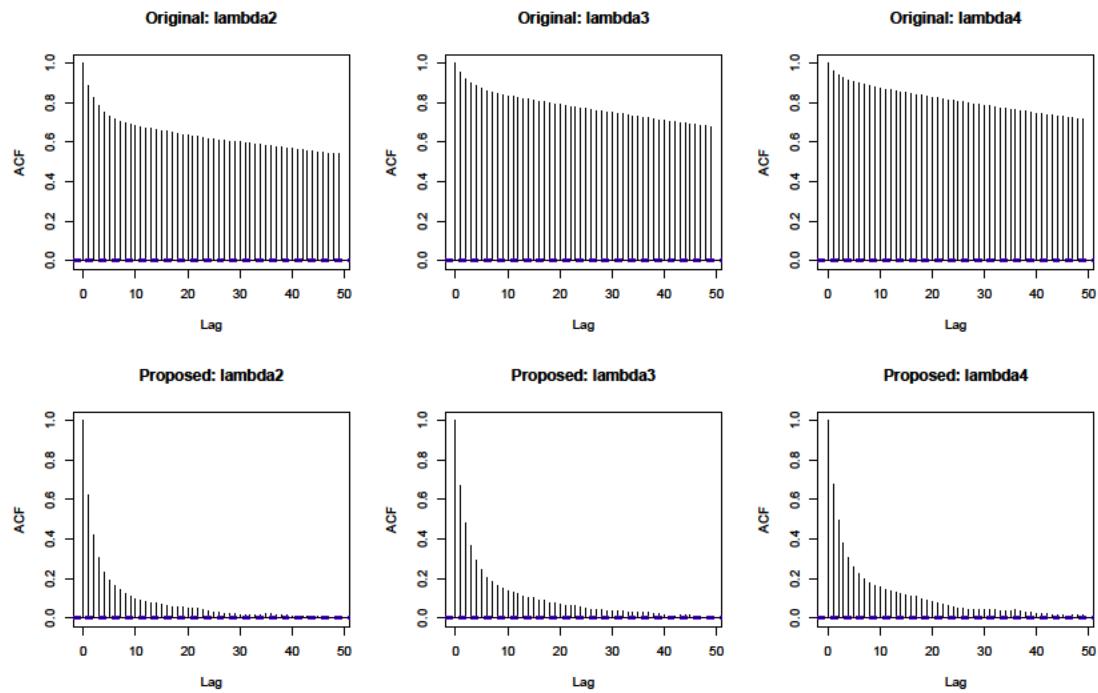


Figure 3.19: Autocorrelation plots for the threshold parameters,  $\lambda_2$ ,  $\lambda_3$ , and  $\lambda_4$ , for the spatial PLMM for ordinal response data under the original parameterization (top) and new proposed parameterization (bottom).

### 3.5 Discussion

We have shown using the mapping between fundamental and reduced-form parameters that the non-spatial common factor probit model with covariates for ordinal response data is identifiable (Section 3.2.2). Using theoretical work for spatial GLMMs and empirical results for spatial LMMs for continuous response data, we investigated identifiability for spatial probit models (Section 3.3.2). The log-likelihood surface plots for spatial PLMMs suggest that the partial sill and range parameter of the exponential covariance function are not identifiable. The positively correlated modal regions of the log-likelihood surfaces indicate that the data are able to estimate the ratio  $\tau/\phi$ . This agrees with the previous work stating that the ratio of the partial sill and range parameter is consistently estimable. We found that the modal region of the surface becomes more localized as the spatial signal increases and the spatial range decreases (i.e.,  $\tau$  increases and  $\phi$  decreases). Further, binary, ordinal, and continuous response data all underestimates  $\tau$  and  $\phi$  where the underestimation is greater for larger  $\phi$ . Lastly, the signal in the response likelihood for  $\tau$  and  $\phi$  increases with data richness where richness is defined as the amount of information in the data.

Weakly identifiable parameters can hinder Bayesian inference by slowing convergence of MCMC. Therefore, appropriate specification of prior distributions to weakly identifiable is extremely important. Priors that are not informative can cause the Markov chain for the weakly identifiable parameters to drift to extreme values leading to inaccurate estimates (Gelfand and Sahu, 1999). Priors that are too informative, however, will limit Bayesian learning from the data. We proposed re-parameterizing the exponential covariance function and assigning proper prior distributions. The re-parameterization defines  $\eta\phi$  as the partial sill and  $\phi$  is the range parameter. This parameterization led to better mixing of the MCMC by decreasing the autocorrelation in the chain between iterations. Our results of the proposed parameterization, however, were only for a small simulation study. Whereas the proposed parameterization is dramatically better in terms of convergence for the spa-

tial PLMM, additional analyses will be performed as part of future work on this subject. This includes investigating the prior distributions and their hyperpriors assigned to  $\eta$  and  $\phi$  as well as using the proposed parameterization of the exponential covariance function for different models and types of data. Further, comparisons could be made between our re-parameterization and the similar parameterizations of Christensen et al. (2006) and Diggle and Ribeiro (2007, Chapter 5.4). One benefit of our parameterization is that  $\eta$  has a conjugate update and therefore does not require a Metropolis-Hastings step.

Xie and Carlin (2006) compute estimates of Bayesian learning of weakly identifiable parameters for Gaussian hierarchical linear models with focus on conditional autoregression spatial models. We would like to extend Bayesian learning estimation to the spatial parameters of a geostatistical spatial model. This includes quantifying the amount of Bayesian learning for spatial PLMM models for binary and ordinal response data. Comparisons could be made between the different response data types as well as between the the different parameterizations. This would lead to enhanced modeling for a general class of spatial models in terms of parameter inference and computational efficiency.

## CHAPTER 4

### COMPARING AND CONTRASTING FIRST-STAGE AND SECOND-STAGE SPATIAL PROBIT MODELS

Multiple forms of the probit model for spatially-correlated binary and ordinal response data have been adopted in the literature (e.g. De Oliveira, 2000; Gelfand et al., 2000; Higgs and Hoeting, 2010; Schliep and Hoeting, 2013). In this chapter, we compare two model structures that differ in the level, or stage, containing spatial correlation. The *first-stage model* assumes spatial correlation at the data level. That is, the binary or ordinal response data has a spatial covariance matrix. The *second-stage model* assumes spatial correlation in the process level where the mean of the binary ordinal response data is a function of a spatially-correlated random variable. Model identifiability differs slightly between the two models since the second-stage model contains an additional parameter in the spatial covariance function. This allows the second-stage model to be more flexible when fitting the model to observed binary or ordinal data. We show that for certain parameter values, the second-stage spatial probit model can mimic the first-stage spatial model in terms of parameter inference and prediction. The more restrictive first-stage model, however, is unable to resemble the second-stage model in most cases. We discuss the implications of fitting each model and their impact on parameter estimation and prediction. Comparing these two models enhances our knowledge of spatial models for ordinal response data.

#### 4.1 First-stage and second-stage spatial probit models

In this section we compare the first-stage and second-stage spatial probit models for binary response data noting that binary data is a special case of ordinal data. The models and their likelihood functions shown within can easily be modified for ordinal data with more

than 2 categories. For observed response data  $\mathbf{y}(\mathbf{s}) = \{y(\mathbf{s}_1), \dots, y(\mathbf{s}_n)\}$ , covariates  $\mathbf{X}(\mathbf{s}_i)$ , for  $i = 1, \dots, n$ , and coefficient parameter vector  $\boldsymbol{\beta}$ , the likelihood for the traditional probit model for independent binary data is

$$L(\boldsymbol{\beta}; \mathbf{y}(\mathbf{s})) = \prod_{i=1}^n \Phi(\mathbf{X}(\mathbf{s}_i)' \boldsymbol{\beta})^{y(\mathbf{s}_i)} (1 - \Phi(\mathbf{X}(\mathbf{s}_i)' \boldsymbol{\beta}))^{1-y(\mathbf{s}_i)}.$$

The latent variable data augmentation version of this model assumes  $\mathbf{Z}(\mathbf{s}) \sim N(\mathbf{X}(\mathbf{s})\boldsymbol{\beta}, \mathbf{I})$  with

$$p(Z(\mathbf{s}_i)|y(\mathbf{s}_i), \boldsymbol{\beta}) \sim \begin{cases} TN(\mathbf{X}(\mathbf{s}_i)' \boldsymbol{\beta}, 1, -\infty, 0) & y(\mathbf{s}_i) = 0 \\ TN(\mathbf{X}(\mathbf{s}_i)' \boldsymbol{\beta}, 1, 0, \infty) & y(\mathbf{s}_i) = 1, \end{cases}$$

where  $TN(\mu, \sigma^2, \lambda_{lower}, \lambda_{upper})$  specifies a normal distribution with mean  $\mu$ , variance  $\sigma^2$ , and lower and upper truncation points,  $\lambda_{lower}$  and  $\lambda_{upper}$ , respectively.

One approach for modeling spatial correlation in binary data using probit regression is to replace the identity covariance matrix,  $\mathbf{I}$ , with a spatial covariance matrix. This is known as a direct or first-stage spatial probit model since the spatial correlation is assumed directly on  $\mathbf{Z}(\mathbf{s})$ . The deterministic relationship between  $\mathbf{Y}(\mathbf{s})$  and  $\mathbf{Z}(\mathbf{s})$ , implies spatial correlation directly on the binary response data making the model not within the class of PLMMs. The first-stage spatial model is a very natural extension of Albert and Chib (1993). Here,

$$\mathbf{Z}(\mathbf{s}) \sim N(\mathbf{X}(\mathbf{s})\boldsymbol{\beta}, \tau \mathbf{R}) \quad (36)$$

where  $\mathbf{R}$  is a correlation matrix defined by parameter,  $\phi$ , and spatial distance matrix,  $\mathbf{d}$ . For the exponential covariogram,  $\phi$  is univariate and  $\mathbf{R}$  is defined by

$$R(\phi, d_{ij}) = Cor(Z(\mathbf{s}_i), Z(\mathbf{s}_j)) = \exp^{-\frac{1}{\phi}d_{ij}} \quad (37)$$

where  $d_{ij}$  is the distance between locations  $\mathbf{s}_i$  and  $\mathbf{s}_j$  and  $\phi$  is the range parameter. When  $\tau$  is fixed to 1, the first-stage model parameters are identifiable. The likelihood function for the

first-stage spatial probit model is a multivariate integral and a function of the parameters  $\beta$ ,  $\phi$ , and  $\tau$  where  $\tau = 1$ . It is computed

$$L(\beta, \tau = 1, \phi; \mathbf{y}(\mathbf{s})) = \int_{\lambda_{y(\mathbf{s}_1)}}^{\lambda_{y(\mathbf{s}_1)+1}} \cdots \int_{\lambda_{y(\mathbf{s}_n)}}^{\lambda_{y(\mathbf{s}_n)+1}} \left( \frac{1}{2\pi} \right)^{n/2} |\tau \mathbf{R}|^{-1/2} \times \exp \left\{ -\frac{1}{2} (\mathbf{Z}(\mathbf{s}) - \mathbf{X}(\mathbf{s})\beta)' (\tau \mathbf{R})^{-1} (\mathbf{Z}(\mathbf{s}) - \mathbf{X}(\mathbf{s})\beta) \right\} d\mathbf{Z}(\mathbf{s}) \quad (38)$$

where  $\lambda_0 = -\infty$ ,  $\lambda_1 = 0$ , and  $\lambda_2 = \infty$ . The first-stage spatial probit model is a no-nugget model since all variation in the response is being modeled as spatially-correlated variation. The first-stage spatial model implies that the observable binary response,  $Y(\mathbf{s}_i)$ , is highly correlated with  $Y(\mathbf{s}_j)$  for nearby locations,  $\mathbf{s}_i$  and  $\mathbf{s}_j$ . Oliveira (2000) fit clipped Gaussian random field models containing first-stage spatial correlation which are similar and conclude that they are suitable in situations where there is a high degree of smoothness in the binary response.

In contrast, we propose a second-stage spatial regression model with spatial random effect. The model fits within the PLMM framework and is arguably more intuitive than the first-stage spatial model (36). Using the same notation as the first-stage spatial model where  $\mathbf{Z}(\mathbf{s})$  is the augmented data, the second-stage model assumes

$$\mathbf{Z}(\mathbf{s}) \sim N(\mathbf{X}(\mathbf{s})\beta + \mathbf{W}(\mathbf{s}), \mathbf{I}), \quad (39)$$

where  $\mathbf{W}(\mathbf{s})$  is a spatial random effect such that  $\mathbf{W}(\mathbf{s}) \sim N(0, \Sigma_W)$  and  $\Sigma_W = \tau \mathbf{R}$ . This model allows for spatial correlation in the mean of the latent variable  $\mathbf{Z}(\mathbf{s})$  but does not require an artificially smooth binary response surface. In regards to second-stage modeling, Banerjee et al. (2003, p.p. 146-148) states, “Introducing the spatial effects in the mean encourages the mean of the spatial variables at proximate locations to be close to each other, adjusted for covariates. Though marginal dependence is induced between  $Y(\mathbf{s}_i)$  and  $Y(\mathbf{s}_j)$ , the observed  $y(\mathbf{s}_i)$  and  $y(\mathbf{s}_j)$  need not be close to each other.”

The likelihood function for the second-stage spatial probit model as a function of the parameters  $\beta$ ,  $\tau$ ,  $\phi$ , and the random effect,  $\mathbf{W}(\mathbf{s})$ , is

$$\begin{aligned}
L(\beta, \mathbf{W}(\mathbf{s}), \tau, \phi; \mathbf{y}(\mathbf{s})) &= L(\beta, \mathbf{W}(\mathbf{s}); \mathbf{y}(\mathbf{s})) \\
&= \int_{\lambda_{y(s_1)}}^{\lambda_{y(s_1)+1}} \cdots \int_{\lambda_{y(s_n)}}^{\lambda_{y(s_n)+1}} \left( \frac{1}{2\pi} \right)^{n/2} |\mathbf{I}|^{-1/2} \\
&\quad \times \exp \left\{ -\frac{1}{2} (\mathbf{Z}(\mathbf{s}) - \mathbf{X}(\mathbf{s})\beta - \mathbf{W}(\mathbf{s}))' \mathbf{I}^{-1} \right. \\
&\quad \left. (\mathbf{Z}(\mathbf{s}) - \mathbf{X}(\mathbf{s})\beta - \mathbf{W}(\mathbf{s})) \right\} d\mathbf{Z}(\mathbf{s})
\end{aligned} \tag{40}$$

where the first line holds since the parameters  $\tau$  and  $\phi$  are in the second stage of the latent process and thus do not appear in the likelihood (first stage). Therefore, conditional on  $\mathbf{W}(\mathbf{s})$ , the likelihood does not contain the spatial parameters  $\tau$  and  $\phi$ . When we marginalize the second-stage spatial model (40) over the spatial random effect,  $\mathbf{W}(\mathbf{s})$ , the likelihood is

$$\begin{aligned}
L(\beta, \tau, \phi; \mathbf{y}(\mathbf{s})) &= \int_{\lambda_{y(s_1)}}^{\lambda_{y(s_1)+1}} \cdots \int_{\lambda_{y(s_n)}}^{\lambda_{y(s_n)+1}} \left( \frac{1}{2\pi} \right)^{n/2} |\tau \mathbf{R} + \mathbf{I}|^{-1/2} \\
&\quad \times \exp \left\{ -\frac{1}{2} (\mathbf{Z}(\mathbf{s}) - \mathbf{X}(\mathbf{s})\beta)' (\tau \mathbf{R} + \mathbf{I})^{-1} (\mathbf{Z}(\mathbf{s}) - \mathbf{X}(\mathbf{s})\beta) \right\} d\mathbf{Z}(\mathbf{s}) \\
&= \int_{\lambda_{y(s_1)}}^{\lambda_{y(s_1)+1}} \cdots \int_{\lambda_{y(s_n)}}^{\lambda_{y(s_n)+1}} \left( \frac{1}{2\pi} \right)^{n/2} |\Sigma_W + \mathbf{I}|^{-1/2} \\
&\quad \times \exp \left\{ -\frac{1}{2} (\mathbf{Z}(\mathbf{s}) - \mathbf{X}(\mathbf{s})\beta)' (\Sigma_W + \mathbf{I})^{-1} (\mathbf{Z}(\mathbf{s}) - \mathbf{X}(\mathbf{s})\beta) \right\} d\mathbf{Z}(\mathbf{s}).
\end{aligned} \tag{41}$$

The marginalized likelihood for the second-stage spatial probit model is the multivariate integral of the likelihood of a Gaussian process covariance-plus-nugget spatial model for continuous response data (Banerjee et al., 2003, p.p. 130-133).

An important differentiation between the likelihood of the first-stage (38) and second-stage (41) spatial probit model is that the first-stage model assumes a no-nugget covariance. This is a major assumption since it means that all variation in the binary response is the

result of the spatial random effect. Therefore, the stochasticity of the binary response is purely spatial as there is no non-spatial component. First-stage spatial modeling may be appropriate when proximate observations are assumed to be close whereas second-stage modeling is appealing when the interest is in the spatial explanation in the mean. Since binary and ordinal data are discrete, we find second-stage spatial modeling via the PLMM (39) to be more appropriate as it limits artificial smoothing, or under-dispersion, in the discrete response. The PLMM also includes an additional spatial covariance parameter,  $\tau$ , the partial sill of the marginal covariance of the latent variable  $\mathbf{Z}(\mathbf{s})$ . Whereas this makes the model more flexible in allowing the stochasticity in the binary response to contain both a spatial and non-spatial component, it can also cause issues with model near nonidentifiability as discussed in Chapter 3. This is true for ordinal response data as well since binary data is a simple case of ordinal data. The likelihood functions in (38), (40), and (41) also hold for ordinal response data. When the ordinal response contains  $K$  categories,  $\lambda_1 = -\infty$ ,  $\lambda_2 = 0$ ,  $\lambda_{K+1} = \infty$ , and  $\lambda_3, \dots, \lambda_K$  are parameters to be estimated.

In Section 4.2 we compare the likelihoods of the first-stage and second-stage model using data simulated under both models. Similar to the parameter identifiability investigations in Chapter 3, we examine the ability of each model's likelihood to detect the signal of the spatial parameters. In Section 4.3 we discuss various methods for predicting ordinal response data at unobserved locations using the posterior predictive distribution. We discuss a method, albeit computationally expensive, for obtaining a posterior distribution of the density function of the unobserved ordinal response. This produces a posterior distribution of the expected value and variance of the unobserved ordinal response. We then offer an efficient and accurate method for approximating the posterior distribution of the expected value and variance and evaluate the method using an example. In Section 4.4 we compare the first-stage and second-stage model in terms of prediction using the approximation. This includes showing that the prediction estimates are equivalent under certain limiting conditions and parameters values

of the first-stage and second-stage model. The chapter concludes with a discussion in Section 4.5.

## 4.2 Comparing first-stage and second-stage model likelihoods via simulation

In this section, we compare the likelihood functions of the first-stage and second-stage spatial probit models via simulation. To start, we naively fit the first-stage spatial model (36) to data simulated using the second-stage, or PLMM, model (39). Figure 4.1 shows the log-likelihood values for the first-stage spatial model for binary response data versus  $\phi$  for data simulated at different values of  $\tau$  and  $\phi$ . For each combination of  $\tau$  and  $\phi$ , the first-stage model greatly underestimates the range parameter,  $\phi$ , to a point of weak or no spatial signal. This is not surprising since the variance of the PLMM contains both a spatially-structured component and an independent noise component. Therefore, the first-stage model having only spatially-structured variance will underestimate  $\phi$  to compensate for being unable to model the extra noise in the response field. As  $\tau$  increases in the second-stage model used for simulating the data, the underestimation of  $\phi$  decreases slightly. This is because large values of  $\tau$  correspond to small nugget-to-sill ratios. When  $\tau = 4$ , for example, the additional noise of the PLMM is muted. We further address the similarity of the first-stage and second-stage spatial models for large  $\tau$  in Section 4.4 in terms of their limiting predictive distributions proposed in Section 4.3.

Figure 4.2 compares the log-likelihood plots when  $\tau = 1$  when the data are fitted using both the first-stage and second-stage model. This shows that the second-stage model is able to capture the spatial signal in the data where the first-stage model could not. Overall, the severe underestimation of  $\phi$  indicates that the first-stage spatial model is inappropriate for discrete response fields that contain unstructured stochasticity.

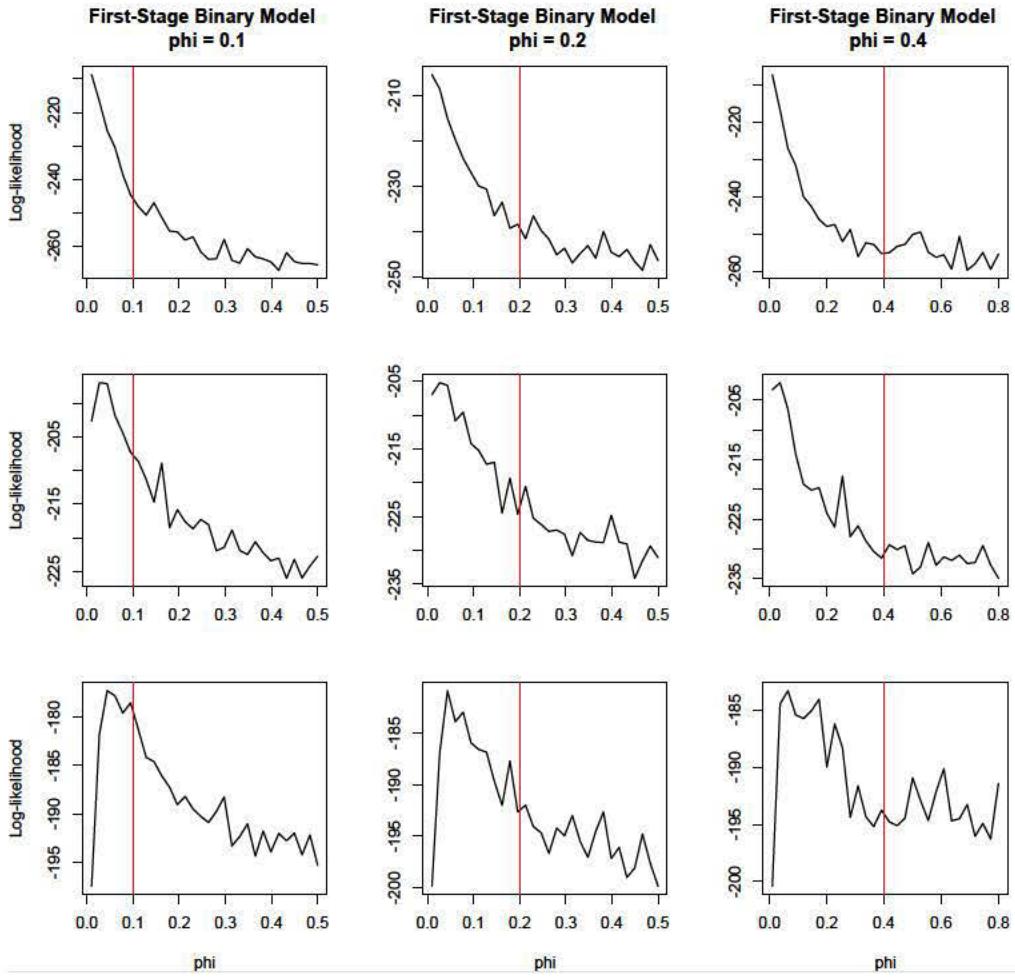


Figure 4.1: Log-likelihood values plotted versus  $\phi$  for binary response data using the first-stage model when data were generated using the PLMM model with  $\tau = 0.25$  (top), 1 (middle), and 4 (bottom) and  $\phi = 0.1$  (left), 0.2 (middle), and 0.4 (right). The vertical lines show the true values of  $\phi$ . Note that for the first-stage model,  $\tau = 1$ .

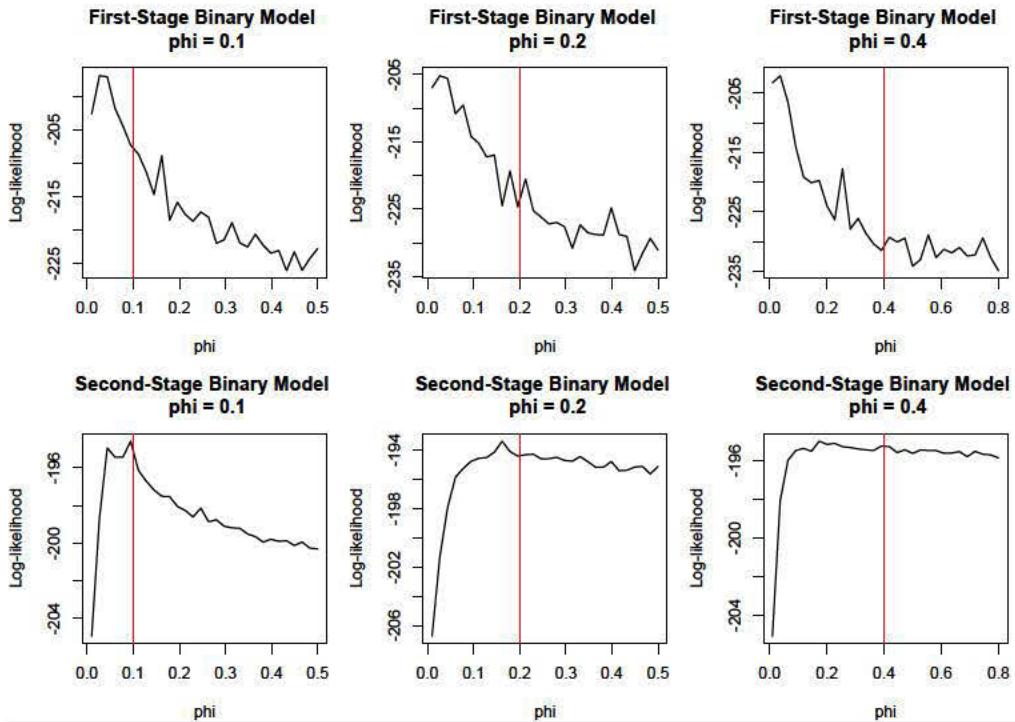


Figure 4.2: Log-likelihood values plotted versus  $\phi$  for binary response data fitted using the first-stage spatial model (top) and second-stage spatial model (bottom). Data were simulated using the second-stage spatial model with  $\tau = 1$  and  $\phi = 0.1$  (left), 0.2 (middle), and 0.4 (right). The vertical lines show the true values of  $\phi$ . Note that for the first-stage model,  $\tau = 1$ .

To more fully compare the first-stage and second-stage model, we simulated data from the first-stage spatial model and fit both the first-stage and second-stage model. Data simulated from the first-stage model assume  $\tau = 1$  for identifiability. Figure 4.3 shows the log-likelihood values versus  $\phi$  for the binary response data fitted using the first-stage model and second-stage model. Even when the data are generated from the first-stage model, ML still underestimates  $\phi$  for both the first-stage and second-stage models. The log-likelihood values of the second-stage model are plotted versus  $\phi$  for fixed values of  $\tau$ . Shown are  $\tau$  set to 1 and its MLE. The log-likelihood trend is similar for both values of  $\tau$  when fitting the second-stage model where ML underestimates  $\phi$ . The underestimation of  $\phi$  appears similar between the first-stage and second-stage model and agrees with the work of Zhang (2004) and Irvine et al. (2007), as well as the results in Chapter 3. There does appear to be more signal in the log-likelihood of the second-stage model compared to the first stage model, however, in that the log-likelihood decreases at a faster rate.

The log-likelihood surfaces of the second-stage model for binary, ordinal, and continuous response data are given in Figure 4.4 for data simulated using the first-stage model. Whereas  $\phi$  is underestimated,  $\tau$  is greatly overestimated. For the same reasons as given above, larger values of  $\tau$  correspond to smaller nugget-to-sill ratios. Therefore, the second-stage model more closely resembles the first-stage model at larger values of  $\tau$ . The binary probit model is able to overestimate  $\tau$  since the variance is unidentifiable. The continuous response model is unable to compensate by increasing  $\tau$  since it is an identifiable parameter. Further, recall that larger values of  $\phi$  lead to less localized spatial correlation where estimation at a particular location is more affected by neighbors and by neighbors at further distances. Therefore, the likelihood surface for the continuous response data overestimates  $\phi$  to adjust for the under-dispersion of the first-stage model. The log-likelihood surface for ordinal data is similar to the continuous response surface. This is because in our simulation the threshold vector is assumed fixed and known putting an upper limit on  $\tau$ . In practice, the threshold vector is estimated, in which case  $\tau$  can increase to compensate for data that has less dispersion.

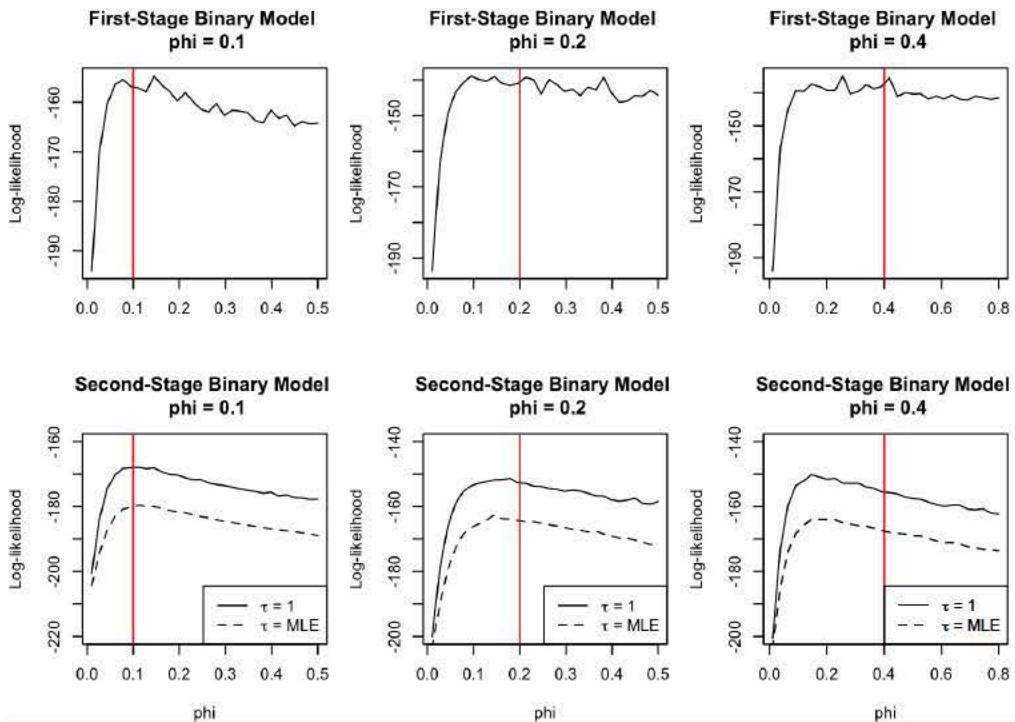


Figure 4.3: Log-likelihood values plotted versus  $\phi$  for binary response data fitted using the first-stage spatial model (top) and second-stage spatial model (bottom). Data were simulated using the first-stage spatial model with  $\tau = 1$  and  $\phi = 0.1$  (left), 0.2 (middle), and 0.4 (right). The log-likelihood values plotted for the second-stage spatial model are for fixed values of  $\tau$ .

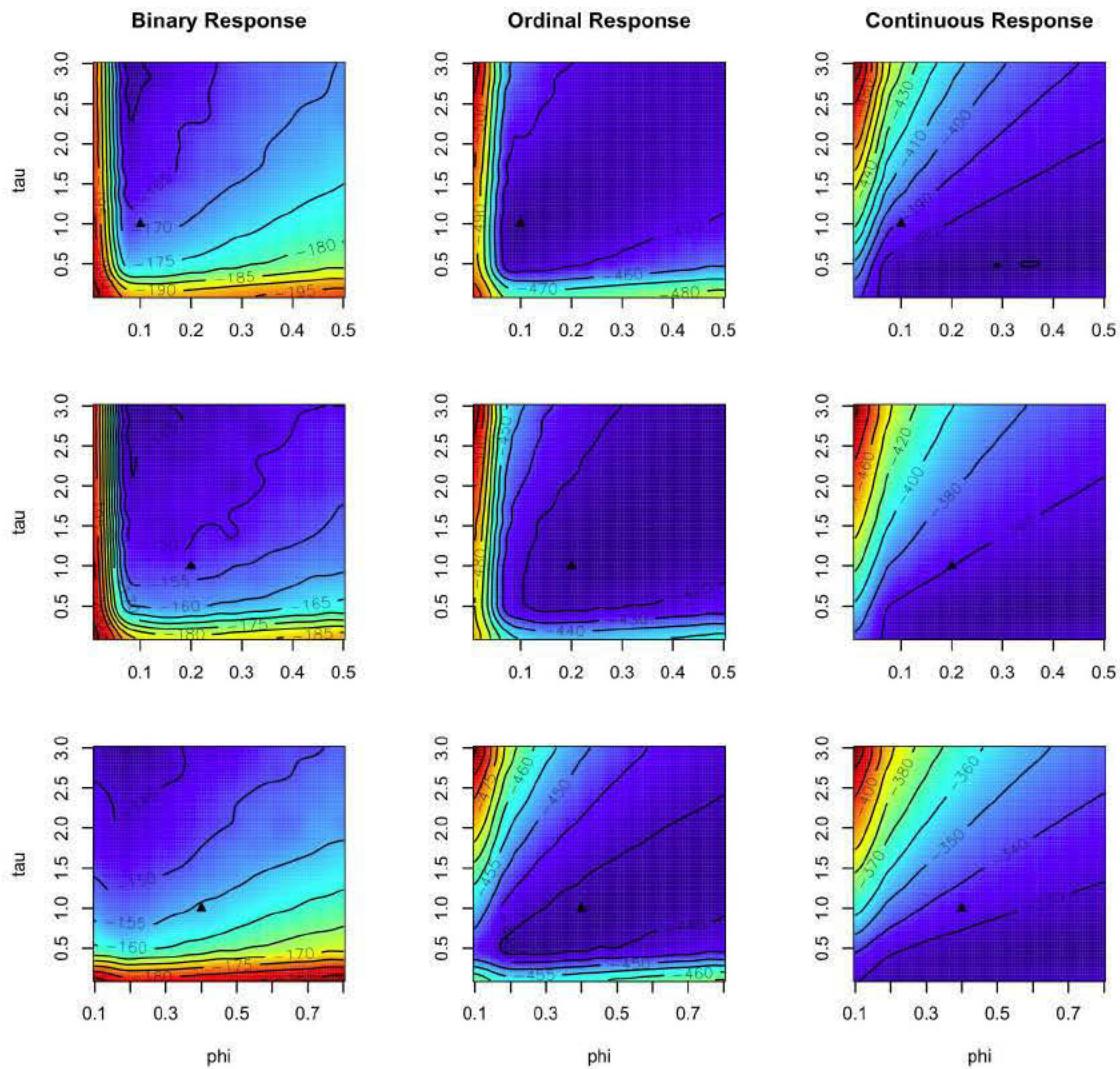


Figure 4.4: Log-likelihood surfaces for binary (left), ordinal (middle) and continuous (right) response data using spatial PLMM. Data were simulated using the first-stage spatial model with  $\tau = 1$  and  $\phi = 0.1$  (left), 0.2 (middle), and 0.4 (right).

Figure 4.3 demonstrates that the first-stage spatial model is only appropriate when it can be assumed that the binary spatial response surface contains only spatially-correlated stochasticity. The second-stage model is more flexible in allowing there to be non-spatial variation in the response field which can limit artificial smoothing. The second-stage model can be similar to the first-stage model for large values of  $\tau$  and can perform equally well in terms of parameter estimation of  $\phi$ . Whereas we focused on binary response data in this section, the same results hold for spatially-correlated ordinal response data with unknown thresholds. Before we make a general recommendation of fitting the second-stage model, even when it is assumed that all variation in the binary or ordinal response is spatially-correlated, we want to compare prediction of the response at unobserved locations using the two models. If the second-stage model is able to predict the binary or ordinal response equally well or better than the first-stage model, we will consider it the optimal probit model for spatially-correlated binary and ordinal data.

### 4.3 Latent variable approach to prediction of ordinal response data

One of the advantages of fitting a geostatistical spatial model, as opposed to a Markov random field, is improved predictions at unobserved locations. Prediction over space, commonly referred to as spatial interpolation, is sought after in many applications. Kriging is a geostatistical estimator at an unobserved location within a random field that is inferred from sample data (see Stein (1999) for theoretical results). Linear estimation draws inference on the response variable at an unobserved location using a linear combination of the observed response variables and weights. The simple kriging estimate is the best linear predictor of the unobserved response in that it minimizes the mean squared prediction error (MSPE). The estimate of the response that minimizes MSPE is the conditional expectation of the unobserved response variable given the observed data. Let  $\mathbf{s}_0$  be the new location of interest and  $\mathbf{s} = (\mathbf{s}_1, \dots, \mathbf{s}_n)$  the observed locations within our sample. The kriging estimate of  $Y(\mathbf{s}_0)$  given the observed data  $\mathbf{y}(\mathbf{s}) = (y(\mathbf{s}_1), \dots, y(\mathbf{s}_n))$  is

$$\hat{Y}(\mathbf{s}_0) = \text{E}(Y(\mathbf{s}_0)|\mathbf{y}(\mathbf{s})).$$

For ordinal response data with  $K$  categories,

$$\begin{aligned}\text{E}(Y(\mathbf{s}_0)|\mathbf{y}(\mathbf{s})) &= \sum_{k=1}^K k P(Y(\mathbf{s}_0) = k|\mathbf{y}(\mathbf{s})) \\ &= \sum_{k=1}^K \frac{k P(Y(\mathbf{s}_0) = k, \mathbf{y}(\mathbf{s}) = \mathbf{y})}{P(\mathbf{y}(\mathbf{s}) = \mathbf{y})}.\end{aligned}\tag{42}$$

For  $k = 1, \dots, K$ , define the threshold vector  $\boldsymbol{\lambda} = \{\lambda_1 = \infty, \lambda_2 = 0, \lambda_3, \dots, \lambda_{K+1} = \infty\}$ . We compute the joint probability of  $Y(\mathbf{s}_0) = k$  and the observed data,  $\mathbf{y}(\mathbf{s})$  as

$$\begin{aligned}P(Y(\mathbf{s}_0) = k, \mathbf{y}(\mathbf{s})) &= \int_{\lambda_k}^{\lambda_{k+1}} \int_{\lambda_{y(\mathbf{s}_1)}}^{\lambda_{y(\mathbf{s}_1)+1}} \cdots \int_{\lambda_{y(\mathbf{s}_n)}}^{\lambda_{y(\mathbf{s}_n)+1}} \left(\frac{1}{2\pi}\right)^{n/2} |\Sigma + \mathbf{I}|^{-1/2} \\ &\quad \times \exp\left\{-\frac{1}{2}(\mathbf{Z}(\mathbf{s}_0, \mathbf{s}) - \mathbf{X}(\mathbf{s}_0, \mathbf{s})\boldsymbol{\beta})'(\Sigma + \mathbf{I})^{-1}(\mathbf{Z}(\mathbf{s}_0, \mathbf{s}) - \mathbf{X}(\mathbf{s}_0, \mathbf{s})\boldsymbol{\beta})\right\} d\mathbf{Z}(\mathbf{s}_0, \mathbf{s})\end{aligned}\tag{43}$$

where  $\mathbf{X}(\mathbf{s}_0, \mathbf{s}) = [\mathbf{X}(\mathbf{s}_0), \mathbf{X}(\mathbf{s}_1), \dots, \mathbf{X}(\mathbf{s}_n)]'$ , and  $\mathbf{Z}(\mathbf{s}_0, \mathbf{s}) = (Z(\mathbf{s}_0), Z(\mathbf{s}_1), \dots, Z(\mathbf{s}_n))'$ .

The covariance matrix for  $\mathbf{Z}(\mathbf{s}_0, \mathbf{s})$ ,  $\Sigma$ , can be partitioned such that

$$\Sigma = \begin{bmatrix} \Sigma_{00} & \Sigma_{01} \\ \Sigma_{10} & \Sigma_{11} \end{bmatrix},$$

where  $\Sigma_{00} = \text{Var}(Z(\mathbf{s}_0))$ ,  $\Sigma_{01} = \Sigma'_{10} = \text{Cov}(Z(\mathbf{s}_0), \mathbf{Z}(\mathbf{s}))$ , and  $\Sigma_{11} = \text{Cov}(\mathbf{Z}(\mathbf{s}), \mathbf{Z}(\mathbf{s}))$ . Each covariance is computed using a valid covariance function (e.g. (31)).

The denominator in (42) contains only the observed data and therefore need only be computed once. It is a multivariate integral of the observed data and is the same as the likelihood function in (41). It is computed

$$P(\mathbf{y}(\mathbf{s}) = \mathbf{y}) = \int_{\lambda_{y(\mathbf{s}_1)}}^{\lambda_{y(\mathbf{s}_1)+1}} \cdots \int_{\lambda_{y(\mathbf{s}_n)}}^{\lambda_{y(\mathbf{s}_n)+1}} \left( \frac{1}{2\pi} \right)^{n/2} |\Sigma_{11} + \mathbf{I}|^{-1/2} \\ \times \exp \left\{ -\frac{1}{2} (\mathbf{Z}(\mathbf{s}) - \mathbf{X}(\mathbf{s})\boldsymbol{\beta})' (\Sigma_{11} + \mathbf{I})^{-1} (\mathbf{Z}(\mathbf{s}) - \mathbf{X}(\mathbf{s})\boldsymbol{\beta}) \right\} d\mathbf{Z}(\mathbf{s}). \quad (44)$$

To calculate  $E(Y(\mathbf{s}_0)|\mathbf{y}(\mathbf{s}))$  in (42), we must compute (43) for  $k = 1, \dots, K$ . This is computationally expensive for even moderate  $n$ .

We propose a method for approximating the conditional expectation (42) using the latent variables  $Z(\mathbf{s}_0)$  and  $\mathbf{Z}(\mathbf{s})$  and the properties of the normal distribution. We write the joint marginal distribution of  $Z(\mathbf{s}_0)$  and  $\mathbf{Z}(\mathbf{s})$  as

$$\begin{pmatrix} Z(\mathbf{s}_0) \\ \mathbf{Z}(\mathbf{s}) \end{pmatrix} \sim N \left( \begin{pmatrix} \mathbf{X}(\mathbf{s}_0)\boldsymbol{\beta} \\ \mathbf{X}(\mathbf{s})\boldsymbol{\beta} \end{pmatrix}, \begin{pmatrix} \Sigma_{00} & \Sigma_{01} \\ \Sigma_{10} & \Sigma_{10} \end{pmatrix} \right). \quad (45)$$

The conditional distribution of  $Z(\mathbf{s}_0)$  given  $\mathbf{Z}(\mathbf{s})$  is

$$Z(\mathbf{s}_0)|\mathbf{Z}(\mathbf{s}) \sim N(\mathbf{X}(\mathbf{s}_0)'\boldsymbol{\beta} + \Sigma_{01}\Sigma_{11}^{-1}(\mathbf{Z}(\mathbf{s}) - \mathbf{X}(\mathbf{s})\boldsymbol{\beta}), \Sigma_{00} - \Sigma_{01}\Sigma_{11}^{-1}\Sigma_{10}). \quad (46)$$

Recall that for ordinal data modeled using the probit link function and the augmented data,

$$P(Y = k) = P(\lambda_k \leq Z < \lambda_{k+1}) = \Phi \left( \frac{\lambda_{k+1} - E(Z)}{\sqrt{\text{Var}(Z)}} \right) - \Phi \left( \frac{\lambda_k - E(Z)}{\sqrt{\text{Var}(Z)}} \right).$$

Therefore,

$$P(Y(\mathbf{s}_0) = k|\mathbf{y}(\mathbf{s})) = P(\lambda_k \leq Z(\mathbf{s}_0) < \lambda_{k+1})|\mathbf{y}(\mathbf{s})).$$

We approximate the conditional probability of  $Y(\mathbf{s}_0)$  given the observed response  $\mathbf{y}(\mathbf{s})$  using the conditional probability of  $Y(\mathbf{s}_0)$  given the latent variable  $\mathbf{Z}(\mathbf{s})$ . That is,

$$\begin{aligned} P(Y(\mathbf{s}_0) = k | \mathbf{y}(\mathbf{s})) &\approx P(Y(\mathbf{s}_0) = k | \mathbf{Z}(\mathbf{s})) \\ &= P(\lambda_k \leq Z(\mathbf{s}_0) < \lambda_{k+1}) | \mathbf{Z}(\mathbf{s})) \end{aligned} \tag{47}$$

Using (47), we approximate the expected value of  $Y(\mathbf{s}_0)$  given  $\mathbf{y}(\mathbf{s})$  as

$$\begin{aligned} \text{E}(Y(\mathbf{s}_0) | \mathbf{y}(\mathbf{s})) &= \sum_{k=1}^K k P(Y(\mathbf{s}_0) = k | \mathbf{y}(\mathbf{s})) \\ &\approx \sum_{k=1}^K k P(\lambda_k \leq Z(\mathbf{s}_0) < \lambda_{k+1} | \mathbf{Z}(\mathbf{s})) \\ &= \sum_{k=1}^K k \left[ \Phi \left( \frac{\lambda_{k+1} - \text{E}(Z(\mathbf{s}_0) | \mathbf{Z}(\mathbf{s}))}{\sqrt{\text{Var}(Z(\mathbf{s}_0) | \mathbf{Z}(\mathbf{s}))}} \right) - \Phi \left( \frac{\lambda_k - \text{E}(Z(\mathbf{s}_0) | \mathbf{Z}(\mathbf{s}))}{\sqrt{\text{Var}(Z(\mathbf{s}_0) | \mathbf{Z}(\mathbf{s}))}} \right) \right] \\ &= \text{E}(Y(\mathbf{s}_0) | \mathbf{Z}(\mathbf{s})) \end{aligned} \tag{48}$$

Thus the approximation proposed in 48 avoids the computationally intensive integrations required to predict  $Y(\mathbf{s}_0)$  given  $\mathbf{y}(\mathbf{s})$  based on (43) and (44).

To evaluate the approximation, we simulate ordinal response data with  $K = 5$  from the PLMM in (39) at 310 locations within the unit square where  $n = 300$  are considered observed observations and the remaining  $m = 10$  are unobserved locations for out-of-sample prediction. We assume an intercept and one covariate simulated from  $\text{Unif}(-1, 1)$  where we fix  $(\beta_0, \beta_1) = (3, 5)$ ,  $\tau = 1$ ,  $\phi = 0.2$ , and  $\boldsymbol{\lambda} = (-\infty, 0, 2, 4, 6, \infty)$ . For each of the  $m$  unobserved locations, we compute the expected value of  $Y(\mathbf{s}_0)$  using (42) and the multivariate integrals (43 and 44) and the approximation method (48) conditioned on the latent variables. Table (4.1) gives the expected value of  $Y$  at the unobserved locations using the two approaches. The estimates are very similar for all  $m$  unobserved locations. Therefore, we deem the approximation method a suitable alternative for estimating ordinal response data at unobserved locations for the spatial PLMM. Whereas the locations in this example

Table 4.1: The conditional expectation of  $Y$  at  $m = 10$  unobserved locations using the multivariate integral approach (42) and the approximation approach (48).

Location	True Value	Integral approach	Approximation approach
1	2	1.969	2.025
2	1	2.231	2.152
3	2	1.046	1.032
4	1	1.016	1.012
5	5	4.709	4.745
6	1	1.044	1.028
7	5	4.882	4.877
8	4	3.469	3.567
9	3	3.424	3.305
10	5	4.225	4.235

were simulated from a bivariate uniform distribution, the results may vary for other spatial designs.

The mean squared error (MSE) under the two prediction approaches for the 10 observed locations is also similar. The integral approach has a mean squared error (MSE) of 0.359 and the latent approach has MSE of 0.321. The approximation method appears to out-perform the multivariate integral method because the latent variable  $\mathbf{Z}$  used in (48) is assumed known. In practice however, the latent variable  $\mathbf{Z}$  will be estimated, leading to additional uncertainty. As future work we would like to compare the two methods in terms of their prediction error.

The main advantage of the approximation method is that computing the approximate conditional expectation of the unobserved response takes less than one-tenth the time of computing the multivariate integrals. The difference in computing time will increase as the number of observed locations increases. It is worth mentioning that the computational cost of the multivariate integral is not a product of the dense spatial covariance matrix. Thus, we also recommend the approximation method for predicting ordinal response data under the probit model with a sparse covariance matrix.

We must make a few comments regarding computing the  $E(Y(\mathbf{s}_0)|\mathbf{y}(\mathbf{s}))$  in practice. First, the estimates computed for the  $m = 10$  unobserved locations were done individually. Whereas it is common to jointly estimate the kriging estimate at more than one site simultaneously and is straightforward in the Gaussian case using the conditional distribution given in (46), joint estimation is not particularly useful for binary or ordinal response variables. In computing the conditional expectation of  $Y(\mathbf{s}_0)$  in (42), the marginal probability,  $P(Y(\mathbf{s}_0) = k|\mathbf{y}(\mathbf{s}))$ , is computed for  $k = 1, \dots, K$ . To estimate  $\mathbf{Y}(\mathbf{s}_0)$ , which is now a vector-valued random variable at unobserved locations, there are  $K^m$  marginal probabilities that could be computed. For example, we could compute the probability that  $Y(\mathbf{s}_0) = k$  for all  $m$  unobserved locations, which is rarely of interest even when  $m = 2$ . Therefore, we recommend predicting  $E(Y(\mathbf{s}_0)|\mathbf{y}(\mathbf{s}))$  using (42) or (47) separately for each unobserved location.

Second, the parameter values  $\boldsymbol{\beta}$ ,  $\boldsymbol{\lambda}$ ,  $\tau$  and  $\phi$  are unknown and need be estimated. In a frequentist context, the parameters might be estimated using maximum likelihood estimation. The estimates can then be plugged into either the multivariate integral equation (42) or its approximation (47) to obtain estimates of the binary or ordinal response variable at an unobserved location.

In the Bayesian context, MCMC can be used to obtain samples from the posterior distribution for each of the unknown parameters. We estimate the unobserved binary or ordinal response variable using the posterior predictive distribution. Computing  $E(Y(\mathbf{s}_0)|\mathbf{y}(\mathbf{s}))$  using our samples from the posterior distribution is computationally infeasible since it requires the multivariate integrals (43) and (44) to be evaluated for each iteration of the chain retained after burn-in. Therefore, the posterior predictive distribution of  $Y(\mathbf{s}_0)$  is traditionally obtained by sampling from the posterior predictive distribution of the latent continuous variable,  $Z(\mathbf{s}_0)$ . De Oliveira (1997, 2000) and Higgs and Hoeting (2010) give detailed information on how this is done for binary and ordinal response data, respectively. The traditional approach produces realizations of the posterior predictive distribution of the unobserved  $Y(\mathbf{s}_0)$ .

We can estimate  $E(Y(\mathbf{s}_0)|\mathbf{y}(\mathbf{s}))$  using the mean of the realizations of the posterior predictive distribution. Further, we can compute  $P(Y(\mathbf{s}_0) = k|\mathbf{y}(\mathbf{s}))$  for all ordinal categories  $k$  by

$$P(Y(\mathbf{s}_0) = k|\mathbf{y}(\mathbf{s})) = \frac{1}{M} \sum_{m=1}^M I_{[y_m(\mathbf{s}_0)=k]},$$

where  $y_m(\mathbf{s}_0)$  is the value of the  $m$ th realization from the posterior predictive distribution of  $Y(\mathbf{s}_0)$  and  $M$  is the number of iterations of the chain.

Notice that in the traditional approach,  $E(Y(\mathbf{s}_0)|\mathbf{y}(\mathbf{s}))$  is a point estimate computed from  $M$  realizations of the posterior predictive distribution. The same is true for the posterior probabilities  $P(Y(\mathbf{s}_0) = k|\mathbf{y}(\mathbf{s}))$  for each  $k$ . Had we been able to evaluate (42) using the multivariate integrals (43 and 44), we would have obtained  $M$  realizations of  $P(Y(\mathbf{s}_0) = k|\mathbf{y}(\mathbf{s}))$  for each  $k$ . This differs from the traditional approach because here we don't obtain realizations of the predictive distribution of  $Y(\mathbf{s}_0)$ . Rather, we get  $M$  realizations of the density function of  $Y(\mathbf{s}_0)$  given  $\mathbf{y}(\mathbf{s})$ , which we refer to as a distribution of a density functions. Using the density functions, we can compute  $E(Y(\mathbf{s}_0)|\mathbf{y}(\mathbf{s}))$  for  $m = 1, \dots, M$ . Thus, we obtain  $M$  realizations of the distribution of  $E(Y(\mathbf{s}_0)|\mathbf{y}(\mathbf{s}))$ . We can estimate the binary or ordinal random variable at an unobserved location with a point estimate *and* credible interval limits. This is invaluable as it quantifies the uncertainty in the estimate of the unobserved response which we are unable to do using the traditional approach. This highly motivates using (48) to approximate the multivariate integrals in the Bayesian framework. It allows us to obtain samples from the approximate posterior distribution of the density functions and  $E(Y(\mathbf{s}_0)|\mathbf{y}(\mathbf{s}))$  without the computational complexity of the multivariate integrals.

## 4.4 Limiting prediction distributions of first-stage and second-stage spatial models

We showed in Section 4.2 that the second-stage model performs equally well or better than the first-stage model in terms of parameter estimation regardless of the assumptions about

structured and unstructured stochasticity. The first-stage model was unable to estimate the spatial range parameter when the true model contained unstructured stochasticity. This implies that the first-stage model is more restrictive than the second-stage model. We also noticed that for large values of  $\tau$ , the empirical likelihood functions of the two models were similar. In this section, we compare the first-stage and second-stage model in terms of prediction at unobserved locations.

We compare the first-stage and second-stage spatial models introduced in Section 4.1 by their predicted values at unobserved locations using the approximation (48). The general expressions for the conditional expectation and variance of  $Z(\mathbf{s}_0)$  given the latent variable  $\mathbf{Z}(\mathbf{s})$  are

$$\begin{aligned} \text{E}(Z(\mathbf{s}_0)|\mathbf{Z}(\mathbf{s})) &= \mathbf{X}(\mathbf{s}_0)'\boldsymbol{\beta} + \Sigma_{01}\Sigma_{11}^{-1}(\mathbf{Z}(\mathbf{s}) - \mathbf{X}(\mathbf{s})\boldsymbol{\beta}) \quad \text{and} \\ \text{Var}(Z(\mathbf{s}_0)|\mathbf{Z}(\mathbf{s})) &= \Sigma_{00} - \Sigma_{01}\Sigma_{11}^{-1}\Sigma_{10}. \end{aligned} \tag{49}$$

The partition of the covariance matrix,  $\Sigma$ , under the first-stage spatial model is

$$\Sigma = \begin{bmatrix} \Sigma_{00} & \Sigma_{01} \\ \Sigma_{10} & \Sigma_{11} \end{bmatrix} = \begin{bmatrix} 1 & \mathbf{R}_{01} \\ \mathbf{R}_{10} & \mathbf{R}_{11} \end{bmatrix}.$$

Under the second-stage model, the covariance matrix is

$$\Sigma = \begin{bmatrix} \Sigma_{00} & \Sigma_{01} \\ \Sigma_{10} & \Sigma_{11} \end{bmatrix} = \begin{bmatrix} \tau + 1 & \tau\mathbf{R}_{01} \\ \tau\mathbf{R}_{10} & \tau\mathbf{R}_{11} + \mathbf{I} \end{bmatrix}.$$

Therefore, the conditional expectation and variance of  $Z(\mathbf{s}_0)$  under the first-stage spatial probit model are

$$\begin{aligned} E_1(Z(s_0)|\mathbf{Z}(s)) &= \mathbf{X}(s_0)'\boldsymbol{\beta} + \mathbf{R}_{01}\mathbf{R}_{11}^{-1}(\mathbf{Z}(s) - \mathbf{X}(s)'\boldsymbol{\beta}) \\ \text{Var}_1(Z(s_0)|\mathbf{Z}(s)) &= 1 - \mathbf{R}_{01}\mathbf{R}_{11}^{-1}\mathbf{R}_{10}. \end{aligned} \tag{50}$$

where the subscript denotes the stage of the spatial model. The conditional expectation and variance under the second-stage spatial probit model, or PLMM, are

$$\begin{aligned} E_2(Z(s_0)|\mathbf{Z}(s)) &= \mathbf{X}(s_0)'\boldsymbol{\beta} + \tau\mathbf{R}_{01}(\tau\mathbf{R}_{11} + \mathbf{I})^{-1}(\mathbf{Z}(s) - \mathbf{X}(s)'\boldsymbol{\beta}) \\ \text{Var}_2(Z(s_0)|\mathbf{Z}(s)) &= (\tau + 1) - \tau^2\mathbf{R}_{01}(\tau\mathbf{R}_{11} + \mathbf{I})^{-1}\mathbf{R}_{10}. \end{aligned} \tag{51}$$

We compare the first-stage and second-stage spatial models by comparing their kriging estimates at an unobserved location. For both models, the approximate conditional expectation (48) of the ordinal response,  $Y(s_0)$ , given the observed data  $\mathbf{y}(s)$  depends only on the conditional expectations and variances given in (50) and (51). Let  $\mathbf{U}$  represent the known covariate and location information of the observed data such that  $\mathbf{U} = (\mathbf{X}(s), s)$  where  $s = (s_1, \dots, s_n)$ . For the unobserved location,  $s_0$ , define  $\mathbf{U}_0 = (\mathbf{X}(s_0), s_0)$ . Let the parameter vector  $\boldsymbol{\theta}$  contain the set of model parameters  $\boldsymbol{\beta}$ ,  $\tau$ , and  $\phi$ , where  $\tau = 1$  under the first-stage spatial model. Let  $\mathbf{Z}_1(\boldsymbol{\theta}, \mathbf{U})$  and  $\mathbf{Z}_2(\boldsymbol{\theta}, \mathbf{U})$  be functions of the parameters and covariate information under the first-stage and second-stage model, respectively. The quantities of interest for the two models are the conditional expectations and variances given in (50) and (51) rewritten as,

$$E(Z_1(\boldsymbol{\theta}, \mathbf{U}_0)|\mathbf{Z}_1(\boldsymbol{\theta}, \mathbf{U})) \quad \text{and} \quad E(Z_2(\boldsymbol{\theta}, \mathbf{U}_0)|\mathbf{Z}_2(\boldsymbol{\theta}, \mathbf{U}))$$

and

$$\text{Var}(Z_1(\boldsymbol{\theta}, \mathbf{U}_0)|\mathbf{Z}_1(\boldsymbol{\theta}, \mathbf{U})) \quad \text{and} \quad \text{Var}(Z_2(\boldsymbol{\theta}, \mathbf{U}_0)|\mathbf{Z}_2(\boldsymbol{\theta}, \mathbf{U})).$$

The conditional expectations and variances are equivalent under certain limiting distributions of the spatial covariance parameters, leading us to the following lemmas and subsequent propositions.

**Lemma 1.** *Equivalent limiting distributions of the approximate conditional expectation and variance for the first-stage and second-stage model as  $\phi \rightarrow 0$  in the first-stage model.*

$$(a) \lim_{\phi \rightarrow 0} E(Z_1(\boldsymbol{\theta}, \mathbf{U}_0) | \mathbf{Z}_1(\boldsymbol{\theta}, \mathbf{U})) = \lim_{\phi \rightarrow 0} E(Z_2(\boldsymbol{\theta}, \mathbf{U}_0) | \mathbf{Z}_2(\boldsymbol{\theta}, \mathbf{U}))$$

$$(b) \lim_{\phi \rightarrow 0} E(Z_1(\boldsymbol{\theta}, \mathbf{U}_0) | \mathbf{Z}_1(\boldsymbol{\theta}, \mathbf{U})) = \lim_{\tau \rightarrow 0} E(Z_2(\boldsymbol{\theta}, \mathbf{U}_0) | \mathbf{Z}_2(\boldsymbol{\theta}, \mathbf{U}))$$

$$(c) \lim_{\phi \rightarrow 0} \text{Var}(Z_1(\boldsymbol{\theta}, \mathbf{U}_0) | \mathbf{Z}_1(\boldsymbol{\theta}, \mathbf{U})) = \lim_{\tau \rightarrow 0} \text{Var}(Z_2(\boldsymbol{\theta}, \mathbf{U}_0) | \mathbf{Z}_2(\boldsymbol{\theta}, \mathbf{U}))$$

**Proposition 1.** *The following propositions lead to the results of Lemma 1:*

$$(a) \lim_{\phi \rightarrow 0} E(Z_1(\boldsymbol{\theta}, \mathbf{U}_0) | \mathbf{Z}_1(\boldsymbol{\theta}, \mathbf{U})) = \mathbf{X}(\mathbf{s}_0)' \boldsymbol{\beta}$$

**Proof:** As  $\phi \rightarrow 0$ , the correlation between  $Z(\mathbf{s}_0)$  and  $\mathbf{Z}(\mathbf{s})$  goes to 0. Therefore,  $\mathbf{R}_{01} \rightarrow \mathbf{0}$  and  $\mathbf{R}_{11} \rightarrow \mathbf{I}$ , and the limit of the conditional expectation is

$$\begin{aligned} \lim_{\phi \rightarrow 0} E(Z_1(\boldsymbol{\theta}, \mathbf{U}_0) | \mathbf{Z}_1(\boldsymbol{\theta}, \mathbf{U})) &= \lim_{\phi \rightarrow 0} \{ \mathbf{X}(\mathbf{s}_0)' \boldsymbol{\beta} + \mathbf{R}_{01} \mathbf{R}_{11}^{-1} (\mathbf{Z}(\mathbf{s}) - \mathbf{X}(\mathbf{s})' \boldsymbol{\beta}) \} \\ &= \mathbf{X}(\mathbf{s}_0)' \boldsymbol{\beta} \end{aligned}$$

$$(b) \lim_{\phi \rightarrow 0} E(Z_2(\boldsymbol{\theta}, \mathbf{U}_0) | \mathbf{Z}_2(\boldsymbol{\theta}, \mathbf{U})) = \mathbf{X}(\mathbf{s}_0)' \boldsymbol{\beta}$$

**Proof:** Again, as  $\phi \rightarrow 0$ ,  $\mathbf{R}_{01} \rightarrow \mathbf{0}$  while  $\mathbf{R}_{11} \rightarrow \mathbf{I}$ . From (51), the limit of the conditional expectation for the second-stage model is

$$\begin{aligned} \lim_{\phi \rightarrow 0} E(Z_2(\boldsymbol{\theta}, \mathbf{U}_0) | \mathbf{Z}_2(\boldsymbol{\theta}, \mathbf{U})) &= \lim_{\phi \rightarrow 0} \{ \mathbf{X}(\mathbf{s}_0)' \boldsymbol{\beta} + \tau \mathbf{R}_{01} (\tau \mathbf{R}_{11} + \mathbf{I})^{-1} (\mathbf{Z}(\mathbf{s}) - \mathbf{X}(\mathbf{s})' \boldsymbol{\beta}) \} \\ &= \mathbf{X}(\mathbf{s}_0)' \boldsymbol{\beta} \end{aligned}$$

$$(c) \lim_{\tau \rightarrow 0} E(Z_2(\boldsymbol{\theta}, \mathbf{U}_0) | \mathbf{Z}_2(\boldsymbol{\theta}, \mathbf{U})) = \mathbf{X}(\mathbf{s}_0)' \boldsymbol{\beta}$$

**Proof:** As  $\tau \rightarrow 0$ ,  $\tau \mathbf{R}_{01} \rightarrow \mathbf{0}$  and  $(\tau \mathbf{R}_{11} + \mathbf{I}) \rightarrow \mathbf{I}$ . Therefore,

$$\begin{aligned} \lim_{\tau \rightarrow 0} E(Z_2(\boldsymbol{\theta}, \mathbf{U}_0) | \mathbf{Z}_2(\boldsymbol{\theta}, \mathbf{U})) &= \lim_{\tau \rightarrow 0} \{ \mathbf{X}(\mathbf{s}_0)' \boldsymbol{\beta} + \tau \mathbf{R}_{01} (\tau \mathbf{R}_{11} + \mathbf{I})^{-1} (\mathbf{Z}(s) - \mathbf{X}(s)' \boldsymbol{\beta}) \} \\ &= \mathbf{X}(\mathbf{s}_0)' \boldsymbol{\beta} \end{aligned}$$

$$(d) \lim_{\phi \rightarrow 0} \text{Var}(Z_1(\boldsymbol{\theta}, \mathbf{U}_0) | \mathbf{Z}_1(\boldsymbol{\theta}, \mathbf{U})) = 1$$

**Proof:** As  $\phi \rightarrow 0$ ,  $\mathbf{R}_{01} = \mathbf{R}'_{10} \rightarrow \mathbf{0}$  while  $\mathbf{R}_{11} \rightarrow \mathbf{I}$ . Therefore, from (50), the limit of the conditional variance is

$$\begin{aligned} \lim_{\phi \rightarrow 0} \text{Var}(Z_1(\boldsymbol{\theta}, \mathbf{U}_0) | \mathbf{Z}_1(\boldsymbol{\theta}, \mathbf{U})) &= \lim_{\phi \rightarrow 0} \{ 1 - \mathbf{R}_{01} \mathbf{R}_{11}^{-1} \mathbf{R}_{10} \} \\ &= 1. \end{aligned}$$

$$(e) \lim_{\phi \rightarrow 0} \text{Var}(Z_2(\boldsymbol{\theta}, \mathbf{U}_0) | \mathbf{Z}_2(\boldsymbol{\theta}, \mathbf{U})) = \tau + 1$$

**Proof:** As  $\phi \rightarrow 0$ ,  $\mathbf{R}_{01} = \mathbf{R}'_{10} \rightarrow \mathbf{0}$  while  $\mathbf{R}_{11} \rightarrow \mathbf{I}$ . Therefore, from (51), the limit of the conditional variance is

$$\begin{aligned} \lim_{\phi \rightarrow 0} \text{Var}(Z_2(\boldsymbol{\theta}, \mathbf{U}_0) | \mathbf{Z}_2(\boldsymbol{\theta}, \mathbf{U})) &= \lim_{\phi \rightarrow 0} \{ (\tau + 1) - \tau^2 \mathbf{R}_{01} (\tau \mathbf{R}_{11} + \mathbf{I})^{-1} \mathbf{R}_{10} \} \\ &= \tau + 1. \end{aligned}$$

$$(f) \lim_{\tau \rightarrow 0} \text{Var}(Z_2(\boldsymbol{\theta}, \mathbf{U}_0) | \mathbf{Z}_2(\boldsymbol{\theta}, \mathbf{U})) = 1$$

**Proof:** From Proposition 1(e),  $\phi \rightarrow 0$  does not result in the same limiting distribution of the conditional variance under the second-stage model as the first stage model. As  $\tau \rightarrow 0$ , however,  $\tau \mathbf{R}_{01} \rightarrow \mathbf{0}$  and  $(\tau \mathbf{R}_{11} + \mathbf{I}) \rightarrow \mathbf{I}$ . Therefore, the limit of the conditional variance is

$$\begin{aligned} \lim_{\phi \rightarrow 0} \text{Var}(Z_2(\boldsymbol{\theta}, \mathbf{U}_0) | \mathbf{Z}_2(\boldsymbol{\theta}, \mathbf{U})) &= \lim_{\tau \rightarrow 0} \{ (\tau + 1) - \tau^2 \mathbf{R}_{01} (\tau \mathbf{R}_{11} + \mathbf{I})^{-1} \mathbf{R}_{10} \} \\ &= 1. \end{aligned}$$

The proof of Proposition 1(f) does not depend on the value of  $\phi$  indicating that  $\tau \rightarrow 0$  is sufficient for equivalence in the conditional variances of the two models in Lemma 1(c).

The simulation results of Section 4.2 indicated that the first-stage and second-stage model become more similar in terms of estimating  $\phi$  as  $\tau$  increases in the second-stage model. This is because the nugget is fixed to 1 for identifiability of the probit model and will become negligible relative to  $\tau$ , the spatial component of the variance, as  $\tau$  increases. Our investigation of  $\tau$  lead us to the following Lemma:

**Lemma 2.** *Limiting distributions of the approximate conditional expectation and variance of the second-stage model as  $\tau \rightarrow \infty$ .*

$$(a) \lim_{\tau \rightarrow \infty} E(Z_2(\boldsymbol{\theta}, \mathbf{U}_0) | \mathbf{Z}_2(\boldsymbol{\theta}, \mathbf{U})) = E(Z_1(\boldsymbol{\theta}, \mathbf{U}_0) | \mathbf{Z}_1(\boldsymbol{\theta}, \mathbf{U}))$$

**Proof:** As  $\tau \rightarrow \infty$ ,  $(\tau \mathbf{R}_{11} + \mathbf{I}) \rightarrow \tau \mathbf{R}_{11}$ . Therefore,

$$\begin{aligned} \lim_{\tau \rightarrow \infty} E(Z_2(\boldsymbol{\theta}, \mathbf{U}_0) | \mathbf{Z}_2(\boldsymbol{\theta}, \mathbf{U})) &= \lim_{\tau \rightarrow \infty} \{ \mathbf{X}(\mathbf{s}_0)' \boldsymbol{\beta} + \tau \mathbf{R}_{01} (\tau \mathbf{R}_{11} + \mathbf{I})^{-1} (\mathbf{Z}(\mathbf{s}) - \mathbf{X}(\mathbf{s})' \boldsymbol{\beta}) \} \\ &= \lim_{\tau \rightarrow \infty} \{ \mathbf{X}(\mathbf{s}_0)' \boldsymbol{\beta} + \tau \mathbf{R}_{01} (\tau \mathbf{R}_{11})^{-1} (\mathbf{Z}(\mathbf{s}) - \mathbf{X}(\mathbf{s})' \boldsymbol{\beta}) \} \\ &= \mathbf{X}(\mathbf{s}_0)' \boldsymbol{\beta} + \mathbf{R}_{01} \mathbf{R}_{11}^{-1} (\mathbf{Z}(\mathbf{s}) - \mathbf{X}(\mathbf{s})' \boldsymbol{\beta}) \\ &= E(Z_1(\boldsymbol{\theta}, \mathbf{U}_0) | \mathbf{Z}_2(\boldsymbol{\theta}, \mathbf{U})) \end{aligned}$$

$$(b) \lim_{\tau \rightarrow \infty} \text{Var}(Z_2(\boldsymbol{\theta}, \mathbf{U}_0) | \mathbf{Z}_2(\boldsymbol{\theta}, \mathbf{U})) = \infty$$

**Proof:** For fixed  $\phi$ ,

$$\begin{aligned} \lim_{\tau \rightarrow \infty} \text{Var}(Z_2(\boldsymbol{\theta}, \mathbf{U}_0) | \mathbf{Z}_2(\boldsymbol{\theta}, \mathbf{U})) &= \lim_{\tau \rightarrow \infty} \{ (\tau + 1) - \tau^2 \mathbf{R}_{01} (\tau \mathbf{R}_{11} + \mathbf{I})^{-1} \mathbf{R}_{10} \} \\ &= \lim_{\tau \rightarrow \infty} \{ 1 + \tau(1 - \mathbf{R}_{01} \mathbf{R}_{11}^{-1} \mathbf{R}_{10}) \} \\ &= \infty. \end{aligned}$$

Therefore, the approximate conditional expectation is equivalent under the two models as  $\tau \rightarrow \infty$ . However, the conditional variance in the second-stage model goes to  $\infty$  as  $\tau \rightarrow \infty$ .

To estimate the ordinal response variable at a new location given the observed data using the approximation (48), we need to compute

$$\mathbb{E}(Y(\mathbf{s}_0)|\mathbf{y}(\mathbf{s})) \approx \sum_{k=1}^K kP(\lambda_k \leq Z(\mathbf{s}_0) < \lambda_{k+1}|\mathbf{Z}(\mathbf{s})).$$

If  $P(\lambda_k \leq Z(\mathbf{s}_0) < \lambda_{k+1}|\mathbf{Z}(\mathbf{s}))$  for each  $k = 1, \dots, K$  under two different models, we say that the models are equivalent in their approximate prediction of  $Y(\mathbf{s}_0)$ .

**Theorem 1.** *Let the spatial range parameter,  $\phi$ , and the coefficients,  $\beta_1, \dots, \beta_p$ , be equal under the first-stage and second-stage model. Define  $\beta_{01}$  and  $\boldsymbol{\lambda}_1$  as the intercept and threshold vector of the first-stage model, and assume  $\tau_1 = 1$  for identifiability. There exists an intercept,  $\beta_{02}$ , threshold vector,  $\boldsymbol{\lambda}_2$ , and partial sill,  $\tau_2 < \infty$ , under the second-stage model such that*

$$E(Y_2(\mathbf{s}_0)|\mathbf{Z}_2(\mathbf{s})) = E(Y_1(\mathbf{s}_0)|\mathbf{Z}_1(\mathbf{s})).$$

This is an important results for two reasons. First, the approximated kriging estimate at an unobserved location given the sample data will be the same under the two models. Second, the models result in the same parameter inference for the coefficients of the fixed effects and the spatial range. To prove Theorem 1, we first assume that the parameters of the first-stage model,  $\boldsymbol{\beta}_1$ ,  $\phi_1$ ,  $\boldsymbol{\lambda}_1$ , and  $\tau_1$ , are fixed and known where  $\boldsymbol{\beta}_1 = (\beta_{01}, \beta_{11}, \dots, \beta_{p1})$ ,  $\boldsymbol{\lambda}_1 = (\lambda_{11}, \dots, \lambda_{(K+1)1})$ , and  $\tau = 1$ . We will show that the second-stage spatial model has equivalent prediction when  $(\beta_{12}, \dots, \beta_{p2}) = (\beta_{11}, \dots, \beta_{p1})$  and  $\phi_2 = \phi_1$ . The conditional expectation of  $Z(\mathbf{s}_0)$  is equivalent under the first-stage and second-stage model as  $\tau \rightarrow \infty$  (Lemma 2(a)). This implies that for  $\epsilon > 0$ , there exists  $\tau_\epsilon < \infty$  such that

$$|\mathbb{E}(Z_2(\boldsymbol{\theta}_2, \mathbf{U}_0)|\mathbf{Z}_2(\boldsymbol{\theta}_2, \mathbf{U})) - \mathbb{E}(Z_1(\boldsymbol{\theta}_1, \mathbf{U}_0)|\mathbf{Z}_1(\boldsymbol{\theta}_1, \mathbf{U}))| < \epsilon \quad (52)$$

where  $\boldsymbol{\theta}_2 = (\boldsymbol{\beta}, \tau = \tau_\epsilon, \boldsymbol{\lambda})$  and  $\boldsymbol{\theta}_1 = (\boldsymbol{\beta}, \tau = 1, \boldsymbol{\lambda})$ . This is important step since under the second-stage model the conditional variance of  $Z(\mathbf{s}_0) \rightarrow \infty$  as  $\tau \rightarrow \infty$  (Lemma 2(b)). The conditional variance under the second-stage model is finite by setting  $\tau = \tau_\epsilon$ .

To show equivalence in prediction of the two models, we need

$$\Phi\left(\frac{\lambda_k - \text{E}(Z_2(\boldsymbol{\theta}_2, \mathbf{U}_0) | \mathbf{Z}_2(\boldsymbol{\theta}_2, \mathbf{U}))}{\sqrt{\text{Var}(Z_2(\boldsymbol{\theta}_2, \mathbf{U}_0) | \mathbf{Z}_2(\boldsymbol{\theta}_2, \mathbf{U}))}}\right) = \Phi\left(\frac{\lambda_k - \text{E}(Z_1(\boldsymbol{\theta}_1, \mathbf{U}_0) | \mathbf{Z}_1(\boldsymbol{\theta}_1, \mathbf{U}))}{\sqrt{\text{Var}(Z_1(\boldsymbol{\theta}_1, \mathbf{U}_0) | \mathbf{Z}_1(\boldsymbol{\theta}_1, \mathbf{U}))}}\right) \quad (53)$$

for  $k = 1, \dots, K+1$ . We define the intercept and threshold vector of the second-stage model as

$$\begin{aligned} \beta_{02} &= \frac{\sqrt{\text{Var}_2}}{\sqrt{\text{Var}_1}} E_1 - (\beta_1 X_1(\mathbf{s}_0) + \dots + \beta_p X_p(\mathbf{s}_0)) \\ &\quad - \tau \mathbf{R}_{01} (\tau \mathbf{R}_{11} + \mathbf{I})^{-1} (\mathbf{Z}(\mathbf{s}) - (\beta_1 \mathbf{X}_1(\mathbf{s}) + \dots + \beta_p \mathbf{X}_p(\mathbf{s}))) \end{aligned} \quad (54)$$

and

$$\lambda_{k2} = \lambda_{k1} \frac{\sqrt{\text{Var}_2}}{\sqrt{\text{Var}_1}} \quad (55)$$

for  $k = 1, \dots, K+1$ , where  $m$  denotes the stage of the spatial model,  $\tau = \tau_\epsilon$ ,  $\text{E}_m = \text{E}(Z_m(\boldsymbol{\theta}_m, \mathbf{U}_0) | \mathbf{Z}_m(\boldsymbol{\theta}_m, \mathbf{U}))$ , and  $\text{Var}_m = \text{Var}(Z_m(\boldsymbol{\theta}_m, \mathbf{U}_0) | \mathbf{Z}_m(\boldsymbol{\theta}_m, \mathbf{U}))$ . Therefore, we compute

$$P(Z_2(\boldsymbol{\theta}_2, \mathbf{U}_0) < \lambda_{k2} | \mathbf{Z}_2(\boldsymbol{\theta}_2, \mathbf{U}))$$

for  $k = 1, \dots, K$  by plugging in (54) and (55) into (53). Starting with plugging (55) in for  $\lambda_{k2}$ , we have

$$\begin{aligned}
P(Z_2(\boldsymbol{\theta}_2, \mathbf{U}_0) < \lambda_{k2} | \mathbf{Z}_2(\boldsymbol{\theta}_2, \mathbf{U})) &= \Phi\left(\frac{\lambda_{k2} - E_2}{\sqrt{\text{Var}_2}}\right) \\
&= \Phi\left(\frac{\lambda_{k1}\frac{\sqrt{\text{Var}_2}}{\sqrt{\text{Var}_1}} - E_2}{\sqrt{\text{Var}_2}}\right) \\
&= \Phi\left(\frac{1}{\sqrt{\text{Var}_2}}\left(\lambda_{k1}\frac{\sqrt{\text{Var}_2}}{\sqrt{\text{Var}_1}} - (\beta_{02} + \beta_1 X_1(\mathbf{s}_0) + \cdots + \beta_p X_p(\mathbf{s}_0))\right.\right. \\
&\quad \left.\left. + \tau \mathbf{R}_{01}(\tau \mathbf{R}_{11} + \mathbf{I})^{-1}(\mathbf{Z}(\mathbf{s}) - (\beta_{02}\mathbf{1} + \beta_1 \mathbf{X}_1(\mathbf{s}) + \cdots + \beta_p \mathbf{X}_p(\mathbf{s})))\right)\right).
\end{aligned}$$

Then, plugging (54) in for  $\beta_{02}$ ,

$$\begin{aligned}
P(Z_2(\boldsymbol{\theta}_2, \mathbf{U}_0) < \lambda_{k2} | \mathbf{Z}_2(\boldsymbol{\theta}_2, \mathbf{U})) \\
&= \Phi\left(\frac{1}{\sqrt{\text{Var}_2}}\left(\lambda_{k1}\frac{\sqrt{\text{Var}_2}}{\sqrt{\text{Var}_1}} - \frac{\sqrt{\text{Var}_2}}{\sqrt{\text{Var}_1}}E_1\right.\right. \\
&\quad \left.\left. - (\beta_1 X_1(\mathbf{s}_0) + \cdots + \beta_p X_p(\mathbf{s}_0)) - \tau \mathbf{R}_{01}(\tau \mathbf{R}_{11} + \mathbf{I})^{-1}(\mathbf{Z}(\mathbf{s}) - (\beta_1 \mathbf{X}_1(\mathbf{s}) + \cdots + \beta_p \mathbf{X}_p(\mathbf{s})))\right.\right. \\
&\quad \left.\left. + (\beta_1 X_1(\mathbf{s}_0) + \cdots + \beta_p X_p(\mathbf{s}_0)) + \tau \mathbf{R}_{01}(\tau \mathbf{R}_{11} + \mathbf{I})^{-1}(\mathbf{Z}(\mathbf{s}) - (\beta_1 \mathbf{X}_1(\mathbf{s}) + \cdots + \beta_p \mathbf{X}_p(\mathbf{s})))\right)\right) \\
&= \Phi\left(\frac{1}{\sqrt{\text{Var}_2}}\left(\lambda_{k1}\frac{\sqrt{\text{Var}_2}}{\sqrt{\text{Var}_1}} - \frac{\sqrt{\text{Var}_2}}{\sqrt{\text{Var}_1}}E_1\right.\right. \\
&\quad \left.\left. - (\beta_1 X_1(\mathbf{s}_0) + \cdots + \beta_p X_p(\mathbf{s}_0)) - \tau \mathbf{R}_{01}(\tau \mathbf{R}_{11} + \mathbf{I})^{-1}(\mathbf{Z}(\mathbf{s}) - (\beta_1 \mathbf{X}_1(\mathbf{s}) + \cdots + \beta_p \mathbf{X}_p(\mathbf{s})))\right.\right. \\
&\quad \left.\left. + (\beta_1 X_1(\mathbf{s}_0) + \cdots + \beta_p X_p(\mathbf{s}_0)) + \tau \mathbf{R}_{01}(\tau \mathbf{R}_{11} + \mathbf{I})^{-1}(\mathbf{Z}(\mathbf{s}) - (\beta_1 \mathbf{X}_1(\mathbf{s}) + \cdots + \beta_p \mathbf{X}_p(\mathbf{s})))\right)\right).
\end{aligned}$$

The last two lines cancel and we are left with

$$\Phi\left(\frac{\lambda_{k2} - E_2}{\sqrt{\text{Var}_2}}\right) = \Phi\left(\frac{\lambda_{k1} - E_1}{\sqrt{\text{Var}_2}}\right).$$

Therefore,

$$P(\lambda_{k2} \leq Z_2(\mathbf{s}_0) < \lambda_{(k+1)2} | \mathbf{Z}_2(\mathbf{s})) = P(\lambda_{k1} \leq Z_1(\mathbf{s}_0) < \lambda_{(k+1)1} | \mathbf{Z}_1(\mathbf{s}))$$

for all  $k$  which implies

$$\text{E}(Y_2(\mathbf{s}_0) | \mathbf{Z}_2(\mathbf{s})) = \text{E}(Y_1(\mathbf{s}_0) | \mathbf{Z}_1(\mathbf{s})).$$

Theorem 1 holds for binary data as well where both threshold vectors,  $\boldsymbol{\lambda}_1$  and  $\boldsymbol{\lambda}_2$  are fixed and only the intercept terms differ.

Whereas we assume an exponential covariance function for the spatially structured stochasticity in the model, the theoretical results of this section hold for most isotropic parametric covariance functions. One notable exception is the spherical covariance function. Lemma 1 does not hold for the spherical covariance function because  $\mathbf{R}_{01} \rightarrow \mathbf{1}$  and  $\mathbf{R}_{11} \rightarrow \mathbf{J}$  as  $\phi \rightarrow 0$ . However, both Lemma 2 and Theorem 1 do hold for the spherical covariance function.

## 4.5 Discussion

The simulation results in Section 4.2 and the theoretical results in Section 4.4 both indicate that the second-stage spatial model, or spatial PLMM, is more flexible than the first-stage model for ordinal data. This is because the spatial PLMM allows for both spatially-correlated and independent components of stochasticity in the ordinal, or binary, response. In practice, estimation and parameter inference under the first-stage model can be mimicked by the PLMM where the converse is not true. This was illustrated in Section 4.2 when the first-stage model poorly estimated the spatial range parameter,  $\phi$ , when the data were generated under the second-stage model. However, when the data were generated under the first-stage model, the second stage model was able to capture the spatial range parameter equally well as the first-stage model. We noticed in the log-likelihood surfaces (Figure 4.4) for binary response data that the estimate of the partial sill parameter,  $\tau$ , was large to compensate for the data being generated under no-nugget first-stage model. Our simulation results directly correspond to our theoretical results in Section 4.4 since for large values of

$\tau$ , the approximate prediction and parameter inference under the second-stage spatial model can be equivalent to that of the first-stage spatial model.

This analysis motivates an important recommendation for those modeling binary and ordinal data. We advocate that the first-stage spatial model is only appropriate when it can be assumed that the binary or ordinal spatial response surface contains only spatially-correlated stochasticity. Even then, the second-stage spatial model can perform equally well in estimating the range parameter as the first-stage spatial model and make comparable predictions of the ordinal response variable at an unobserved location. Therefore, we recommend fitting the more general flexible second-stage spatial probit model because it is equally good or better than the first-stage spatial probit model for spatially-correlated binary and ordinal data.

## CHAPTER 5

# MULTILEVEL LATENT GAUSSIAN PROCESS MODEL FOR MIXED DISCRETE AND CONTINUOUS MULTIVARIATE RESPONSE DATA

### 5.1 Introduction

Latent variable modeling has become common practice in a variety of scientific research fields where the latent variables are not directly observed but instead inferred from other values that are observed. These models are particularly relevant when the observed data are assumed to be driven by some underlying, unobservable process. Oftentimes in the biological and ecological sciences, for example, multiple measurements are reported for each sampling unit or at each sampled location within a spatial domain and the goal is to understand the underlying latent variable(s) generating the measurements. Here, these measurements make up a multivariate response. In spatial statistics, a latent variable could be used to model a random field, or process. Chakraborty et al. (2010) applied a latent spatial process model to model species abundance across a large region of South Africa. Christensen and Amemiya (2002) developed a general framework for multivariate latent variable models that incorporates spatial correlation among the latent variables.

We focus on ordered categorical, or ordinal data where measurements for each observation are reported on a specified scale, (e.g., low, medium, high). Some discrete data are ordinal in nature. For example, in survey data, respondents are asked to characterize their opinions on a Likert scale ranging from strongly disagreeing to strongly agreeing. In other situations, data will be ordinal when a researcher reports the response as a discretized continuous variable instead of as the actual continuous variable due to constraints on the data collection process.

This may be the case when reporting sediment size in streams or surface area of leaves on individual plants, especially when the data are to be collected over a large spatial domain.

We propose a model for drawing inference about mixed ordinal and continuous multivariate response data. We refer to the model as a multilevel latent process model because we introduce latent variables at two levels within the hierarchy. The first level of latency is introduced by assuming there is a continuous latent process that generates each variable of the multivariate response. The model extends the multivariate latent health factor model proposed by Chiu et al. (2011) by allowing dependence on the site effect to vary across response variables.

The second level of latency is introduced by assuming there exists an underlying univariate latent spatial process, or latent random field, that is generating the multivariate response. We assume a linear relationship between each of the latent continuous response processes (first level of latency) and the latent spatial process (second level of latency). Refer to Figure 5.1 for a diagram of the multilevel latency. This model provides estimates of the latent spatial process in order to compare different locations within a specified region of interest. Second, the model allows quantification of the relationship between the spatial latent variable and each of the variables of the multivariate response. Lastly, we can determine which of the variables of the multivariate response are most closely associated with the latent spatial process. In doing so, we can establish weights for each of the response variables to be used in weighted averaging for estimating the underlying latent spatial process. By incorporating point-referenced covariate information, we can predict the value for the latent spatial variable as well as the mixed ordinal and continuous multivariate response at new locations.

In Section 5.2 we motivate the model with an application of assessing the condition of wetlands in Colorado. In Section 5.3 we introduce the mixed ordinal and continuous multivariate latent Gaussian process model; we also describe methods of inference and estimation of the model parameters under the Bayesian framework. In Section 5.4 we develop methods

to predict the latent random variable and ranking procedures for the multivariate response. The methodology is applied in Section 5.5 through the evaluation of wetland condition in the North Platte and Rio Grande River Basins of Colorado. Section 5.6 concludes with a brief discussion and recommendations for future work.

## 5.2 Motivating example

The proposed model was motivated by a program to assess the condition of wetlands in Colorado. Limited data exist on the location, type, and condition of Colorado's wetlands hindering wetland management. The long-term viability and integrity of Colorado's wetland resources are threatened due to increased demand from major urban areas for water development and storage projects, growth in the oil and gas industry, and changes in forest health (Dahl, 2011). The data considered here were collected through a partnership between Colorado Parks and Wildlife (CPW)'s Wetlands Program and the Colorado Natural Heritage Program (CNHP) to assess the condition of wetlands in Colorado. The specific data used in this model were collected in Colorado's North Platte and Rio Grande River Basins (Lemly et al., 2011; Lemly and Gillian, 2012). One of the major goals of the CPW-CNHP partnership is to model the spatial distribution of wetland ecological condition throughout each river basin in the state. Our goal was to improve spatial modeling techniques in order to help land managers effectively maintain and improve critical wetland habitats.

In order to implement effective wetland protection strategies and to establish restoration and management plans, wetlands must be assessed and then potential threats or stressors identified. There are many different in-field measurements, known as metrics, that reflect various aspects of wetland condition. These metrics can be of any variable type including continuous, count data, ordinal, etc. Overall scores that are computed based on multiple measurements are referred to as multi-metric indices. When the metrics are of the same variable type, one index to evaluate overall wetland condition is an average metric score. However, difficulty arises when trying to compute an index that encompasses metrics of dif-

ferent variable types. In this work, we propose using continuous latent variables as consistent measures across all metric types. Appropriate link functions can map the continuous latent variables to the different metrics.

One popular index that incorporates 12 metrics to evaluate ecological condition is the index of biotic integrity, or IBI (Karr, 1981). It is of great interest to ecologists to determine whether the particular metrics that are used in computing the IBI are useful in evaluating wetland condition. Of equal importance, ecologists are interested in identifying which of the measurements taken during in-field data collection are most representative of overall wetland condition. This is beneficial as it will not only increase accuracy in gauging wetland condition but will also save time and resources for future data collection by requiring fewer measurements.

There are tens of thousands of acres of reported wetlands in Colorado's North Platte and Rio Grande River Basins and sampling time and resources are limited. One of the major goals of the wetland profiling project is to model the spatial distribution of the ecological condition of wetlands throughout the basins and determine the optimal metrics for measuring key habitat features for wetland-dependent wildlife species. We compare the ecological condition of the wetlands based on five metrics in both the North Platte and Rio Grande River Basins.

## 5.3 Model and inference

### 5.3.1 Multivariate mixed response data

One of the main goals of this work is to use observed mixed ordinal and continuous multivariate responses from a finite number of point-referenced locations to draw inference on an underlying latent spatial process. We wish to make predictions of the latent spatial process as well as quantify uncertainty. The model consists of first representing each of the multivariate response variables as a continuous response. For the ordinal response variables, this continuous response is latent. We then define a linear relationship between each of the

(latent) continuous response variables and the underlying latent spatial process of interest. We assume that each of the response variables contains information about this latent spatial process. Refer to Figure 5.1 for a diagram of the multilevel latent model.

For the spatial domain of interest,  $D$ , define  $\{\mathbf{Y}(\mathbf{s}) = [Y_1(\mathbf{s}), \dots, Y_J(\mathbf{s})], \mathbf{s} \in D\}$  as a mixed ordinal and continuous multivariate random field at location  $\mathbf{s}$  having  $J$  responses. Each response at location  $\mathbf{s}$ ,  $\{Y_j(\mathbf{s}), \mathbf{s} \in D\}$  for  $j = 1, \dots, J$  is modeled by a random field of either continuous or ordinal values. Let  $J_c$  denote the number of continuous response variables and  $J_o$  denote the number of ordinal response variables, where  $J_o \geq 1$ . Therefore,  $J = J_o + J_c$ . For all ordinal variables variables  $j$  in  $1, \dots, J_o$ , the observable response  $Y_j(\mathbf{s}) \in \{1, \dots, K\}$  for every location  $\mathbf{s}$ . The model can easily be generalized to include observable response variables with varying number of categories, e.g.  $Y_j(\mathbf{s}) \in \{1, \dots, K_j\}$ . In such a case, parameter constraints, discussed below, will need to be modified to maintain model identifiability.

We assume there exists an underlying continuous multivariate Gaussian process,  $\{\mathbf{Z}(\mathbf{s}) = [Z_1(\mathbf{s}), \dots, Z_J(\mathbf{s})], \mathbf{s} \in D\}$ , that over the region of interest is generating  $\mathbf{Y}(\mathbf{s})$ . Dropping the dependence on  $\mathbf{s}$  for ease of notation, we denote  $\mathbf{Y} = [\mathbf{Y}_1, \dots, \mathbf{Y}_J]$  and  $\mathbf{Z} = [\mathbf{Z}_1, \dots, \mathbf{Z}_J]$  where respectively  $\mathbf{Y}_j$  and  $\mathbf{Z}_j$  are the  $j$ th observable response and underlying continuous Gaussian process. For  $j = 1, \dots, J$ , we define  $F_j$  as the mapping of the continuous variable  $\mathbf{Z}_j$  to the observable response  $\mathbf{Y}_j$ . Whereas the observable response data presented in this work are continuous and ordinal, the model holds for other types of response variables, e.g. binary, Poisson, etc. The mapping function  $F_j$  can take on any form as long as it is reasonable to assume that an underlying continuous Gaussian process is generating the response. For an ordinal response, the continuous variable  $\mathbf{Z}_j$  is latent. Here, the mapping  $F_j$  is defined as a function with parameter vector  $\boldsymbol{\lambda}_j$ , a  $((K + 1) \times 1)$ -dimensional vector of thresholds, that assigns the latent continuous random variables  $\mathbf{Z}_j$  to the ordered categories  $1, \dots, K$  of the observable data  $\mathbf{Y}_j$  (Muthen, 1984). The threshold parameter vector is constrained such that  $-\infty = \lambda_{j,0} \leq \lambda_{j,1} \leq \dots \lambda_{j,K} = \infty$  for each ordinal metric. We define a mapping,

$F_j$ , of  $Z_j(\mathbf{s})$  to  $Y_j(\mathbf{s})$  as

$$Y_j(\mathbf{s}) = F_j(Z_j(\mathbf{s}), \boldsymbol{\lambda}_j) = \sum_{k=1}^K k I_{\{\lambda_{j,k-1} < Z_j(\mathbf{s}) \leq \lambda_{j,k}\}}, \quad j = 1, \dots, J_o, \quad \mathbf{s} \in D. \quad (56)$$

For continuous response variables, the mapping  $F_j$  is taken as the identity function since  $\mathbf{Z}_j$  would be observed directly.

### 5.3.2 Multilevel latency

We assume that the latent random process is expressed by a mixed model. For the  $j$ th random process,  $\mathbf{Z}_j$ , we assume a multivariate Gaussian process where

$$\mathbf{Z}_j \sim GP(\theta_j \mathbf{1} + \omega_j \mathbf{H}, \sigma_j^2 \mathbf{I}). \quad (57)$$

We define the mean of each  $\mathbf{Z}_j$  as a metric-specific linear combination of the 1-vector and a latent random field  $\mathbf{H}$ . The latent random field  $\mathbf{H}$  is the process of interest and encompasses the latent measure of wetland condition in our application. The fixed effect  $\theta_j$  is the intercept for metric  $j$  and the fixed effect  $\omega_j$  is the factor loading of the spatial random field  $\mathbf{H}$ . Both  $\boldsymbol{\theta}$  and  $\boldsymbol{\omega}$  are  $(1 \times J)$ -dimensional vectors. The parameter  $\boldsymbol{\omega}$  allows us to quantify the relationship between each of the response variables and  $\mathbf{H}$ . The variance of  $\mathbf{Z}_j$  is specific to each metric  $j$ , which we define as  $\sigma_j^2 \mathbf{I}$  where  $\mathbf{I}$  is the identity matrix. For  $j \neq l$ ,  $\mathbf{Z}_j$  and  $\mathbf{Z}_l$  are conditionally independent given  $\mathbf{H}$ ,  $\boldsymbol{\theta}$ , and  $\boldsymbol{\omega}$ .

The spatial dependence of the multivariate random field is modeled through the latent spatial process,  $\mathbf{H}$ . Note that the inclusion of the additional latent process  $\mathbf{H}$  makes this a multilevel latent process model. We assume this latent spatial process is driving the mixed ordinal and continuous multivariate observable response,  $\mathbf{Y}$ . Therefore,  $\mathbf{H}$  provides a univariate summary measure for each location from which we will draw inference across space. We assume  $\mathbf{H}$  to be a Gaussian process with covariates in the mean structure and a

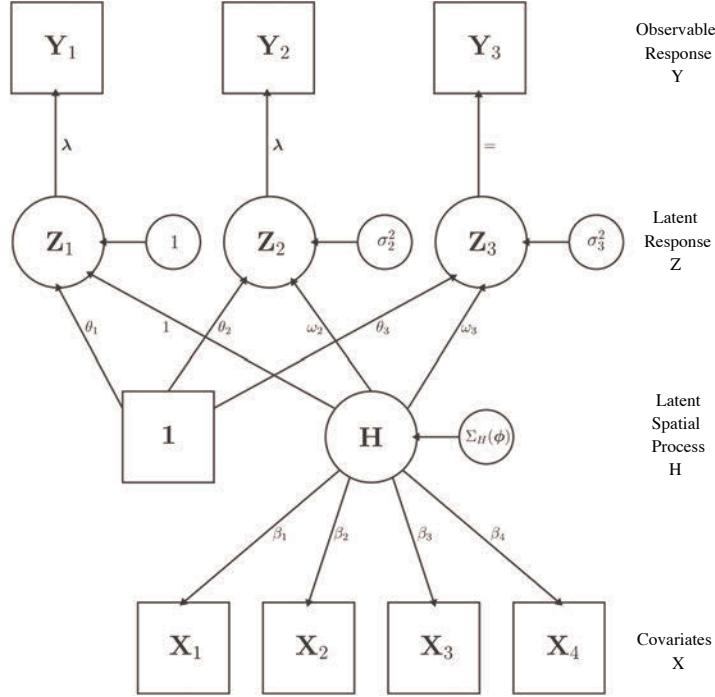


Figure 5.1: Diagram of a multilevel latent model with two ordinal observable response variables  $\mathbf{Y}_1$  and  $\mathbf{Y}_2$  and a continuous observable response variable  $\mathbf{Y}_3$ . Here,  $\mathbf{Z}_1$ ,  $\mathbf{Z}_2$ , and  $\mathbf{Z}_3$ , represent the first level of latency as the latent continuous response variables.  $\mathbf{H}$  is the second level of latency and is the latent spatial random field of interest. There are 4 covariates in the model,  $\mathbf{X}_1$ ,  $\mathbf{X}_2$ ,  $\mathbf{X}_3$ , and  $\mathbf{X}_4$ . The  $\square$  indicates an observable value and the  $\circ$  indicates a random variable. Additional parameters are shown next to the links.

covariance matrix defined by a spatial correlation function. Let

$$\mathbf{H} \sim GP(\mathbf{X}\boldsymbol{\beta}, \Sigma_{\mathbf{H}}(\phi)) \quad (58)$$

where  $\mathbf{X}$  contains  $p$  location-specific observable covariates and  $\boldsymbol{\beta}$  is a  $p \times 1$  vector of coefficients. The covariance matrix  $\Sigma_{\mathbf{H}}(\phi)$  is described by a function  $\Sigma_{\mathbf{H}}(\phi) = \rho(||\mathbf{s}_i - \mathbf{s}_l||; \phi)$  where  $\rho$  is a covariance function with parameters  $\phi$  that produces a valid covariance matrix depending only on the spatial distance matrix.

### 5.3.3 Bayesian framework

The observed multivariate data matrix  $\mathbf{y}$  is of dimension  $n \times J$  where  $n$  is the number of point-referenced locations in our sample and  $J$  is the number of metrics or responses at each location. For  $i = 1, \dots, n$  and ordinal response variables  $j = 1, \dots, J_o$ , the density of  $y_{ij}$  is the integral from  $\lambda_{j,y_{ij}-1}$  to  $\lambda_{j,y_{ij}}$  of the normal distribution defined for  $Z_{ij}$ . Whereas we first defined  $\mathbf{Z}_j$  as a Gaussian process for each  $j = 1, \dots, J$ , realizations of these processes have a multivariate normal distribution. Denoting the multivariate ordinal observed response  $\mathbf{y}_o = [\mathbf{y}_1, \dots, \mathbf{y}_{J_o}]$ , we write the likelihood of the  $j$ th vector of  $\mathbf{J}_o$ ,  $\mathbf{y}_j$ , as the integral of an  $n$ -dimensional multivariate normal distribution. Therefore

$$p_o(\mathbf{y}_j | \mathbf{H}, \theta_j, \omega_j, \boldsymbol{\lambda}_j, \sigma_j^2) = \int_{\lambda_{j,y_{1j}-1}}^{\lambda_{j,y_{1j}}} \cdots \int_{\lambda_{j,y_{nj}-1}}^{\lambda_{j,y_{nj}}} (2\pi)^{-n/2} |\sigma_j^2 \mathbf{I}_n|^{-1/2} \times \exp \left\{ -\frac{1}{2} [\mathbf{Z}_j - (\theta_j \mathbf{1} + \omega_j \mathbf{H})]' [\sigma_j^2 \mathbf{I}_n]^{-1} [\mathbf{Z}_j - (\theta_j \mathbf{1} + \omega_j \mathbf{H})] \right\} d\mathbf{Z}_j. \quad (59)$$

For the multivariate continuous observed response  $\mathbf{y}_c = [\mathbf{y}_1, \dots, \mathbf{y}_{J_c}]$ , the likelihood of the  $j$ th vector of  $\mathbf{y}_c$  is the multivariate normal density,  $p_c$ . Denoting  $\mathbf{y} = [\mathbf{y}_o, \mathbf{y}_c]$ , the likelihood for all observations is given by

$$p(\mathbf{y} | \mathbf{H}, \boldsymbol{\theta}, \boldsymbol{\omega}, \boldsymbol{\lambda}, \boldsymbol{\sigma}^2) = \prod_{j=1}^{J_o} p_o(\mathbf{y}_j | \mathbf{H}, \theta_j, \omega_j, \boldsymbol{\lambda}_j, \sigma_j^2) \times \prod_{j=1}^{J_c} p_c(\mathbf{y}_j | \mathbf{H}, \theta_j, \omega_j, \sigma_j^2)$$

We define prior distributions for all model parameters and latent random variables to complete the Bayesian model specification. We aim to assign proper yet vague prior distributions to unknown parameters to maintain generality of the model. When applicable, conjugate priors are assigned to ease computational complexity.

To ensure identifiability of the intercept parameter vector  $\boldsymbol{\theta}$ , it is necessary to place a restriction on one of the threshold parameters. Where the lower and upper cut points are defined as  $\lambda_{j,0} = -\infty$  and  $\lambda_{j,K} = \infty$ , we assume without loss of generality that  $\lambda_{j,1} = 0$  for  $j = 1, \dots, J_o$ . Therefore, we are left to estimate  $J_o \times (K - 2)$  threshold parameters. A

uniform prior can be assigned to the cut parameters as shown by Albert and Chib (1993), where  $p(\lambda_{j,k} | \lambda_{j,k-1}, \lambda_{k,k+1}) \propto I_{(\lambda_{j,k-1}, \lambda_{j,k+1})}$ , for  $k = 2, \dots, k - 1$  and  $j = 1, \dots, J_o$ . However, the constraint that  $\lambda_{j,k-1} \leq \lambda_{j,k}$  can lead to poor mixing in the Markov chain. We transform the parameter  $\lambda_{j,1}, \dots, \lambda_{j,k-1}$  to a new space with parameters  $\alpha_{j,1}, \dots, \alpha_{j,k-1}$  (Albert and Chib, 1997). The transformation is performed by setting  $\alpha_{j,1} = \lambda_{j,1} = 0$ ,  $\alpha_{j,2} = \log(\lambda_{j,2})$ , and letting  $\alpha_{j,k} = \log(\lambda_{j,k} - \lambda_{j,k-1})$  for  $k = 3, \dots, K - 1$ . The inverse transformation is expressed as  $\lambda_{j,k} = \sum_{i=2}^k e^{\alpha_{j,i}}$ . We then impose an unrestricted multivariate normal prior distribution to the  $((K - 2) \times 1)$ -dimensional vector  $\boldsymbol{\alpha}$  for each  $j = 1, \dots, J_o$  with mean  $\mathbf{a}$  and covariance matrix  $\mathbf{A}$ .

As denoted above, each of the latent response vectors  $\mathbf{Z}_j$  for  $j = 1, \dots, J$  is a Gaussian process with mean  $\theta_j \mathbf{1} + \omega_j \mathbf{H}$  and covariance matrix  $\sigma_j^2 \mathbf{I}$ . Due to the multivariate multilevel latent structure of the model, some parameters will be fixed to ensure identifiability of the other parameters of interest. When the threshold vectors are metric-specific, as shown in (56), the scale parameter,  $\sigma_j^2$ , for  $j = 1, \dots, J_o$  of the covariance of the continuous multivariate random variables  $\mathbf{Z}_j$  will have to be fixed (Skrondal and Rabe-Hesketh, 2004). When all of the ordinal metrics have the same number of categories, the threshold parameter vector  $\boldsymbol{\lambda}$  can be assumed the same across all metrics. In this case, the parameter  $\boldsymbol{\sigma}^2$  is identifiable for the ordinal metrics if just one element,  $\sigma_j^2$  is fixed. Fixing thresholds to be equal for all metrics is not overly restrictive when the number of categories of the ordered response is small. Indeed, it can be helpful when some of the metrics have few responses in some categories. Also, the mean and variance of the latent continuous response are able to vary across metrics which allows the model to be flexible. However, this assumption becomes more restrictive as the number of categories per metric increases because the model may not be sufficiently flexible to preserve the proportions in each category for the different metrics. Without loss of generality, we set the variance of the first ordinal response variable,  $\sigma_1^2 = 1$  and drop the metric dependence on the thresholds. The remaining parameters,  $\sigma_j^2$  for  $j = 2, \dots, J$ , are assigned inverse-Gamma prior distributions with hyper-parameters  $a_z$  and  $b_z$ .

The mean of the distribution of the latent process  $\mathbf{H}$  is defined as  $\mathbf{X}\boldsymbol{\beta}$ , where the covariate matrix  $\mathbf{X}$  is centered and scaled and does not include the one vector in order to estimate  $\boldsymbol{\theta}$  in (57). The conjugate prior distribution for the  $p \times 1$  vector  $\boldsymbol{\beta}$  is  $N(\mathbf{0}, \sigma_\beta^2 \mathbf{I}_p)$ . Let  $\Sigma_{\mathbf{H}}(\boldsymbol{\phi})$  be the covariance of the distribution of  $\mathbf{H}$  where the vector  $\boldsymbol{\phi}$  represents the parameters of the covariance function. Here we choose an exponential covariance function and write  $\rho(\mathbf{s}_i - \mathbf{s}_l; \boldsymbol{\phi}) = \phi_1 \exp^{-d_{il}\phi_2}$  where  $d_{il}$  represents the Euclidean distance between locations  $i$  and  $l$ . The conjugate inverse-Gamma prior distribution is assigned to  $\phi_1$  and a Gamma prior distribution is assigned to  $\phi_2$ . The shape and scale hyper-parameters of these distributions are  $a_{\phi_1}$  and  $b_{\phi_1}$  and  $a_{\phi_2}$  and  $b_{\phi_2}$ , respectively. For identifiability, however,  $\phi_1$  is set to 1 when all response variables are ordinal. Specification of the prior distribution of  $\phi_2$  and its corresponding hyper-parameters can be challenging and must be chosen with careful consideration to keep it non-informative. (see e.g., Schmidt et al., 2008).

The parameters  $\boldsymbol{\theta}$  and  $\boldsymbol{\omega}$  are each assigned a multivariate normal prior distribution with mean vector  $\mathbf{0}$  and covariance matrix  $\sigma^2 \mathbf{I}_J$ . The scale parameters of both covariance matrices,  $\sigma_\theta^2$  and  $\sigma_\omega^2$ , are chosen to be large such that the prior distributions are vague. To ensure identifiability of the model parameters one value of the  $(1 \times J)$ -dimensional vector  $\boldsymbol{\omega}$  must be fixed. Without loss of generality we set  $\omega_1 = 1$ . Fixing  $\omega_1$  establishes a point of reference for the relationship between  $\mathbf{Z}$  and the parameter of interest,  $\mathbf{H}$ .

### 5.3.4 Inference

We make inference about the parameters of the model using the Bayesian paradigm incorporating Gibbs and Metropolis-Hastings sampling techniques. This approach allows estimation of both the model parameters and the multilevel and multivariate latent variables, as well as their uncertainty. Due to the constrained threshold parameter vector  $\boldsymbol{\lambda}$ , the model proposed in this work is computationally complex.

The joint posterior distribution of the unknown parameters of interest and the latent variables given the observed data can be factored and written as

$$\begin{aligned} p(\mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\omega}, \mathbf{H}, \boldsymbol{\lambda}, \sigma^2, \boldsymbol{\beta}, \boldsymbol{\phi} | \mathbf{y}) &\propto p(\mathbf{y} | \mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\omega}, \mathbf{H}, \boldsymbol{\lambda}, \sigma^2, \boldsymbol{\beta}, \boldsymbol{\phi}) p(\mathbf{Z} | \boldsymbol{\theta}, \boldsymbol{\omega}, \mathbf{H}, \boldsymbol{\lambda}, \Sigma, \boldsymbol{\beta}, \boldsymbol{\phi}) \\ &\quad \times p(\mathbf{H} | \boldsymbol{\beta}, \boldsymbol{\phi}) p(\boldsymbol{\theta}, \boldsymbol{\omega}, \boldsymbol{\lambda}, \sigma^2, \boldsymbol{\beta}, \boldsymbol{\phi}) \end{aligned}$$

where  $p(\mathbf{y} | \cdot)$  is the distribution of the mixed ordinal and continuous multivariate random variables given the model parameters and latent variables,  $p(\mathbf{Z} | \cdot)$  is the conditional distribution of the continuous latent random variable,  $p(\mathbf{H} | \boldsymbol{\beta}, \boldsymbol{\phi})$  is the distribution of the latent spatial field of interest, and  $p(\boldsymbol{\theta}, \boldsymbol{\omega}, \boldsymbol{\lambda}, \sigma^2, \boldsymbol{\beta}, \boldsymbol{\phi})$  is the joint prior distribution for the parameters  $\boldsymbol{\theta}$ ,  $\boldsymbol{\omega}$ ,  $\boldsymbol{\lambda}$ ,  $\sigma^2$ ,  $\boldsymbol{\beta}$ , and  $\boldsymbol{\phi}$ .

The Markov chain Monte Carlo (MCMC) algorithm proceeds as follows:

1. Update the spatial covariance scale and range parameters,  $\phi_1$  and  $\phi_2$ , respectively. Parameter  $\phi_1$  can be drawn directly from its complete conditional distribution whereas  $\phi_2$  requires a Metropolis-Hastings step to sample from its complete conditional distribution.
2. Update the regression parameter vector  $\boldsymbol{\beta}$  and the latent spatial multivariate normal,  $\mathbf{H}$ , from their complete conditional distributions.
3. Update the metric-specific parameters  $\boldsymbol{\theta}$  and  $\boldsymbol{\omega}$  and variance parameter  $\sigma^2$  each in block form from their complete conditional distributions.
4. Update the threshold parameters,  $\boldsymbol{\lambda}$  by drawing  $\boldsymbol{\alpha}$  from  $p(\boldsymbol{\alpha} | \mathbf{y}_o, \mathbf{Z}_o)$  and inverse mapping to get  $\boldsymbol{\lambda}$ . See Higgs and Hoeting (2010) for explicit details on the re-parameterization and updating scheme for  $\boldsymbol{\lambda}$ .
5. Update the latent multivariate normal  $\mathbf{Z}_o$  from the complete conditional distribution.

The samples from the posterior distribution can then be used to draw inference on both the model parameters and latent variables.

## 5.4 Posterior inference

### 5.4.1 Posterior prediction

The model can be used to make predictions for the mixed ordinal and continuous multivariate response as well as the underlying latent spatial process at unobserved locations. The multivariate response at  $m$  unobserved locations will be denoted  $\tilde{\mathbf{Y}} = [\tilde{\mathbf{Y}}_1, \dots, \tilde{\mathbf{Y}}_J]$  where  $\tilde{\mathbf{Y}}_j = [\tilde{Y}_{1j}, \dots, \tilde{Y}_{mj}]'$ . Similarly, predictions of the latent spatial process at the  $m$  unobserved locations will be written as  $\tilde{\mathbf{H}} = [\tilde{H}_1, \dots, \tilde{H}_m]'$ . Predictions can be made using the Bayesian posterior predictive distributions  $p(\tilde{\mathbf{Y}}|\mathbf{y})$  and  $p(\tilde{\mathbf{H}}|\mathbf{y})$  for the multivariate response and latent spatial process, respectively.

In most applications, the value of the latent variable  $H_i$  at location  $i$  will be inconsequential but the comparison of  $\mathbf{H}$  across locations may be of interest. For example, wetland condition encompasses many variables. If a latent variable  $H_i$  summarizes wetland condition at site  $i$ , comparisons among sites will be useful to many agencies and individuals. For each location, we obtain draws from the distributions  $H_i|\mathbf{y}$  and  $\tilde{H}_i|\mathbf{y}$  for each iteration of the Markov chain. We then examine the distribution of the posterior ranks for each location to draw inference and conduct comparisons across the region of interest.

Other model parameters of particular interest include the parameters of the latent spatial field  $\mathbf{H}$ ,  $\boldsymbol{\beta}$  and  $\boldsymbol{\phi}$ , as well as the metric-specific parameters of  $\mathbf{Z}$ ,  $\boldsymbol{\omega}$ , and  $\sigma^2$ . Estimating the parameter vector of coefficients of the linear model,  $\boldsymbol{\beta}$ , enables us to evaluate the relationship between the point-referenced covariates and the latent random variable  $\mathbf{H}$ . The unaccounted for spatial correlation of the latent random variable  $\mathbf{H}$  can be estimated by drawing inference on  $\phi_1$  and  $\phi_2$  as well as the effective range,  $3/\phi_2$ . The effective range is the distance at which the correlation function does not exceed 0.05 times the variance.

### 5.4.2 Multivariate correlation statistics

We estimate the relationship between latent variables  $\mathbf{Z} = [\mathbf{Z}_1, \dots, \mathbf{Z}_J]$  and  $\mathbf{H}$  by computing multiple correlation values. Due to the deterministic relationship between latent  $\mathbf{Z}$  and observed  $\mathbf{Y}$ , we assume that the relationship we are estimating will capture that of the relationship between  $\mathbf{H}$  and the multivariate response  $\mathbf{Y}$ . This is a method used in canonical correlation analysis to evaluate the level of linear relationship between two sets of variables (Rencher, 2002). It is useful to first partition the covariance matrix of the matrix  $\mathbf{Z}$  and vector  $\mathbf{H}$  as

$$\mathbf{S} = \begin{pmatrix} \mathbf{S}_{ZZ} & \mathbf{S}_{ZH} \\ \mathbf{S}_{HZ} & \mathbf{S}_{HH} \end{pmatrix} \quad (60)$$

where  $\mathbf{S}_{ZZ}$  is the  $J \times J$  sample covariance matrix of  $\mathbf{Z}$ ,  $\mathbf{S}_{ZH}$  is the  $J \times 1$  matrix of sample covariances between  $\mathbf{Z}$  and  $\mathbf{H}$ , and  $\mathbf{S}_{HH}$  is the sample covariance of  $\mathbf{H}$ . The  $(j, j')$  element of  $\mathbf{S}_{ZZ}$  is the covariance between the  $(n \times 1)$ -dimensional vectors  $\mathbf{Z}_j$  and  $\mathbf{Z}_{j'}$ . Similarly, the  $j$ th element of  $\mathbf{S}_{ZH}$  is the covariance between the  $(n \times 1)$ -dimensional vectors  $\mathbf{Z}_j$  and  $\mathbf{H}$ . A measure of association between  $\mathbf{Z}$  and  $\mathbf{H}$  as a whole is  $R_M^2 = |\mathbf{S}_{ZZ}^{-1} \mathbf{S}_{ZH} \mathbf{S}_{HH}^{-1} \mathbf{S}_{HZ}|$ . This value is analogous to  $R^2$  in linear regression. This value can also be expressed in terms of the canonical correlations between  $\mathbf{Z}$  and  $\mathbf{H}$ . However, we would like to evaluate the correlation between each of the responses and  $\mathbf{H}$  separately. The correlation between  $\mathbf{Z}_j$  and  $\mathbf{H}$  is defined as the square root of

$$R_{Z_j|H}^2 = S_{Z_j H} S_{HH}^{-1} S_{HZ_j} S_{Z_j Z_j}^{-1} \quad (61)$$

where  $S_{Z_j H}$  is the  $j$ th element of  $\mathbf{S}_{ZH}$ ,  $S_{HZ_j}$  is the  $j$ th element of  $\mathbf{S}_{HZ}$ , and  $S_{Z_j Z_j}$  is the  $j$ th element of the diagonal of  $\mathbf{S}_{ZZ}$ .

We evaluate the multiple correlation for each metric using the posterior simulations. Therefore, at each simulation draw of the model parameters, we first compute the covariance matrix  $\mathbf{S}$ . Then, for  $j = 1, \dots, J$ , we compute the correlation between the posterior draw of

$\mathbf{Z}_j$  and  $\mathbf{H}$  using (61). Larger values of  $R_{Z_j|H}$  (i.e., closer to 1) suggest that metric  $j$  is more correlated with the underlying latent variable  $\mathbf{H}$ . In application, a large  $R_{Z_j|H}$  value means that metric  $j$  is a good measurement or predictor for the unobserved latent spatial process. We use the multiple correlation values to rank the importance of each of the response metrics in measuring the latent spatial process of wetland condition. Further details on the multiple correlation statistic are given in Section A.5

#### 5.4.3 Model evaluation

Mixed ordinal and continuous multivariate response models present a unique problem for model evaluation. Whereas there are multiple methods to measure predictive ability for discrete response models or continuous response models, the difficulty arises when we wish to compare mixed response models with both continuous and discrete variables. Multi-category loss functions like those presented by Higgs and Hoeting (2010) cannot be applied when  $J_c \neq 0$ . Therefore, we direct our attention to loss functions for continuous data since we have a continuous latent variable for all  $J$  metrics. In the Bayesian framework, the loss is computed by comparing the true value to draws from the posterior predictive distribution. Therefore, we first need to determine the “true” value for the ordinal variable on the continuous scale. The posterior mean or median of the latent continuous response could be used as the “true” value but we feel this favors the discrete response metrics. We propose setting the “true” value for the continuous representation of the observed ordinal variable  $y$  as the value  $\hat{Z}$  such that

$$\frac{\int_{\lambda_{y-1}}^{\hat{Z}} \frac{1}{\sqrt{2\pi\sigma_z^2}} \exp^{\frac{-1}{2\sigma_z^2}(Z-\mu_z)^2} dZ}{\int_{\lambda_{y-1}}^{\lambda_y} \frac{1}{\sqrt{2\pi\sigma_z^2}} \exp^{\frac{-1}{2\sigma_z^2}(Z-\mu_z)^2} dZ} = 0.50 \quad (62)$$

where  $\mu_z$  and  $\sigma_z^2$  are the mean and variance of the posterior distribution of  $Z$ , respectively. Therefore,  $\hat{Z}$  is the 50th percentile of the estimated normal distribution between the thresholds  $\lambda_{y-1}$  and  $\lambda_y$ . We can estimate both  $\mu_z$  and  $\sigma_z^2$  for  $i = 1, \dots, n$  and  $j = 1, \dots, J_o$  using the posterior draws of the parameters  $\omega_j, \theta_j, H_i$  and  $\sigma_j^2$ . We apply this method to perform model comparison in Section 5.5 under squared error loss.

## 5.5 Assessing wetland condition

### 5.5.1 Data and model specification

The data were collected at 95 locations within the North Platte River Basin and 137 locations within the Rio Grande River Basin, resulting in  $n = 232$  locations (Figure 5.2). The surveyed parcel consisted of a 0.5-hectare area around each target location. These locations were sampled randomly using a Generalized Random Tessellation Stratified (GRTS) survey design (Stevens and Olsen, 2004). Details of the GRTS design differed between the basins (Lemly et al., 2011; Lemly and Gillian, 2012). We applied the multivariate multilevel latent Gaussian process model to each river basin separately and to the basins together and reached similar conclusions. The results presented here are those from the river basins modeled together as one data set.

The data include measurements to evaluate the biotic integrity of the wetland, as well as the surrounding landscape, soil, and water conditions. Here we apply our multilevel latent model to evaluate the biotic integrity of wetlands. We refer to the biotic integrity as a proxy for wetland condition because it is the biotic condition that drives the overall condition of the wetland. Five measurements, or metrics, were derived from detailed vegetation surveys conducted at each field location. The five metrics include native plant cover, noxious weed cover, aggressive native cover, structural complexity, and floristic quality assessment (Lemly and Gillian, 2012). It is assumed that each of these metrics represents a component of the biotic integrity of the wetland. It is current practice for wetland condition assessment

to use a method of weighted averages to evaluate the biotic condition using these metrics. Whereas these weights are often thought to be assigned based on best professional judgement or without statistical support, our goal is to use the data within the multivariate multilevel latent Gaussian process model to rank the metrics in a hierarchy of most important to least important to assess wetland condition. We can then identify a subset of the metrics that are most valuable for future data collection.

Each metric was reported on a five-category ordinal scale from “poor” to “excellent,” to which we assign integer values from 1 to 5, respectively (Appendix A.3). The floristic quality assessment, native plant cover, noxious weed cover, and aggressive native cover metrics are discretized continuous variables (see Lemly and Gillian (2012) for more details on discretization). The floristic quality metric evaluates the overall floristic quality and fidelity of the plant community at each location to natural, or undisturbed, conditions (Rocchio, 2007). Each species in the Colorado flora has been assigned a coefficient of conservatism (C value: 0-10) that reflects the species tolerance or intolerance to disturbance (Swink and Wilhem, 1994; Taft et al., 1997). The continuous value is an average of C values assigned to the plants present at the wetland site. The ordinal value at each location is assigned by applying a threshold to the continuous metric value. However, this thresholding scheme is dependent on wetland type because the natural vegetation differs between wetland types with some naturally containing plant species with lower values of floristic quality. Structural complexity is Likert-like and has no tangible underlying continuous variable. Here, we fit a discrete-only model with  $J_o = 5$  and  $J_c = 0$  as well as a mixed response model with  $J_o = 4$  and  $J_c = 1$  where the continuous metric is floristic quality and compare the results. For all  $J_o$  ordinal responses, the observed value  $Y_i \in \{1, \dots, K = 5\}$  for  $i = 1, \dots, 232$ .

The variance of  $Z_{i1}$  is fixed and held constant across all locations at  $\sigma_1^2 = 1$  for model identifiability. The hyper-parameters of the inverse-Gamma distributions of the metric specific variance parameters  $\sigma_j^2$  are  $a_z = b_z = 1$  for  $j = 2, \dots, 5$ . The metric-specific parameters

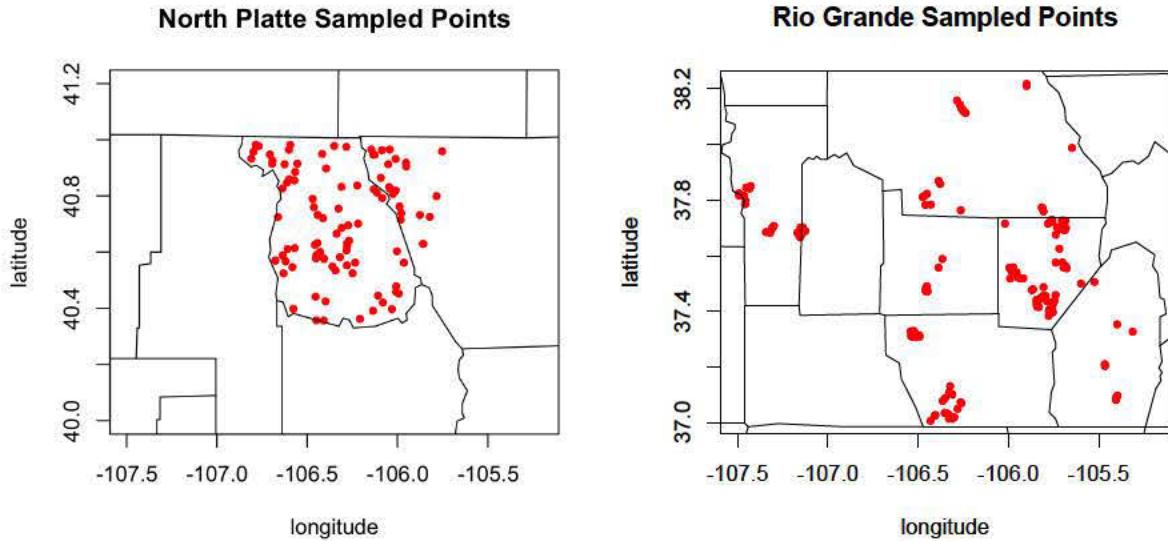


Figure 5.2: The  $n = 232$  locations of observed data within the North Platte and Rio Grande River Basins.

$\theta$  and  $\omega$  are of dimension  $1 \times 5$ . We set the variance hyper-parameters  $\sigma_\theta^2 = \sigma_\omega^2 = 100$ . For identifiability of the coefficient vector  $\beta$ , we fix  $\omega_1 = 1$ .

Elevation and percent of closed tree canopy vegetation are two continuous point-referenced covariates used to model the mean of the Gaussian process  $\mathbf{H}$  (Appendix A.3). We also included wetland type as a categorical covariate with five levels: riparian shrublands and woodlands, saline wetlands, marshes, wet meadows, and fens. The prior distribution of the coefficient vector  $\beta$  is  $N(\mathbf{0}, \sigma_\beta^2 \mathbf{I}_p)$  with  $\sigma_\beta^2 = 100$  and  $p = 6$ . The exponential covariance function for the latent random variable  $\mathbf{H}$  is defined as  $\phi_1 \exp^{-d_{il}\phi_2}$  where  $d_{il}$  is the Euclidean distance between locations  $i$  and  $l$ . In the mixed response model, we assign an Inv.Gamma(1, 1) prior for  $\phi_1$  and fix  $\phi_1 = 1$  in the discrete-only model. In Chapter 3 we discussed identifiability of univariate response probit regression models for spatially correlated ordinal data and concluded that Bayesian learning can exist for both parameters of the exponential covariogram. However, due to the already complex structure of the model, we fixed the partial-sill parameter in the discrete-only model in this work. Fixing the partial-sill to 1 implies that for metric 1, where  $\sigma^2 = 1$ , the amount of variation explained by the spatial

process is the same as the noise for the metric. That is, the signal-to-noise ratio is 1 for metric 1. In both models,  $\phi_2$  is assigned a Gamma(2, 2) prior distribution. The prior of  $\phi_2$  was chosen such that the effective range,  $3/\phi_2$ , could reach the maximum distance between sites.

### 5.5.2 Model results

The Markov chain Monte Carlo algorithm was run for 100,000 iterations using R software (R Development Core Team, 2007). The first 10,000 iterations for both models were discarded as burn-in. We ran multiple chains from different starting values to evaluate convergence of our Gibbs sampler. The Gelman (2004) potential scale reduction factor for each parameter was below 1.2. Similarly, other standard diagnostics showed no indications of lack of convergence.

The posterior estimates from both the discrete-only model and the mixed response model indicate that wetland condition scores are higher for locations at higher elevations and with higher percentages of closed tree canopy (Table 5.1 for discrete-only response model, Table 5.2 for mixed response model). The coefficients  $\beta_3$ ,  $\beta_4$ ,  $\beta_5$ , and  $\beta_6$  represent the effect for saline, marsh, wet meadow, and fen wetland types, respectively, relative to riparian shrublands and woodlands. These values vary greatly between models due to the discretization of the floristic quality assessment metric. The discretization process includes additional information about the condition of each site based on its wetland type and thus, the ordinal values for this metric are not uniformly assigned across all locations (Lemly and Gillian, 2012). For example, a riparian wetland with a floristic quality value of 5.6 on the continuous scale would be assigned a 4 on the ordinal scale, whereas a marsh wetland with the same continuous value would be assigned an ordinal value of 5. For this reason, the coefficients for marsh and saline wetland type vary between the two models.

All estimates of the factor loading (57)  $\omega$  are positive indicating that the linear relationship between latent wetland condition and each of the individual metrics is positive (Tables

Table 5.1: Posterior estimates and 95% credible intervals for discrete-only model parameters.

Parameter		Estimate	95 % CI
$\beta_1$	Elevation	0.54	(0.22, 0.89)
$\beta_2$	Closed tree canopy	0.40	(0.20, 0.62)
$\beta_3$	Saline	0.62	(0.16, 1.10)
$\beta_4$	Marsh	0.60	(0.26, 0.97)
$\beta_5$	Wet meadow	-0.03	(-0.28, 0.21)
$\beta_6$	Fen	1.00	(0.55, 1.53)
$3/\phi_2$	Effective Range	0.88	(0.45, 1.84)
$\omega_1$	Native plant cover	1.00	
$\omega_2$	Noxious weed cover	1.37	(1.00, 1.90)
$\omega_3$	Aggressive native cover	2.54	(0.89, 6.01)
$\omega_4$	Structural diversity	0.21	(0.11, 0.33)
$\omega_5$	Floristic quality	1.59	(1.33, 1.91)
$\sigma_1^2$	Native plant cover	1.00	
$\sigma_2^2$	Noxious weed cover	1.34	(0.86, 2.16)
$\sigma_3^2$	Aggressive native cover	20.46	(8.00, 67.64)
$\sigma_4^2$	Structural diversity	0.89	(0.67, 1.18)
$\sigma_5^2$	Floristic quality	0.36	(0.22, 0.57)

5.1 and 5.2). Based on the 95% credible intervals these estimates are all significantly different from 0.

The estimates of effective range of spatial correlation for the two models are comparable at 88 and 67 km. The overall maximum distance between the 232 observed locations is  $d_{max} = 470$  km whereas the maximum distance within the North Platte and Rio Grande River Basins is 93 km and 202 km, respectively. The minimum distance between sampled locations from the two river basins is 240 km. Not surprisingly, the estimate of the effective range indicates that the spatial correlation of wetland condition is only of interest within the river basins and not between them.

To compare the performance of the discrete-only model to the mixed response model, we compute the median squared error loss using the latent response  $\mathbf{Z}$ . For the ordinal metrics, we estimate the “true” value of  $\mathbf{Z}$  using (62). The squared error loss for each metric is similar

Table 5.2: Posterior estimates and 95% credible intervals for mixed response model parameters.

Parameter		Estimate	95 % CI
$\beta_1$	Elevation	0.39	(0.23, 0.57)
$\beta_2$	Closed tree canopy	0.17	(0.07, 0.28)
$\beta_3$	Saline	-0.21	(-0.51, 0.07)
$\beta_4$	Marsh	-0.22	(-0.43, -0.03)
$\beta_5$	Wet meadow	-0.26	(-0.42, -0.12)
$\beta_6$	Fen	0.22	(0.05, 0.42)
$3/\phi_2$	Effective Range	0.67	(0.31, 3.08)
$\omega_1$	Native plant cover	1.00	
$\omega_2$	Noxious weed cover	1.21	(0.86, 1.69)
$\omega_3$	Aggressive native cover	5.37	(2.28, 10.57)
$\omega_4$	Structural diversity	0.38	(0.25, 0.54)
$\omega_5$	Floristic quality	1.52	(1.28, 1.83)
$\sigma_1^2$	Native plant cover	1.00	
$\sigma_2^2$	Noxious weed cover	1.36	(0.89, 2.18)
$\sigma_3^2$	Aggressive native cover	11.83	(4.41, 33.67)
$\sigma_4^2$	Structural diversity	0.61	(0.46, 0.81)
$\sigma_5^2$	Floristic quality	0.18	(0.13, 0.24)

Table 5.3: Discrete-only model: For each metric, estimates and 95% credible intervals of multiple correlation, estimates of the percent contribution, and rank in evaluating wetland condition.

Metric	Parameter	Est.	95 % CI	% Contrib.	95 % CI	Rank	Index
Native plant cover	$R_{Z_1 H}$	0.80	(0.68, 0.88)	0.23	(0.20, 0.26)	3	20%
Noxious weed cover	$R_{Z_2 H}$	0.84	(0.70, 0.92)	0.24	(0.21, 0.28)	2	0 or 20%
Aggressive native cover	$R_{Z_3 H}$	0.58	(0.23, 0.83)	0.17	(0.08, 0.22)	4	0 or 20%
Structural diversity	$R_{Z_4 H}$	0.28	(0.09, 0.49)	0.08	(0.03, 0.13)	5	20%
Floristic quality	$R_{Z_5 H}$	0.96	(0.92, 0.98)	0.28	(0.25, 0.32)	1	40%

between the discrete-only model and the mixed response model (see Table A.5 in Appendix A.4).

The remaining results presented here are for the discrete-only model because it is of interest to the ecologists. The multiple correlation statistics (61) suggest that metric 5, floristic quality assessment, is most closely correlated with wetland condition (Table 5.3) and should be ranked most important in evaluating wetland condition. The assessments of native plant cover, noxious weed cover, and aggressive native cover are slightly less correlated with wetland condition. The structural diversity measurement (metric 4) appears to be the least correlated with wetland condition of the five measurements and therefore is ranked last. Estimates of percent contribution are also given in Table 5.3 where the values are calculated based on the estimate of  $R_{Z_j|H}$  divided by the sum of all estimates of  $R_{Z_j|H}$  for  $j = 1, \dots, 5$ . The percent contribution estimates can be used as weights for each of the metrics in estimating the underlying wetland condition. The last column in Table 5.3 reports the current index weights that were selected by a group of wetland experts (Lemly and Gillian, 2012). The scientists believe floristic quality assessment to be the most important. The weight “0 or 20%” assigns 20% weight to the lower of the noxious weed cover and aggressive native cover metrics. Our estimates improve on the current weighting scheme by being statistically derived weights for each of the metrics with confidence limits.

We estimate the latent spatial process  $\mathbf{H}$  of wetland condition within the North Platte and Rio Grande River Basins by drawing from the posterior distribution  $p(H_i|\mathbf{y})$  for  $i =$

$1, \dots, n$ . Since the values of  $\mathbf{H}$  hold no intrinsic value, we rank the locations from draws of the posterior distribution. For each draw  $t$  in  $1, \dots, T$ , the posterior value of  $H_i$  is ranked across all  $i = 1, \dots, n$  assigning a posterior rank to each location for each draw. We estimate the latent spatial process of wetland condition by computing the median of the posterior ranks at each location. A location with a median posterior rank falling in the top 20% of ranks indicates that the wetland at this particular location is in the top 20% of all wetlands in the region in terms of biotic condition. Figure 5.3 shows the median of the posterior ranks across all locations within the North Platte and Rio Grande River Basins. Linear interpolation is used to provide a relatively smooth surface over the two river basins. Note, however, that wetlands are not found continuously over the regions. The color scale and contours of the surface are based on the percentile of the median of the posterior ranks over all locations. Wetland management efforts should be directed towards areas within the river basins with low posterior ranks. For example, the wetlands in the eastern region of the Rio Grande River Basin may be of concern. Conversely, land managers may wish to preserve wetlands in good condition such as those shown in red in Figure 5.3. Similar plots can be made for the estimates of uncertainty. We performed a simulation study to evaluate the model and out-of-sample predictive performance (Appendix A.6). The results indicate that our method provides accurate parameter estimates, predictions, and predictive coverage for the simulation scenarios that we considered.

## 5.6 Discussion

The multilevel multivariate latent Gaussian process model presented in this paper provides a method for evaluating a continuous latent Gaussian process using mixed ordinal and continuous multivariate response data. A multivariate latent variable is used as the continuous representation of the multivariate mixed response. A second latent variable depending on site-specific covariates models the continuous random field that is assumed to be driving the multivariate response. If we were to estimate the partial-sill parameter of the continuous

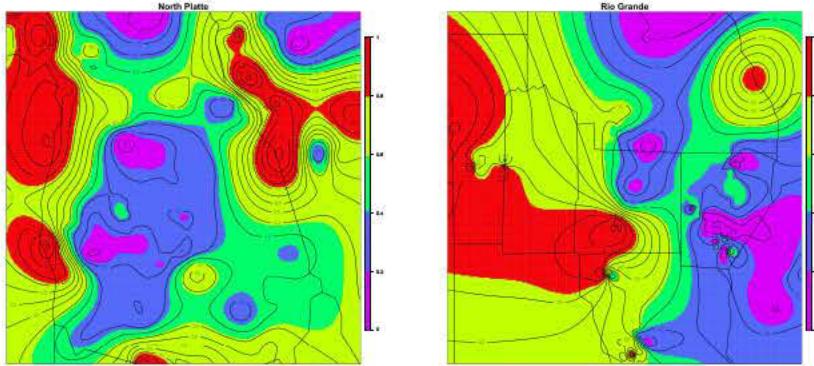


Figure 5.3: Median of the posterior ranks of the latent spatial process encompassing wetland condition ( $\mathbf{H}$ ) across space from the discrete-only response model.

random field in the future using MCMC, we would utilize the re-parameterization of the exponential covariogram proposed in Chapter 3.4. The continuous latent random field was modeled in this work using a Gaussian process. Lindgren et al. (2011) present an approximation to the Gaussian field using a Gaussian Markov random field. Their approach could accelerate estimation of the parameters of the spatial covariance function.

Our multilevel multivariate latent variable model is used to evaluate the ecological condition of wetlands or other natural resources. Whereas Liu et al. (2005) gave a general framework for spatial structural equation modeling, the model presented here for multivariate response data could be easily replicated or modified for other applications. The model is advantageous as it allowed for comparisons of the condition of wetlands in two river basins in Colorado across space. Further, in-field measurements, or metrics, were ranked when evaluating the wetland condition score at each particular location. These rankings allow assignment of statistically valid weights to the five measurements or metrics. These results will lead to a decrease in the time and effort needed for future wetland evaluation. They will also help land managers to design and implement effective protocols for maintaining and restoring wetland habitats.

While we have described and applied the model to a problem related to wetland condition, the model holds in much larger context. For example, in human health, doctors apply a panel

of tests to a subject to evaluate health. In this case, the multivariate response would be the outcomes of the tests and the covariates would be individual information such as gender and body mass index (BMI).

## CHAPTER 6

### CONCLUSION AND FUTURE WORK

In this work we proposed a multivariate multilevel latent variable model to evaluate the condition of wetlands in Colorado. Latent variables were used in the model for three distinct purposes. First, the multivariate response data we observed were both continuous and ordinal. Therefore, in assuming the ordinal response data were generated by continuous latent variables, we were able to combine the two types of data in one model. Second, the probit linear model (PLM) can be defined using continuous latent variables. In Chapter 2 we showed that the latent variable approach to the probit model can be efficiently fitted within the Bayesian framework. This is because the latent variables allow for simple and feasible simulation from the posterior distribution. Lastly, latent variables can also be used to draw inference on an unmeasurable variable or quantity. In the multivariate multilevel latent variable model proposed in Chapter 5, we assumed a latent variable, or common factor, for the univariate unmeasurable quantity of wetland condition. In the example, wetland condition summarized the multivariate response at each location. We were able to draw inference on the latent variable using mixed continuous and discrete multivariate response data and covariate information at observed locations. Further, we were able to predict the univariate quantity of wetland condition at unobserved locations.

There were two overarching themes throughout this work for spatial probit models: computational efficiency and parameter identifiability. Whereas latent variable models are an extremely flexible class of models, they tend to have a high-dimensional parameter space leading to slow mixing when running Markov chain Monte Carlo (MCMC). Further, multivariate and multilevel latent variable models can easily suffer from parameter nonidentifiability. The issues of computational efficiency and parameter identifiability are magnified

when fitting geostatistical spatial models. MCMC is slow for geostatistical spatial models since they require matrix inversion at each iteration and contain parameters that are weakly identifiable.

## 6.1 Computational efficiency: Conclusion and future work

In Chapter 2, we proposed a set of data augmentation (DA) and parameter-expanded data augmentation (PX-DA) algorithms for the spatial PLMM for binary and ordinal data. We showed that DA algorithms increase the ease of sampling from the posterior distribution. PX-DA algorithms increase the rate of convergence of MCMC by increasing variation between iterations within the chain. Other methods and approximations for more efficient inference of spatial Bayesian models have been introduced in the literature. Integrated nested Laplacian approximations use both Laplacian and Gaussian approximations for fitting latent Gaussian models (Rue et al., 2009). Gaussian Markov random fields can be used to approximate continuous spatial fields (Lindgren et al., 2011). Fixed-rank kriging (Cressie and Johannesson, 2008), predictive processes (Banerjee et al., 2008), and covariance tapering (Furrer et al., 2006) are other dimension-reduction techniques that have sped up computation for geostatistical spatial models. The benefit of PX-DA over some of the other methods for increased computation efficiency is that it uses MCMC for inference. PX-DA MCMC algorithms should obtain better approximation because they better estimate uncertainty. The nonidentifiable parameter(s) used in PX-DA algorithms is(are) integrated out after each iteration of MCMC. Therefore, we obtain samples from the posterior distribution of interest.

### 6.1.1 MCMC algorithm using marginalized latent variable

DA and PX-DA algorithms were motivated by the multivariate multilevel latent variable model proposed in Chapter 5. Our aim was to increase computational efficiency for multivariate response data fit using a spatially correlated common factor model. When fitting

the spatial PLMM in a Bayesian framework using a Gibbs sampler, Banerjee et al. (2003, Chapter 5) recommend that algorithms avoid drawing from the full conditional of the spatial random effect,  $\mathbf{W}$ . Instead, they suggest integrating over  $\mathbf{W}$  and drawing the latent continuous response,  $\mathbf{Z}$ , from its marginal distribution. This is because the marginal variance of  $\mathbf{Z}$ , which is equal to  $\Sigma_W + \mathbf{I}$ , has better mixing properties than the variance of  $\mathbf{W}$ ,  $\Sigma_W$ .

The model proposed in Chapter 5 follows the framework of a common factor model where the response is multivariate. Therefore, the marginal distribution of the latent continuous response,  $\mathbf{Z}$ , has dimension  $nJ \times 1$ , where  $n$  is the number of observed locations and  $J$  is the number of response variables. The covariance matrix of  $\mathbf{Z}$  is dense with dimension  $nJ \times nJ$ . Using the suggested marginalization over the latent spatial variable, each iteration of the MCMC requires inversion of the  $nJ \times nJ$  matrix. This has high computation cost for even reasonable values of  $n$  and  $J$ . By not marginalizing over  $\mathbf{W}$  in the MCMC algorithm, matrix inversion is only required for the  $n \times n$  covariance matrix of  $\mathbf{W}$ . Therefore, there is a trade-off between high computation cost of matrix inversion and better mixing properties of MCMC algorithms that marginalize over the spatial random effect. We would like to investigate this trade-off and determine a general rule for when to run MCMC using the marginalized algorithm for the common factor model.

The fact that the marginal variance of  $\mathbf{Z}$  has better mixing properties than the variance of  $\mathbf{W}$  also suggests that the second-stage spatial probit model may out-perform the first-stage probit spatial model. Recall that the first-stage spatial probit model (36) is a no-nugget model where all stochasticity in the binary or ordinal response data is modeled as spatial stochasticity. That is, the variance of  $\mathbf{Z}$  is  $\Sigma_W$  as opposed to  $\Sigma_W + \mathbf{I}$ . In this case there is no random effect to marginalize in order to improve mixing. Unfortunately, this does not automatically imply that the first-stage spatial probit model will be worse in terms of mixing and convergence than the marginalizable second-stage spatial probit model (39). This is because the first-stage spatial probit model is likelihood identifiable when  $\tau$  is fixed (i.e.,  $\Sigma_W = \tau \mathbf{R}$  where  $\tau$  is the spatial variance and  $\mathbf{R}$  is a spatial correlation matrix). We showed

that the second-stage spatial probit model with covariance function containing a partial sill and range parameter is weakly identifiable. Weak identifiability can cause mixing and convergence issues when fitting MCMC. This is why we and others (e.g., (Schmidt et al., 2008)) recommend informative prior distributions for weakly identifiable parameters. We would like to further compare the first-stage and second-stage spatial models in terms of their mixing properties when running MCMC. This comparison could include the benefits of DA and PX-DA algorithms for the first-stage and second-stage spatial probit models discussed in Chapter 2. Berrett and Calder (2012) developed DA and PX-DA algorithms for the first stage spatial probit model for binary response data. These algorithms would need to be extended for ordinal response data.

### 6.1.2 Parameterizations of spatial covariance functions

In Chapter 3.4 we proposed a re-parameterization of the exponential covariance function. Recall that the traditional approach is to assign prior distributions to  $\tau$  and  $\phi$  where the exponential covariance function is defined as

$$\text{Cov}(W(\mathbf{s}_i), W(\mathbf{s}_j)) = \tau \exp^{-\frac{1}{\phi} d_{ij}}. \quad (63)$$

The re-parameterization (35) defined both the partial sill and range parameter as functions of  $\phi$ . The re-parameterization led to better mixing of the MCMC in the small simulation study that we conducted by decreasing the autocorrelation in the chain between iterations. We would like to further investigate the re-parameterization of the covariance function in a few ways. First, we would like to determine whether the re-parameterized exponential covariance function is sensitive to prior distributions. We will run MCMC with different prior distributions and compare mixing between the original and the proposed parameterization. We would also like to see if the benefits of the re-parameterization hold for different covariance functions, such as the Matérn covariance function (of which exponential is a special

case). This would allow for a general recommendation for parameterization of spatial covariance functions when fitting spatial PLMMs. Third, our simulation study was limited to ordinal response data and the spatial PLMM. Therefore, additional simulations will include using the re-parameterized covariance function to fit spatial models to both continuous and count response data. Lastly, we will compare our proposed parameterization to the spatial covariance parameterizations proposed by Christensen et al. (2006) and Diggle and Ribeiro (2007, Chapter 5.4).

Our comparison would include, but is not limited, to the following three parameterizations of the priors for the parameters in (63):

1. Christensen et al. (2006) parameterization: priors assigned to  $\theta_1$  and  $\theta_2$  where  $\theta_1 = \log(\tau^{1/2})$  and  $\theta_2 = \log(\tau/\phi)$ .
2. Diggle and Ribeiro (2007, Chapter 5.4) parameterization: priors assigned to  $\theta_1$  and  $\theta_2$  where  $\theta_1 = \log(\tau/\phi)$  and  $\theta_2 = \log(\phi)$ .
3. Schliep parameterization: Priors assigned to  $\theta_1$  and  $\theta_2$  where  $\theta_1 = \tau/\phi$  and  $\theta_2 = \phi$ .

The comparisons will be made in terms of both mixing of the MCMC algorithm and accuracy of parameter estimates.

### 6.1.3 Efficient prediction via latent variables

In Chapter 4 we proposed an approximation for predicting ordinal response data at unobserved locations for the spatial PLMM. The approximation was shown to be efficient and accurate. In the Bayesian framework, the method resulted in approximate samples from the posterior predictive distribution of the density function of an unobserved ordinal response variable. From these samples we were able to compute estimates of the approximate posterior predictive distributions,  $P(E(Y(\mathbf{s}_0)|\mathbf{Z}(\mathbf{s})))$  and  $P(\text{Var}(Y(\mathbf{s}_0)|\mathbf{Z}(\mathbf{s})))$ , the conditional expectation and variance of the unobserved ordinal response given the continuous

latent variables. We also obtained samples from the true posterior predictive distributions,  $P(\text{E}(Y(\mathbf{s}_0)|\mathbf{y}(\mathbf{s})))$  and  $P(\text{Var}(Y(\mathbf{s}_0)|\mathbf{y}(\mathbf{s})))$ , the conditional expectation and variance of the unobserved ordinal response given the observed data using (42). We compared the approximation estimates to the estimates of the true posterior predictive distribution by computing the mean squared error (MSE) for out of sample prediction. The MSEs were very similar for the two approaches.

In Chapter A.4 we compared the mixed discrete and continuous model to the discrete-only model by computing the loss for the two models across metric. The loss function (68) was complicated by the fact that in the mixed discrete and continuous model there were both ordinal and continuous response variables and in the discrete-only model there were only ordinal response variables. We computed the squared-error loss using draws from the posterior predictive distribution of the latent response variable,  $\mathbf{Z}$ . This required us to compute a “true value” using (62) for the latent response variable (Chapter 5.4.3). The loss for location,  $\mathbf{s}_i$  and response variable  $j$  was computed using the posterior draws as

$$(Z_j(\mathbf{s}_i)^{(m)} - \hat{Z}_j(\mathbf{s}_i))^2$$

where  $Z_j(\mathbf{s}_i)^{(m)}$  is the  $m$ th draw of  $Z_j(\mathbf{s}_i)$  and  $\hat{Z}_j(\mathbf{s}_i)$  is the “true value” of the continuous representation of the observed ordinal response,  $Y_j(\mathbf{s}_i)$ . We believe that the approximation may be a more appropriate approach for comparing models with multivariate response data of different types when each variable is modeled by a latent Gaussian variable. This is because we can use the observed data as the true value for all data types when estimating prediction error as opposed to having to compute a “true value”. Using the approximation, we could compute the loss as

$$(\text{E}(Y_j(\mathbf{s}_i)|\mathbf{Z}_j(\mathbf{s}))^{(m)} - Y_j(\mathbf{s}_i))^2$$

where  $E(Y_j(\mathbf{s}_i)|\mathbf{Z}(\mathbf{s}))^{(m)}$  is the  $m$ th draw of the approximate posterior predictive distribution of the conditional expectation of  $Y_j(\mathbf{s}_i)$  given the continuous latent variables  $\mathbf{Z}_j(\mathbf{s})$  and  $Y_j(\mathbf{s}_i)$  is the true value of the observed ordinal response. Therefore, we would like to investigate how the approximation performs for other types of variables, not just ordinal response data. Further, we would like to explore the approximation as a method for model selection for multivariate response data where the models differ in response variable type.

## 6.2 Identifiability: Conclusion and future work

In Chapter 3 we discussed identifiability as it applies to frequentist and Bayesian inference. We investigated identifiability of the common factor model for multivariate ordinal response data. We applied a mapping approach between fundamental and reduced-form parameters for checking identifiability in probit linear models. The mapping approach did not extend to the common factor model with second-stage spatial correlation. Therefore, we studied parameter identifiability in spatial probit linear mixed models (PLMMs) for binary and ordinal response data based on asymptotic theory and empirical results of parameter identifiability in LMMs and GLMMs. We used simulations to compare the signal in the likelihood functions of PLMMs for the spatial parameters and related our results to the theoretical results previously developed for spatial GLMMs (Zhang, 2004).

### 6.2.1 Identifiability via mapping approach

The mapping approach for checking identifiability focuses on the functional relationship between structural and reduced-form parameters. For PLMs using latent Gaussian variables, the reduced-form parameters are the means and variance-covariances of the latent variables. To achieve identifiability, we proposed fixing one of the variance parameters and one of the factor-loadings of the common factor. We also omitted the intercept term from the mean of the common factor. Our work only pertained to the common factor model where each variable of the multivariate response was ordinal. To apply to a more general common factor

model, we propose extending the mapping approach to other types of response data that can be fitted using latent Gaussian models. Latent Gaussian models are suitable for many types of data, many of which may be of interest when drawing inference on an unobservable common factor. Therefore, universal recommendations for identifiability of latent Gaussian models with common factors would be valuable to many modelers.

### 6.2.2 Identifiability of spatial probit models

We investigated identifiability for spatial probit models using theoretical work for spatial GLMMs and empirical results for spatial LMMs. The theoretical results for GLMMs show that  $\tau$  and  $\phi$  in (63) are not consistently estimable but the ratio  $\tau/\phi$  is consistently estimable (Zhang, 2004). As future work we would like to prove Zhang's results for PLMMs. The empirical results for continuous response data confirm Zhang's results and show that both  $\tau$  and  $\phi$  are underestimated by maximum likelihood (ML) and restricted maximum likelihood (REML) (Irvine et al. (2007) and Chapter 3). The log-likelihood surface plots for spatial PLMMs (Chapter 3.3.2) suggest that the partial sill and range parameter of the exponential covariance function are weakly identifiable. This is indicated by the modal regions of the log-likelihood surfaces (Figures 3.5 - 3.14). The modal regions have positive correlation suggesting that the ratio  $\tau/\phi$  is identifiable for spatial PLMMs. As the spatial signal increases and the spatial range decreases, we found that the modal region becomes more localized. We compared the log-likelihood surfaces for binary, ordinal, and continuous response data and concluded that each data type underestimated both  $\tau$  and  $\phi$ .

The strength of the signal of the spatial parameters in the likelihood of PLMMs is similar to the strength of the spatial parameter signal found in the likelihood of GLMMs and LMMs. As the spatial variance,  $\tau$ , increased, the difference in signal of the spatial parameters between the data types also increased. The amount of signal in the response likelihood is positively correlated with data richness where richness is defined by the amount of information in the data. Continuous data contain the most information and are therefore the most rich. The

log-likelihood surfaces produced by ordinal response data with 5 categories are similar to those produced by continuous response data (Figures 3.5 - 3.14). This suggests that as the number of ordinal categories increases, the signal of the spatial parameters will increase to that of continuous response data.

In Chapter 4 we compared the first-stage and second-stage spatial probit model (PLMM). Whereas the first-stage spatial probit model is identifiable, the PLMM is a more flexible model for fitting binary and ordinal response data. We showed that under different limiting conditions of the parameters of the spatial covariance function, the approximations of the conditional expectation and variance at an unobserved location are equivalent for the two models. In general, we recommend fitting the PLMM when modeling spatial binary or ordinal response data. We only recommend fitting the first-stage spatial model when all stochasticity in the response data can reasonably be assumed to be spatially correlated. In this case, fitting the first-stage model alleviates the issues associated with fitting a weakly identifiable model.

### 6.2.3 Bayesian learning for spatial probit models

Weak identifiability does not preclude Bayesian learning about parameters from the data. Xie and Carlin (2006) proposed a method to gauge the amount of potential Bayesian learning of weakly identifiable parameters focusing on Gaussian response data and conditional autoregressive (CAR) spatial models. Their methods are not readily applicable to geostatistical and discrete response data because they require prior, posterior, and other conditional densities of the latent variables to be known in closed form. Many of these densities will need to be estimated for spatially correlated non-Gaussian response data. We would first like to extend their ideas to models for other types of data, including ordinal response data. Non-Gaussian spatially correlated data are common in many fields. For example, in disease modeling, observed data can be presence/absence of the disease, count of the number of animals with the disease, or disease prevalence, all of which are non-Gaussian. Since the

richness of the response data varies by type, understanding the potential of Bayesian learning for a given model and data set would be extremely useful.

Geostatistical models are a popular class of spatial models that allow for prediction at unobserved locations. We would also like to extend the estimation of potential Bayesian learning to geostatistical spatial models. If we could discern when a weakly identifiable parameter has very little Bayesian learning potential, it would suggest that the parameter should be fixed and not estimated with MCMC. This would resolve convergence issues of MCMC that result from weakly identifiable parameters. It would also help in assigning informative priors. Priors that are not informative can cause the Markov chain for the weakly identifiable parameters to drift to extreme values impeding parameter estimation. Overly informative priors, however, limit Bayesian learning. Therefore, a better understanding of the potential of Bayesian learning would enhance inference and prediction.

### 6.3 Conclusion

This dissertation advanced both computational efficiency and identifiability in fitting models to spatially correlated ordinal response data. The PX-DA algorithms provided faster convergence of spatial PLMMs and can be adapted easily for models for multivariate response data. We presented a clear differentiation between first-stage and second-stage spatial probit models. We showed that weak identifiability exists in spatial PLMMs and is similar to that in LMMs and GLMMs. We also provided general recommendations for identifiability in common factor models with multivariate mixed response data. Lastly, the multivariate multilevel latent variable model was applied to mixed continuous and ordinal measurements of the biotic condition of wetlands in Colorado. The results are valuable to ecologists for management planning and restoration for wetlands.

## REFERENCES

- Agresti, A. (2002). *Categorical data analysis*, volume 359. Wiley-interscience.
- Albert, J. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669–679.
- Albert, J. and Chib, S. (1997). Bayesian methods for cumulative, sequential and two-step ordinal data regression models. Technical report, Citeseer.
- Banerjee, S., Gelfand, A., Finley, A., and Sang, H. (2008). Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(4):825–848.
- Banerjee, S., Gelfand, A. E., and Carlin, B. P. (2003). *Hierarchical modeling and analysis for spatial data*, volume 101. Chapman and Hall/CRC.
- Barankin, E. W. (1960). Sufficient parameters: solution of the minimal dimensionality problem. *Annals of the Institute of Statistical Mathematics*, 12(2):91–118.
- Basu, A. (1983). Identifiability. *Encyclopedia of Statistical Sciences*, 4:2–6.
- Berger, J. O., De Oliveira, V., and Sansó, B. (2001). Objective Bayesian analysis of spatially correlated data. *Journal of the American Statistical Association*, 96(456):1361–1374.
- Berrett, C. and Calder, C. (2012). Data augmentation strategies for the Bayesian spatial probit regression model. *Computational Statistics & Data Analysis*, 56(3):478–490.
- Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 192–236.
- Besag, J., Green, P., Higdon, D., and Mengerson, K. (1995). Bayesian computation and stochastic systems. *Statistical Science*, pages 3–41.
- Cauchemez, S., Carrat, F., Viboud, C., Valleron, A., and Boelle, P. (2004). A Bayesian MCMC approach to study transmission of influenza: application to household longitudinal data. *Statistics in Medicine*, 23(22):3469–3487.
- Chakraborty, A., Gelfand, A., Wilson, A., Latimer, A., and Silander Jr, J. (2010). Modeling large scale species abundance with latent spatial processes. *The Annals of Applied Statistics*, 4(3):1403–1429.
- Chiu, G., Guttorp, P., Westveld, A., Khan, S., and Liang, J. (2011). Latent health factor index: A statistical modeling approach for ecological health assessment. *Environmetrics*, 22(3):243–255.

- Christensen, O. F., Roberts, G. O., and Sköld, M. (2006). Robust Markov chain Monte Carlo methods for spatial generalized linear mixed models. *Journal of Computational and Graphical Statistics*, 15(1):1–17.
- Christensen, W. and Amemiya, Y. (2002). Latent variable analysis of multivariate spatial data. *Journal of the American Statistical Association*, 97(457):302–317.
- Cowles, M. (1996). Accelerating Monte Carlo Markov chain convergence for cumulative-link generalized linear models. *Statistics and Computing*, 6(2):101–111.
- Cressie, N. and Johannesson, G. (2008). Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):209–226.
- Dahl, T. (2011). *Status and trends of wetlands in the conterminous United States 2004 to 2009*. US Dept. of the Interior, US Fish and Wildlife Service.
- Dawid, A. P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–31.
- De Oliveira, V. (1997). Prediction in some classes of non-Gaussian random fields. *Unpublished Ph. D. Thesis, University of Maryland-College Park*.
- De Oliveira, V. (2000). Bayesian prediction of clipped Gaussian random fields. *Computational Statistics & Data Analysis*, 34(3):299–314.
- Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38.
- Diggle, P. J. and Ribeiro, P. J. (2007). *Model-based geostatistics*. Springer.
- Dupacová, J. and Wold, H. (1982). On some identification problems in ML modeling of systems with indirect observation. *Systems Under Indirect Observations: Causality, Structure, Prediction*, 2.
- Eberly, L. E. and Carlin, B. P. (2000). Identifiability and convergence issues for Markov chain Monte Carlo fitting of spatial models. *Statistics in Medicine*, 19(1718):2279–2294.
- Furrer, R., Genton, M., and Nychka, D. (2006). Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics*, 15(3):502–523.
- Gelfand, A. E. and Sahu, S. K. (1999). Identifiability, improper priors, and Gibbs sampling for generalized linear models. *Journal of the American Statistical Association*, 94(445):247–253.
- Gelfand, A. E. and Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):398–409.
- Gelfand, A. J., Ravishanker, N., and Ecker, M. D. (2000). Point-referenced binary spatial data. *Generalized Linear Models: A Bayesian Perspective*, 5:373.

- Gelman, A., Carlin, J., Stern, H., and Rubin, D. (2004). *Bayesian data analysis*. CRC press.
- Givens, G. H. and Hoeting, J. A. (2012). *Computational statistics*, volume 708. Wiley.
- Heagerty, P. J. and Lele, S. R. (1998). A composite likelihood approach to binary spatial data. *Journal of the American Statistical Association*, 93(443):1099–1111.
- Higgs, M. and Hoeting, J. (2010). A clipped latent variable model for spatially correlated ordered categorical data. *Computational Statistics & Data Analysis*, 54(8):1999–2011.
- Imai, K. and Van Dyk, D. (2005). A Bayesian analysis of the multinomial probit model using marginal data augmentation. *Journal of Econometrics*, 124(2):311–334.
- Irvine, K. M., Gitelman, A. I., and Hoeting, J. A. (2007). Spatial designs and properties of spatial correlation: effects on covariance estimation. *Journal of Agricultural, Biological, and Environmental statistics*, 12(4):450–469.
- Johnson, V. E. and Albert, J. H. (1999). *Ordinal data modeling*. Springer.
- Kadane, J. B. (1974). The role of identification in Bayesian theory. *Studies in Bayesian Econometrics and Statistics*, 1.
- Karr, J. (1981). Assessment of biotic integrity using fish communities. *Fisheries*, 6(6):21–27.
- Lemly, J. and Gillian, L. (2012). *North Platte wetland profile and condition assessment*. Colorado Natural Heritage Program, Colorado State University, Fort Collins, Colorado.
- Lemly, J., Gillian, L., and Fink, M. (2011). *Statewide strategies to improve effectiveness in protecting and restoring Colorado's wetland resource*. Colorado Natural Heritage Program, Colorado State University, Fort Collins, Colorado.
- Liang, F., Liu, C., and Carroll, R. (2011). *Advanced Markov chain Monte Carlo methods: learning from past samples*, volume 714. Wiley.
- Lindgren, F., Rue, H., and Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498.
- Lindley, D. V. (1972). *Bayesian statistics: A review*, volume 2. SIAM.
- Liu, C., Rubin, D., and Wu, Y. (1998). Parameter expansion to accelerate EM: the PX-EM algorithm. *Biometrika*, 85(4):755–770.
- Liu, J. and Wu, Y. (1999). Parameter expansion for data augmentation. *Journal of the American Statistical Association*, 94(448):1264–1274.
- Liu, X., Wall, M., and Hodges, J. (2005). Generalized spatial structural equation models. *Biostatistics*, 6(4):539–557.

- Mardia, K. V. and Marshall, R. (1984). Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika*, 71(1):135–146.
- Meng, X.-L. and Van Dyk, D. A. (1999). Seeking efficient data augmentation schemes via conditional and marginal augmentation. *Biometrika*, 86(2):301–320.
- Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, 49(1):115–132.
- Oliveira, V. D. (2000). Bayesian prediction of clipped Gaussian random fields. *Computational Statistics & Data Analysis*, 34(3):299–314.
- R Development Core Team (2007). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org>.
- Rencher, A. (2002). *Methods of Multivariate Analysis, Second Edition*. John Wiley & Sons, Inc.
- Roberts, G. O. and Sahu, S. K. (1997). Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(2):291–317.
- Rocchio, J. (2007). *Floristic quality assessment indices for Colorado plant communities*. Colorado Natural Heritage Program, Colorado State University, Fort Collins, Colorado.
- Rothenberg, T. (1971). Identification in parametric models. *Econometrica: Journal of the Econometric Society*, pages 577–591.
- Royle, J., Dorazio, R., and Link, W. (2007). Analysis of multinomial models with unknown index using data augmentation. *Journal of Computational and Graphical Statistics*, 16(1):67–85.
- Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319–392.
- Schliep, E. M. and Hoeting, J. A. (2013). Multilevel latent Gaussian process model for mixed discrete and continuous multivariate response data. *Journal of Agricultural, Biological, and Environmental Statistics*, pages 1–22.
- Schmidt, A., Conceição, M., and Moreira, G. (2008). Investigating the sensitivity of Gaussian processes to the choice of their correlation function and prior specifications. *Journal of Statistical Computation and Simulation*, 78(8):681–699.
- Schmidt, A. M. and Gelfand, A. E. (2003). A Bayesian coregionalization approach for multivariate pollutant data. *Journal of Geophysical Research: Atmospheres (1984–2012)*, 108(D24).

- Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. CRC Press.
- Spearman, C. (1904). "general intelligence," objectively determined and measured. *The American Journal of Psychology*, 15(2):201–292.
- Stein, M. (1990). Uniform asymptotic optimality of linear predictions of a random field using an incorrect second-order structure. *The Annals of Statistics*, pages 850–872.
- Stein, M. L. (1988). Asymptotically efficient prediction of a random field with a misspecified covariance function. *The Annals of Statistics*, pages 55–63.
- Stein, M. L. (1999). *Interpolation of spatial data: some theory for kriging*. Springer Verlag.
- Stevens, D. and Olsen, A. (2004). Spatially balanced sampling of natural resources. *Journal of the American Statistical Association*, 99(465).
- Swartz, T. B., Haitovsky, Y., Vexler, A., and Yang, T. Y. (2004). Bayesian identifiability and misclassification in multinomial data. *Canadian Journal of Statistics*, 32(3):285–302.
- Swendsen, R. and Wang, J. (1987). Nonuniversal critical dynamics in Monte Carlo simulations. *Physical Review Letters*, 58(2):86.
- Swink, F. and Wilhem, G. (1994). Plants of the Chicago Region. 4th edition. *Indiana Academy of Science, Indianapolis, IN*.
- Taft, J., Wilhelm, G., Ladd, D., and Masters, L. (1997). *Floristic quality assessment for vegetation in Illinois, a method for assessing vegetation integrity*. Illinois Native Plant Society.
- Tanner, M. and Wong, W. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82(398):528–540.
- Wald, A. (1950). Note on the identification of economic relations. *Statistical Inference in Dynamic Economic Models, Cowles Commission Monograph*, 10:238–244.
- Xie, Y. and Carlin, B. P. (2006). Measures of Bayesian learning and identifiability in hierarchical models. *Journal of Statistical Planning and Inference*, 136(10):3458–3477.
- Zhang, H. (2004). Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *Journal of the American Statistical Association*, 99(465):250–261.

## APPENDIX

### A.1 Description of data used in non-spatial probit models

We apply the non-spatial probit model to ordinal response data with 5 categories in Sections 2.3 and 2.4.2. The response variable is a measure of biotic integrity of wetlands in Colorado (Lemly and Gillian, 2012). Each observation also contains covariate information. We include two covariates, elevation and closed-tree canopy, that are known to be correlated with biotic integrity of wetlands. We refer to the covariates as  $X_1$  and  $X_2$ , respectively. Table A.1 gives the counts of observations in each of the 5 categories of the ordinal response. Figure A.1 indicates that both covariates are positively correlated with the ordinal response.

Table A.1: Observed ordinal response values.

	1	2	3	4	5	Total
Observed counts	25	47	65	26	69	232

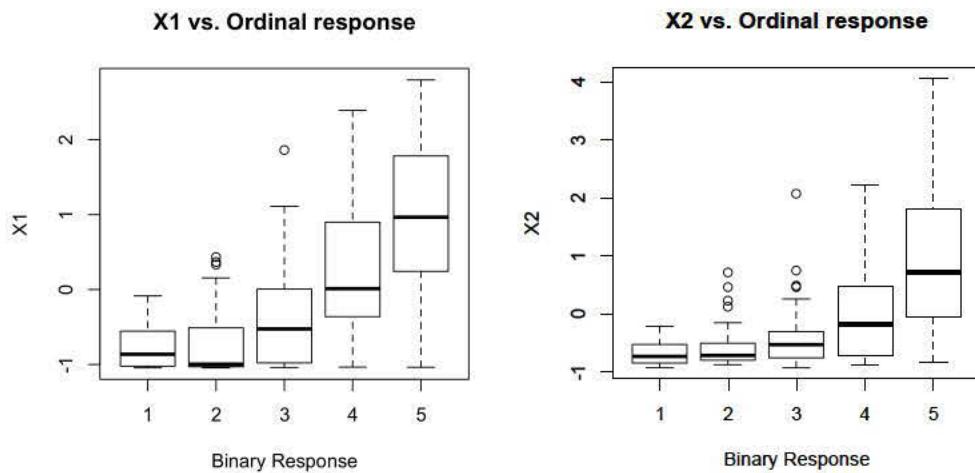


Figure A.1: Boxplots of covariates by ordinal response data showing a positive correlation between both  $X_1$  and  $X_2$  and response,  $y$ .

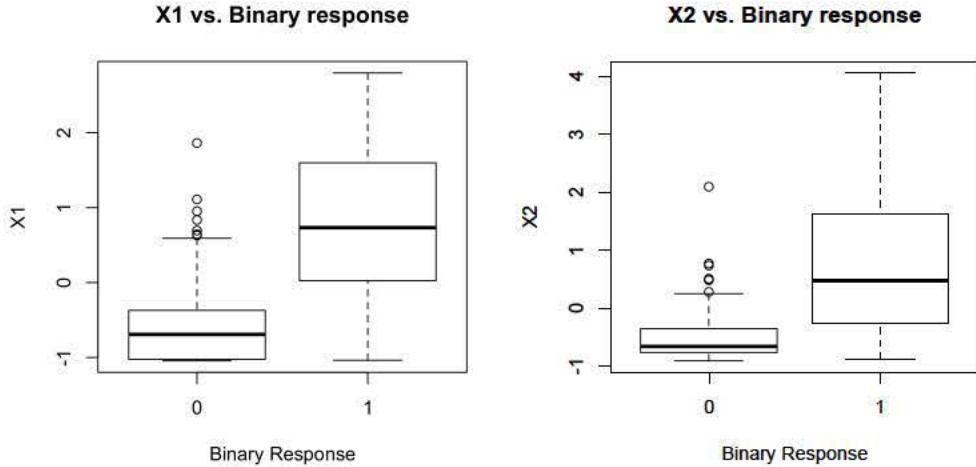


Figure A.2: Boxplots of covariate 1 and 2 by observed binary response. They indicate a positive correlation between both covariates and the binary response,  $y$ .

Table A.2: Observed binary response values.

	0	1	Total
Observed counts	137	95	232

The same data are used in the binary response examples where the response variable takes on one of two values. We further bin the ordinal response such that  $y = 0$  when the ordinal value is in  $\{1, 2, 3\}$  and  $y = 1$  when the ordinal value is in  $\{4, 5\}$ . The counts of the observed data and boxplots for each covariate in the binary response case are shown in Table A.2 and Figure A.2, respectively.

Table A.3: Posterior medians and 95% credible intervals for  $\boldsymbol{\beta}$  and  $\boldsymbol{\lambda}$  using the PDA algorithm (A.5) and both PX-PDA algorithms (A.6 and 9) for ordinal data with five categories.

Parameter	Algorithms		
	PDA (A.5)	RV-PX-PDA (A.6)	RT-PX-PDA (A.9)
$\beta_0$	2.01 (1.76, 2.26)	2.00 (1.73, 2.28)	2.01 (1.77, 2.26)
$\beta_1$	0.90 (0.69, 1.11)	0.89 (0.69, 1.11)	0.90 (0.69, 1.11)
$\beta_2$	0.71 (0.49, 0.94)	0.71 (0.49, 0.94)	0.71 (0.49, 0.94)
$\lambda_2$	1.00 (0.81, 1.20)	0.99 (0.79, 1.21)	1.00 (0.81, 1.20)
$\lambda_3$	2.22 (1.96, 2.49)	2.21 (1.95, 2.48)	2.22 (1.97, 2.48)
$\lambda_4$	2.87 (2.57, 3.18)	2.86 (2.57, 3.17)	2.87 (2.58, 3.17)

## A.2 Posterior inference

### A.2.1 Posterior derivations for $\sigma^2$ in RV-PX-PDA

The conditional posterior distribution for  $\sigma^2$  is

$$\begin{aligned}
p(\sigma^2 | \mathbf{y}, \tilde{\mathbf{Z}}, \boldsymbol{\lambda}) &\propto p(\sigma^2) \int p(\tilde{\mathbf{Z}} | \tilde{\boldsymbol{\beta}}, \sigma^2) p(\tilde{\boldsymbol{\beta}} | \sigma^2) d\tilde{\boldsymbol{\beta}} \\
&= p(\sigma^2) \int \frac{1}{(2\pi)^{n/2}} |\sigma^2 \mathbf{I}|^{-1/2} \exp \left[ -\frac{1}{2} (\tilde{\mathbf{Z}} - \mathbf{X}' \tilde{\boldsymbol{\beta}})' (\sigma^2 \mathbf{I})^{-1} (\tilde{\mathbf{Z}} - \mathbf{X}' \tilde{\boldsymbol{\beta}}) \right] \\
&\quad \times \frac{1}{(2\pi)^{p/2}} |\sigma^2 \mathbf{I}|^{-1/2} \exp \left[ -\frac{1}{2} \tilde{\boldsymbol{\beta}}' (\sigma^2 \Sigma_{\boldsymbol{\beta}})^{-1} \tilde{\boldsymbol{\beta}} \right] d\tilde{\boldsymbol{\beta}} \\
&= p(\sigma^2) \frac{1}{(2\pi)^{n/2}} |\sigma^2 \mathbf{I}|^{-1/2} \exp \left[ -\frac{1}{2} \tilde{\mathbf{Z}}' (\sigma^2 \mathbf{I})^{-1} \tilde{\mathbf{Z}} \right] \int \frac{1}{(2\pi)^{p/2}} |\sigma^2 \mathbf{I}|^{-1/2} \\
&\quad \times \exp \left[ -\frac{1}{2} \left( \tilde{\boldsymbol{\beta}}' (\mathbf{X}' (\sigma^2 \mathbf{I})^{-1} \mathbf{X} + (\sigma^2 \Sigma_{\boldsymbol{\beta}})^{-1}) \tilde{\boldsymbol{\beta}} - 2\tilde{\boldsymbol{\beta}}' \mathbf{X}' (\sigma^2 \mathbf{I})^{-1} \tilde{\mathbf{Z}} \right) \right] d\tilde{\boldsymbol{\beta}} \quad (64) \\
&= p(\sigma^2) \frac{1}{(2\pi)^{n/2}} |\sigma^2 \mathbf{I}|^{-1/2} \\
&\quad \times \exp \left[ -\frac{1}{2} \left( \tilde{\mathbf{Z}}' (\sigma^2 \mathbf{I})^{-1} \tilde{\mathbf{Z}} - (\sigma^2) \tilde{\mathbf{Z}}' \mathbf{X} (\mathbf{X}' \mathbf{X} + \Sigma_{\boldsymbol{\beta}}^{-1})^{-1} \mathbf{X}' \tilde{\mathbf{Z}} \right) \right] \\
&= p(\sigma^2) \frac{1}{(2\pi)^{n/2}} |\sigma^2 \mathbf{I}|^{-1/2} \exp \left[ -\frac{1}{2\sigma^2} \left( \tilde{\mathbf{Z}}' \tilde{\mathbf{Z}} - \tilde{\mathbf{Z}}' \mathbf{X} (\mathbf{X}' \mathbf{X} + \Sigma_{\boldsymbol{\beta}}^{-1})^{-1} \mathbf{X}' \tilde{\mathbf{Z}} \right) \right]
\end{aligned}$$

Therefore, since the conjugate prior distribution for  $\sigma^2$  as  $p(\sigma^2) \sim \text{Inv. Gamma}(\alpha_s, \beta_s)$ ,

$$p(\sigma^2 | \mathbf{y}, \tilde{\mathbf{Z}}, \boldsymbol{\lambda}) \propto (\sigma^2)^{-\frac{n}{2} - \alpha_s - 1} \exp \left[ -\frac{1}{\sigma^2} \left( \beta_s + \frac{1}{2} \left( \tilde{\mathbf{Z}}' \tilde{\mathbf{Z}} - \tilde{\mathbf{Z}}' \mathbf{X} (\mathbf{X}' \mathbf{X} + \Sigma_{\beta}^{-1})^{-1} \mathbf{X}' \tilde{\mathbf{Z}} \right) \right) \right].$$

At iteration  $t$  within the MCMC,  $(\sigma^2)^t$  is drawn from

$$p(\sigma^2 | \mathbf{y}, \tilde{\mathbf{Z}}^t, \boldsymbol{\lambda}^t) \sim \text{Inv. Gamma} \left( \alpha_s + \frac{n}{2}, \beta_s + \frac{1}{2} \left( (\tilde{\mathbf{Z}}^t)' \tilde{\mathbf{Z}}^t - \tilde{\mathbf{Z}}^t \mathbf{X} (\mathbf{X}' \mathbf{X} + \Sigma_{\beta}^{-1})^{-1} \mathbf{X}' \tilde{\mathbf{Z}}^t \right) \right).$$

In the spatial PLMM, similar derivations are required to obtain the conditional posterior distribution of  $\sigma^2$ . First note that the posteriors can be simplified to

$$p(\sigma^2 | \mathbf{y}, \tilde{\mathbf{Z}}, \tau, \phi) \propto p(\sigma^2 | \mathbf{y}, \tilde{\mathbf{Z}})$$

and

$$p(\sigma^2 | \mathbf{y}, \tilde{\mathbf{Z}}, \tau, \phi, \boldsymbol{\lambda}) \propto p(\sigma^2 | \mathbf{y}, \tilde{\mathbf{Z}}, \boldsymbol{\lambda}).$$

Second, the resulting posterior is the same for both binary and ordinal data. Therefore, the posterior for the PX-DA algorithms is written as

$$\begin{aligned} p(\sigma^2 | \mathbf{y}, \tilde{\mathbf{Z}}^t) &\propto p(\sigma^2 | \mathbf{y}, \tilde{\mathbf{Z}}^t, \boldsymbol{\lambda}^t) \\ &\sim \text{Inv. Gamma} \left( \alpha_s + \frac{n}{2}, \beta_s + \frac{1}{2} \left( (\tilde{\mathbf{Z}}^t)' \mathbf{P}^t (\tilde{\mathbf{Z}}^t - \mathbf{X} (\Sigma_{\beta}^{-1} + \mathbf{X}' \mathbf{P}^t \mathbf{X})^{-1} \mathbf{X}' \mathbf{P}^t \tilde{\mathbf{Z}}^t) \right) \right) \end{aligned}$$

where  $\mathbf{P}^t = \mathbf{I} - (\mathbf{I} + (\tau^t \mathbf{R}^t)^{-1})^{-1}$ .

For the PX<sup>2</sup>-DA algorithms (e.g., Algorithm A.12), we must compute both the posterior  $p(\sigma^2 | \mathbf{y}, \tilde{\mathbf{Z}}, \widetilde{\mathbf{W}})$  where  $\tilde{\boldsymbol{\beta}}$  is integrated out and  $p(\sigma^2 | \mathbf{y}, \tilde{\mathbf{Z}}, \widetilde{\boldsymbol{\beta}})$  where  $\widetilde{\mathbf{W}}$  is integrated out. Again, note that

$$p(\sigma^2 | \mathbf{y}, \tilde{\mathbf{Z}}, \tilde{\mathbf{W}}) = p(\sigma^2 | \mathbf{y}, \tilde{\mathbf{Z}}, \tilde{\mathbf{W}}, \boldsymbol{\lambda})$$

and

$$p(\sigma^2 | \mathbf{y}, \tilde{\mathbf{Z}}, \tilde{\boldsymbol{\beta}}) = p(\sigma^2 | \mathbf{y}, \tilde{\mathbf{Z}}, \tilde{\boldsymbol{\beta}}, \boldsymbol{\lambda}).$$

The first posterior is

$$\begin{aligned} p(\sigma^2 | \mathbf{y}, \tilde{\mathbf{Z}}^t, \tilde{\mathbf{W}}^{t-1}) &\sim \text{Inv. Gamma}(\alpha_s + n, \\ &\quad \beta_s + \frac{1}{2} \left( (\tilde{\mathbf{W}}^{t-1})' (\tau^{t-1} \mathbf{R}^{t-1})^{-1} \tilde{\mathbf{W}}^{t-1} + (\tilde{\mathbf{Z}}^t - \tilde{\mathbf{W}}^{t-1})' \mathbf{P}^{t-1} (\tilde{\mathbf{Z}}^t - \tilde{\mathbf{W}}^{t-1}) \right) \end{aligned}$$

where  $\mathbf{P}^t = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X} + \Sigma_{\beta}^{-1})^{-1}\mathbf{X}'$ .

The second posterior is

$$\begin{aligned} p(\sigma^2 | \mathbf{y}, \tilde{\mathbf{Z}}^t, \tilde{\boldsymbol{\beta}}^t) &\sim \text{Inv. Gamma}(\alpha_s + \frac{n}{2} + \frac{m}{2}, \\ &\quad \beta_s + \frac{1}{2} \left( (\tilde{\boldsymbol{\beta}}^t)' \Sigma_{\beta}^{-1} \tilde{\boldsymbol{\beta}}^t + (\tilde{\mathbf{Z}}^t - \mathbf{X}\tilde{\boldsymbol{\beta}}^t)' \mathbf{P}^{t-1} (\tilde{\mathbf{Z}}^t - \mathbf{X}\tilde{\boldsymbol{\beta}}^t) \right) \end{aligned}$$

where  $\mathbf{P}^t = \mathbf{I} - (\mathbf{I} + (\tau^t \mathbf{R}^t)^{-1})^{-1}$  and  $m$  is the number of covariates in the model.

### A.2.2 Posterior derivations for $\lambda$ in PX-DA for binary response data

The conditional posterior distribution for  $\lambda$  is

$$\begin{aligned}
p(\lambda|\mathbf{y}, \mathbf{Z}) &= \frac{p(\mathbf{y}, \mathbf{Z}, \lambda)}{p(\mathbf{y}, \mathbf{Z})} \\
&= \frac{\int p(\mathbf{y}, \mathbf{Z}, \boldsymbol{\beta}, \lambda) d\boldsymbol{\beta}}{p(\mathbf{y}, \mathbf{Z})} \\
&= \frac{\int p(\mathbf{Z}|\mathbf{y}, \boldsymbol{\beta}, \lambda) p(\mathbf{y}, \boldsymbol{\beta}, \lambda) d\boldsymbol{\beta}}{p(\mathbf{y}, \mathbf{Z})} \\
&\propto \int p(\mathbf{Z}|\mathbf{y}, \boldsymbol{\beta}, \lambda) p(\mathbf{y}|\boldsymbol{\beta}, \lambda) p(\boldsymbol{\beta}, \lambda) d\boldsymbol{\beta} \\
&\propto \int p(\mathbf{Z}|\mathbf{y}, \boldsymbol{\beta}, \lambda) p(\mathbf{y}|\boldsymbol{\beta}, \lambda) p(\boldsymbol{\beta}) p(\lambda) d\boldsymbol{\beta}.
\end{aligned} \tag{65}$$

where the last line holds because  $\boldsymbol{\beta}$  and  $\lambda$  are assumed independent a priori. In the binary response setting, we defined the sets  $C_0$  and  $C_1$  such that  $C_0 = \{i : y_i = 0\}$  and  $C_1 = \{i : y_i = 1\}$  for  $i \in 1, \dots, n$  (Section 2.4.1). Assigning a normal prior distribution to  $\lambda$  such that  $p(\lambda) \sim N(0, L)$ , we can write the posterior distribution in (65) as

$$\begin{aligned}
p(\lambda|\mathbf{y}, \mathbf{Z}) &\propto p(\lambda) \int \prod_{i \in C_0} \left[ \frac{\exp \left[ -\frac{1}{2}(Z_i - \lambda - \mathbf{X}_i \boldsymbol{\beta})'(Z_i - \lambda - \mathbf{X}_i \boldsymbol{\beta}) \right]}{\Phi(-\mathbf{X}_i \boldsymbol{\beta})} I_{[Z_i \leq \lambda]} \right] \\
&\quad \times \prod_{i \in C_1} \left[ \frac{\exp \left[ -\frac{1}{2}(Z_i - \lambda - \mathbf{X}_i \boldsymbol{\beta})'(Z_i - \lambda - \mathbf{X}_i \boldsymbol{\beta}) \right]}{1 - \Phi(-\mathbf{X}_i \boldsymbol{\beta})} I_{[Z_i > \lambda]} \right] \\
&\quad \times \prod_{i=1}^n [\Phi(-\mathbf{X}_i \boldsymbol{\beta})^{1-y_i} (1 - \Phi(-\mathbf{X}_i \boldsymbol{\beta}))^{y_i}] \exp \left[ -\frac{1}{2} \boldsymbol{\beta}' \Sigma_{\boldsymbol{\beta}}^{-1} \boldsymbol{\beta} \right] d\boldsymbol{\beta} \\
&\propto p(\lambda) \int \exp \left[ -\frac{1}{2} (\mathbf{Z} - \mathbf{1}\lambda - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Z} - \mathbf{1}\lambda - \mathbf{X}\boldsymbol{\beta}) \right] \\
&\quad \times \exp \left[ -\frac{1}{2} \boldsymbol{\beta}' \Sigma_{\boldsymbol{\beta}}^{-1} \boldsymbol{\beta} \right] d\boldsymbol{\beta} \prod_{i \in C_0} I_{[Z_i \leq \lambda]} \prod_{i \in C_1} I_{[Z_i > \lambda]} \\
&\propto p(\lambda) \int \exp \left[ -\frac{1}{2} (\boldsymbol{\beta}' \mathbf{X}' \mathbf{X} \boldsymbol{\beta} - 2\boldsymbol{\beta}' \mathbf{X}' (\mathbf{Z} - \mathbf{1}\lambda) + (\mathbf{Z} - \mathbf{1}\lambda)'(\mathbf{Z} - \mathbf{1}\lambda) \right. \\
&\quad \left. + \boldsymbol{\beta}' \Sigma_{\boldsymbol{\beta}}^{-1} \boldsymbol{\beta}) \right] d\boldsymbol{\beta} \times \prod_{i \in C_0} I_{[Z_i \leq \lambda]} \prod_{i \in C_1} I_{[Z_i > \lambda]} \\
&\propto p(\lambda) \exp \left[ -\frac{1}{2} ((\mathbf{Z} - \mathbf{1}\lambda)'(\mathbf{Z} - \mathbf{1}\lambda) \right. \\
&\quad \left. - (\mathbf{X}'(\mathbf{Z} - \mathbf{1}\lambda))'(\mathbf{X}' \mathbf{X} + \Sigma_{\boldsymbol{\beta}}^{-1})^{-1} \mathbf{X}(\mathbf{Z} - \mathbf{1}\lambda)) \right] \times \prod_{i \in C_0} I_{[Z_i \leq \lambda]} \prod_{i \in C_1} I_{[Z_i > \lambda]} \\
&\propto \exp \left[ -\frac{1}{2} (\lambda' (L^{-1} + \mathbf{1}' \mathbf{1} - \mathbf{1}' \mathbf{X} (\mathbf{X}' \mathbf{X} + \Sigma_{\boldsymbol{\beta}}^{-1})^{-1} \mathbf{X}' \mathbf{1}) \lambda \right. \\
&\quad \left. - 2\lambda' (\mathbf{1}' \mathbf{Z} - \mathbf{1}' \mathbf{X} (\mathbf{X}' \mathbf{X} + \Sigma_{\boldsymbol{\beta}}^{-1})^{-1} \mathbf{X}' \mathbf{Z})) \right] \times \prod_{i \in C_0} I_{[Z_i \leq \lambda]} \prod_{i \in C_1} I_{[Z_i > \lambda]}
\end{aligned} \tag{66}$$

The resulting distribution is truncated-normal and can be written as

$$p(\lambda|\mathbf{y}, \mathbf{Z}) \sim TN(\mu_{\lambda}, \tau_{\lambda}, \max\{Z_i : i \in A_0\}, \min\{Z_i : i \in C_1\}), \text{ where}$$

$$\begin{aligned}
\mu_{\lambda} &= (L^{-1} + n - \mathbf{1}' \mathbf{X} (\mathbf{X}' \mathbf{X} + \Sigma_{\boldsymbol{\beta}}^{-1})^{-1} \mathbf{X}' \mathbf{1})^{-1} (\mathbf{1}' \mathbf{Z} - \mathbf{1}' \mathbf{X} (\mathbf{X}' \mathbf{X} + \Sigma_{\boldsymbol{\beta}}^{-1})^{-1} \mathbf{X}' \mathbf{Z}) \text{ and} \\
\tau_{\lambda} &= (L^{-1} + n - \mathbf{1}' \mathbf{X} (\mathbf{X}' \mathbf{X} + \Sigma_{\boldsymbol{\beta}}^{-1})^{-1} \mathbf{X}' \mathbf{1})^{-1}.
\end{aligned} \tag{67}$$

In the ordinal response setting, we defined the sets  $A_k$  such that  $A_k = \{i : y_i = k\}$  for  $k \in 1, \dots, K$ . To be consistent with the notation in Algorithm 9, we draw  $\alpha_1$  from the conditional distribution  $p(\alpha_1 | \mathbf{y}, \mathbf{Z}, \boldsymbol{\alpha})$ . That is,

$$p(\alpha_1 | \mathbf{y}, \mathbf{Z}, \boldsymbol{\alpha}) \sim TN(\mu_\alpha, \tau_\alpha, \max\{Z_i : i \in C_1\}, \min\{Z_i : i \in C_2\}),$$

where  $\mu_\alpha$  and  $\tau_\alpha$  are equal to  $\mu_\lambda$  and  $\tau_\lambda$  in (67).

### A.3 Observed Data

The frequency of the observed ordinal response values for each metric over all  $n = 232$  locations is summarized in Table A.4. Figures A.3 and A.4 show univariate summaries between the each ordinal response and the covariates.

Table A.4: Observed response data by metric

Metric	Ordinal response				
	1	2	3	4	5
Native plant cover	9	16	60	69	75
Noxious weed cover	1	6	10	50	165
Aggressive native cover	1	4	4	3	220
Structural diversity	5	15	82	108	22
Floristic quality	24	47	65	26	69

### A.4 Squared error loss

The discrete-only model and the mixed response model are compared by computing the median squared error loss using the posterior predictions of the latent response  $\mathbf{Z}$ . The “true” value of  $\mathbf{Z}$  for the ordinal metrics is estimated using (62). To compare squared error loss across models and metrics, we scale each loss value by the variance of its “true” value of  $\mathbf{Z}$ . The standardized loss for each location  $i$  and metric  $j$  is computed using the posterior

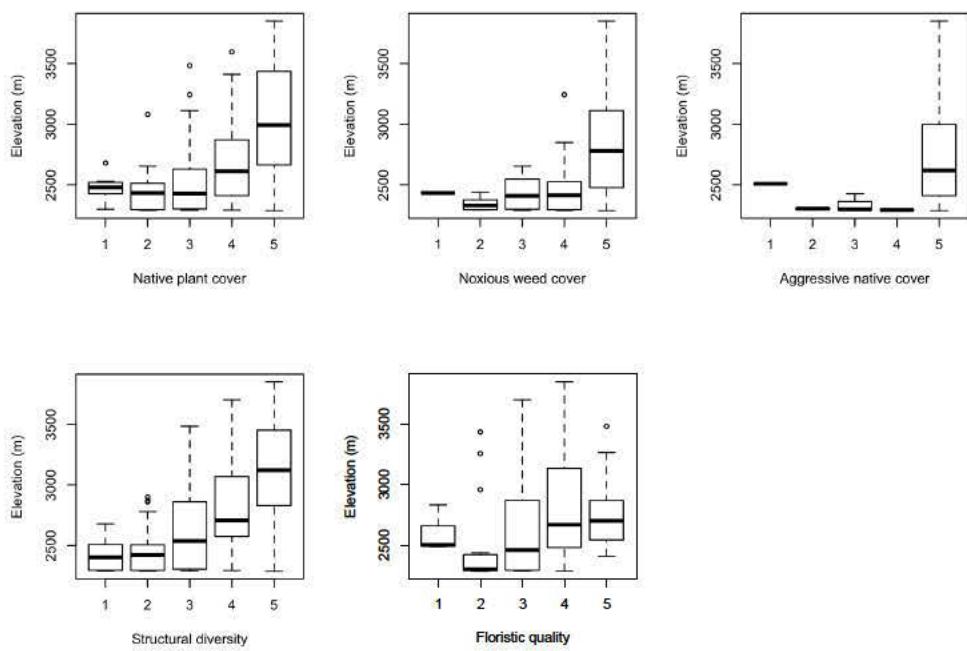


Figure A.3: Boxplots of elevation (y-axis) for each ordinal response (x-axis) for each metric.

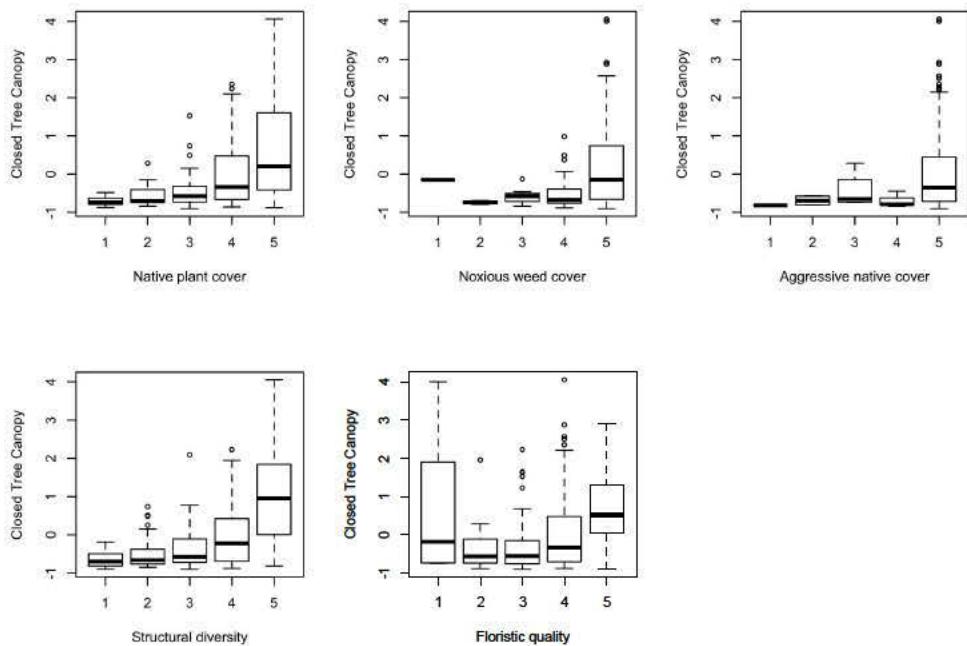


Figure A.4: Boxplots of closed tree canopy (y-axis) for each ordinal response (x-axis) for each metric.

draws as

$$\frac{(Z_{ij}^{(m)} - \hat{Z}_{ij})^2}{\hat{\sigma}_{\hat{\mathbf{Z}}_j}^2} \quad (68)$$

where  $Z_{ij}^{(m)}$  is the  $m$ th draw of  $Z_{ij}$  and  $\hat{Z}_{ij}$  is the true value of the continuous representation of the observed ordinal response,  $Y_{ij}$ . For a continuous response metric,  $\hat{\sigma}_{\hat{\mathbf{Z}}_j}^2$  is the variance of the response vector  $\mathbf{Y}_j$  since  $\hat{\mathbf{Z}}_j$  is observed. For an ordinal response metric,  $\hat{\sigma}_{\hat{\mathbf{Z}}_j}^2$  is the variance of  $\hat{\mathbf{Z}}_j$ , which is based on the MCMC draws. The resulting loss for each metric is similar between the discrete-only model and the mixed response model (Table A.5).

Table A.5: Median squared error loss comparison between the two models.

Metric	Discrete-Only Model Loss Estimate	Mixed Response Model Loss Estimate
Native plant cover	0.73	0.75
Noxious weed cover	0.74	0.83
Aggressive native cover	1.70	1.04
Structural diversity	0.96	0.93
Floristic quality	0.58	0.37

## A.5 Multiple correlation

Canonical correlation analysis is a method used to measure the linear relationship between two sets of variables (Rencher, 2002, Chapter 11). It is an extension of multiple correlation, which is the correlation between  $y$  and a set of  $x$ 's. Canonical correlation is often used to evaluate the relationship between a set of response variables  $\mathbf{y} = \{\mathbf{y}_1, \dots, \mathbf{y}_p\}$  and a set of predictor variables  $\mathbf{x} = \{\mathbf{x}_1, \dots, \mathbf{x}_q\}$ , each of which is measured on the same set of sampling

units. It is useful to first partition the covariance matrix of the vectors  $\mathbf{y}$  and  $\mathbf{x}$  as

$$\mathbf{S} = \begin{pmatrix} \mathbf{S}_{yy} & \mathbf{S}_{yx} \\ \mathbf{S}_{xy} & \mathbf{S}_{xx} \end{pmatrix}$$

where  $\mathbf{S}_{yy}$  is the  $p \times p$  sample covariance matrix of  $\mathbf{y}$ ,  $\mathbf{S}_{yx}$  is the  $p \times q$  sample covariance matrix between  $\mathbf{y}$  and  $\mathbf{x}$ , and  $\mathbf{S}_{xx}$  is the  $q \times q$  sample covariance matrix of  $\mathbf{x}$ . A measure of association between  $\mathbf{y}$  and  $\mathbf{x}$  is  $R_M^2$ . This value is analogous to  $R^2$  in linear regression. It can be written as

$$R_M^2 = |\mathbf{S}_{yy}^{-1} \mathbf{S}_{yx} \mathbf{S}_{xx}^{-1} \mathbf{S}_{xy}| = \prod_{i=1}^m r_i^2$$

where  $m = \min(p, q)$  and  $r_1^2, \dots, r_m^2$  are the eigenvalues of  $\mathbf{S}_{yy}^{-1} \mathbf{S}_{yx} \mathbf{S}_{xx}^{-1} \mathbf{S}_{xy}$ . The eigenvalues provide meaningful measures of association between  $\mathbf{y}$  and  $\mathbf{x}$ . The largest eigenvalue is the best overall measure of association in that it represents the maximum squared correlation between a linear combination of the  $\mathbf{y}'s$  and a linear combination of the  $\mathbf{x}'s$ .

The square roots of the eigenvalues are called canonical correlations. The canonical correlations can also be obtained by computing the eigenvalues of the correlation matrix. We can partition the correlation matrix as

$$\mathbf{R} = \begin{pmatrix} \mathbf{R}_{yy} & \mathbf{R}_{yx} \\ \mathbf{R}_{xy} & \mathbf{R}_{xx} \end{pmatrix} \quad (69)$$

where  $\mathbf{R}_{yy}$  is the  $p \times p$  sample correlation matrix of  $\mathbf{y}$ ,  $\mathbf{R}_{yx}$  is the  $p \times q$  sample correlation matrix between the  $\mathbf{y}$  and  $\mathbf{x}$ , and  $\mathbf{R}_{xx}$  is the  $q \times q$  sample correlation matrix of  $\mathbf{x}$ . The canonical correlations can also be computed from  $\mathbf{R}$  in (69) since the eigenvalues from  $\mathbf{R}$  and the matrix  $\mathbf{S}_{yy}^{-1} \mathbf{S}_{yx} \mathbf{S}_{xx}^{-1} \mathbf{S}_{xy}$  are equal.

While it is useful to look at the overall association between  $\mathbf{y}$  and  $\mathbf{x}$ , we would also like to evaluate the correlation between each  $\mathbf{y}_j$  and  $\mathbf{x}$  separately. The multiple correlation,  $R_{y_j|\mathbf{x}}$ ,

of  $\mathbf{y}_j$  and  $\mathbf{x}$  is computed as the square root of

$$R_{y_j|\mathbf{x}}^2 = \mathbf{S}_{y_jx} \mathbf{S}_{xx}^{-1} \mathbf{S}_{xy_j} S_{y_jy_j}^{-1}$$

where  $\mathbf{S}_{y_jx}$  is the  $j$ th row of  $\mathbf{S}_{yx}$ ,  $\mathbf{S}_{xy_j}$  is the  $j$ th column of  $\mathbf{S}_{xy}$ , and  $S_{y_jy_j}$  is the  $j$ th element of the diagonal of  $\mathbf{S}_{yy}$ . By comparing these multiple correlation values, we can determine which of the response variables is most correlated with the set of predictor variables. Similarly, we can compute the individual correlation of each  $x_k$  and  $\mathbf{y}$  as the square root of

$$R_{x_k|\mathbf{y}}^2 = \mathbf{S}_{x_ky} \mathbf{S}_{yy}^{-1} \mathbf{S}_{yx_k} S_{x_kx_k}^{-1}$$

where  $\mathbf{S}_{x_ky}$  is the  $k$ th row of  $\mathbf{S}_{xy}$ ,  $\mathbf{S}_{yx_k}$  is the  $k$ th column of  $\mathbf{S}_{yx}$ , and  $S_{x_kx_k}$  is the  $k$ th element of the diagonal of  $\mathbf{S}_{kk}$ .

### A.5.1 Multiple correlation of wetland health

We estimate the relationship between the latent continuous response,  $\mathbf{Z} = \{\mathbf{Z}_1, \dots, \mathbf{Z}_J\}$ , and the latent spatial process,  $\mathbf{H}$ , in our model by estimating multiple correlation. Using the notation given above,  $p = J$  and  $q = 1$ . Due to the deterministic relationship between  $\mathbf{Z}$  and  $\mathbf{Y}$ , we assume that the relationship we are estimating will capture that of the relationship between  $\mathbf{H}$  and the multivariate response  $\mathbf{Y}$ . We partition the covariance matrix of the matrix  $\mathbf{Z}$  and vector  $\mathbf{H}$  according to (60).

The multiple correlation statistic is computed for each metric using the posterior simulations described in Section 5.3.4. The posterior median sample covariances for the discrete-only response model are

$$\mathbf{S} = \left( \begin{array}{c|c} \mathbf{S}_{ZZ} & \mathbf{S}_{ZH} \\ \hline \mathbf{S}_{HZ} & \mathbf{S}_{HH} \end{array} \right) = \left( \begin{array}{ccccc|c} 2.72 & 2.38 & 4.33 & 0.37 & 2.76 & 1.73 \\ 2.38 & 4.72 & 5.93 & 0.50 & 3.82 & 2.41 \\ 4.33 & 5.93 & 32.18 & 0.91 & 6.88 & 4.26 \\ 0.37 & 0.50 & 0.91 & 0.99 & 0.57 & 0.36 \\ 2.76 & 3.82 & 6.88 & 0.57 & 4.81 & 2.77 \\ \hline 1.73 & 2.41 & 4.26 & 0.36 & 2.77 & 1.74 \end{array} \right)$$

To measure the overall relationship between  $\mathbf{Z}$  and  $\mathbf{H}$ , we compute  $R_M^2 = 0.9507$ . The resulting canonical correlation is  $r_1 = 0.975$ . This indicates that there is a strong positive relationship between the multivariate response and the latent wetland condition variable. For  $j \in 1, \dots, J$ , we compute the correlation between the posterior draws of  $\mathbf{Z}_j$  and  $\mathbf{H}$  using (61). Larger values of  $R_{Z_j|H}$  (i.e., closer to 1) suggest that metric  $j$  is highly correlated with or explained by the underlying latent variable  $\mathbf{H}$ . The posterior median sample correlations are

$$\mathbf{R} = \left( \begin{array}{c|c} \mathbf{R}_{ZZ} & \mathbf{R}_{ZH} \\ \hline \mathbf{R}_{HZ} & \mathbf{R}_{HH} \end{array} \right) = \left( \begin{array}{ccccc|c} 1.00 & 0.67 & 0.46 & 0.23 & 0.77 & 0.80 \\ 0.67 & 1.00 & 0.49 & 0.24 & 0.81 & 0.84 \\ 0.46 & 0.49 & 1.00 & 0.16 & 0.56 & 0.58 \\ 0.23 & 0.24 & 0.16 & 1.00 & 0.27 & 0.28 \\ 0.77 & 0.81 & 0.56 & 0.27 & 1.00 & 0.96 \\ \hline 0.80 & 0.84 & 0.58 & 0.28 & 0.96 & 1.00 \end{array} \right)$$

The multiple correlation estimates, 95% credible intervals, estimates of percent contribution, and rank in evaluating wetland condition for the discrete-only model are given in Table 5.3.

We can compute the same posterior median sample covariances, sample correlations, and the estimates and credible intervals of the multiple correlation values for the mixed response model. The posterior median sample covariances are

$$\mathbf{S} = \left( \begin{array}{c|c} \mathbf{S}_{ZZ} & \mathbf{S}_{ZH} \\ \hline \mathbf{S}_{HZ} & \mathbf{S}_{HH} \end{array} \right) = \left( \begin{array}{ccccc|c} 1.59 & 0.72 & 3.11 & 0.22 & 0.91 & 0.59 \\ 0.72 & 2.30 & 3.88 & 0.28 & 1.12 & 0.72 \\ 3.11 & 3.88 & 30.31 & 1.15 & 4.90 & 3.14 \\ 0.22 & 0.28 & 1.15 & 0.70 & 0.35 & 0.23 \\ 0.91 & 1.12 & 4.90 & 0.35 & 1.62 & 0.93 \\ \hline 0.59 & 0.72 & 3.14 & 0.23 & 0.93 & 0.60 \end{array} \right)$$

and the posterior median sample correlations are

$$\mathbf{R} = \left( \begin{array}{c|c} \mathbf{R}_{ZZ} & \mathbf{R}_{ZH} \\ \hline \mathbf{R}_{HZ} & \mathbf{R}_{HH} \end{array} \right) = \left( \begin{array}{ccccc|c} 1.00 & 0.38 & 0.46 & 0.22 & 0.58 & 0.61 \\ 0.38 & 1.00 & 0.47 & 0.22 & 0.59 & 0.63 \\ 0.46 & 0.47 & 1.00 & 0.26 & 0.72 & 0.76 \\ 0.22 & 0.22 & 0.26 & 1.00 & 0.33 & 0.36 \\ 0.58 & 0.59 & 0.72 & 0.33 & 1.00 & 0.95 \\ \hline 0.61 & 0.63 & 0.76 & 0.35 & 0.95 & 1.00 \end{array} \right)$$

The multiple correlation estimates for the mixed response model are given in Table A.6. The results appear to be similar to those of the discrete-only response model given in Table 5.3. The differences in the rankings of the metrics, specifically metrics 1, 2, and 3 is the result of the continuous scale representation of floristic quality assessment being discretized differently for different wetland types. Whereas continuous variables tend to contain more information than their discretized counterpart, the discretization contained information relative to wetland health and led to changes in the correlation within the metric response variables.

### A.5.2 Ratio of Fixed Effects

Another way to quantify the relationship between  $\mathbf{Y}$  and  $\mathbf{H}$  is by computing the ratio of  $\omega_j$  to  $\theta_j$  for each of the metrics  $j = 1, \dots, J$  (see equation (57)). Since the parameter  $\theta_j$  is

Table A.6: Mixed response model: Estimates and 95% credible intervals of multiple correlation for each metric as well as the values of percent contribution and rank for each metric in evaluating wetland condition.

Metric	Parameter	Est.	95 % CI	% Contrib.	95 % CI	Rank	Index
Native plant cover	$R_{Z_1 H}$	0.61	(0.45, 0.76)	0.19	(0.15, 0.21)	4	20%
Noxious weed cover	$R_{Z_2 H}$	0.63	(0.42, 0.79)	0.19	(0.15, 0.22)	3	0 or 20%
Aggressive native cover	$R_{Z_3 H}$	0.77	(0.49, 0.93)	0.23	(0.17, 0.27)	2	0 or 20%
Structural diversity	$R_{Z_4 H}$	0.36	(0.18, 0.56)	0.11	(0.06, 0.15)	5	20%
Floristic quality	$R_{Z_5 H}$	0.95	(0.89, 0.97)	0.29	(0.25, 0.34)	1	40%

Table A.7: Estimates of the ratio  $\omega/\theta$  from the discrete-only model as well as % contribution of the metric and ranking.

Metric	Parameter	Estimate	% Contrib.	Rank	Index
Native plant cover	$\omega_1/\theta_1$	0.40	0.20	2	20%
Noxious weed cover	$\omega_2/\theta_2$	0.32	0.17	3	0 or 20%
Aggressive native cover	$\omega_3/\theta_3$	0.23	0.12	4	0 or 20%
Structural diversity	$\omega_4/\theta_4$	0.10	0.05	5	20%
Floristic quality	$\omega_5/\theta_5$	0.88	0.46	1	40%

a fixed effect for metric  $j$ , it represents the intercept in the relationship between  $\mathbf{Z}$  and  $\mathbf{H}$ . The parameter  $\omega_j$  is also a fixed effect and represents the factor loading of the spatial random field  $\mathbf{H}$ . Therefore, the ratio of  $\omega_j/\theta_j$  can be thought of as the ratio of the factor loading of the spatial random effect and the fixed effect for metric  $j$ . Larger values of  $\omega_j$  relative to  $\theta_j$  would indicate that  $\mathbf{H}$  is more closely related to or represented by metric  $j$ . For instance, when  $\omega_j$  is large, slight differences between  $\tilde{H}_i$  and  $\tilde{H}_l$  may result in  $\tilde{Y}_{ij} \neq \tilde{Y}_{lj}$ . When  $\omega_j$  is small, however, even significant differences in the latent health may not change the predicted ordinal response. The metrics can be ranked by comparing  $\omega_j/\theta_j$  for  $j \in 1, \dots, J$  (Table A.7). The results indicate that  $\mathbf{H}$  is predominantly represented by floristic quality. Native plant cover and noxious weed cover are also correlated with  $\mathbf{H}$ . The comparison of the ratio of fixed effects can be made only between metrics having response variables of the same type and scale.

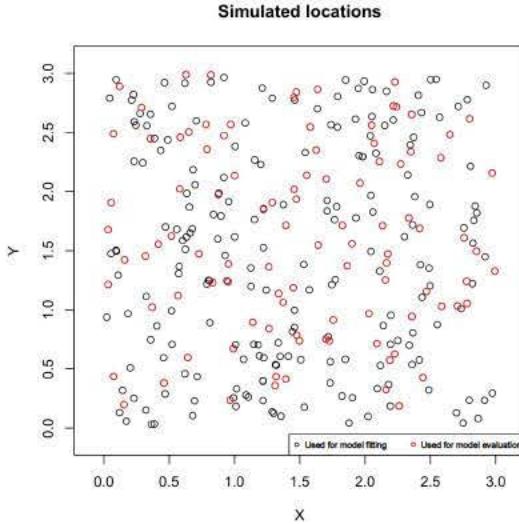


Figure A.5: Three hundred simulated locations within  $3 \times 3$  grid. Two hundred locations used to fit the model and the remaining one hundred were used for model evaluation.

## A.6 Simulation Study

To evaluate the performance of our methods we simulated data based on the multivariate multilevel latent variable model with three discrete response metrics. Three data sets were simulated as outlined below and the model was estimated for each data set. The outcomes of the three simulation were similar and therefore we report the results of only one. We define our spatial domain of interest as a  $3 \times 3$  spatial grid and simulated 300 locations uniformly over the region. We use the first  $n = 200$  locations to fit the proposed model and the remaining  $m = 100$  locations for prediction (Figure A.5).

We randomly simulated two multivariate random variables over the spatial domain to use as covariates  $\mathbf{X}_1$  and  $\mathbf{X}_2$  in the mean of  $\mathbf{H}$  where  $\mathbf{H}$  is drawn according to (58). Values for the coefficient vector  $\beta = (\beta_1, \beta_2)$  were fixed at 0.22 and 0.95, respectively. The true parameter of the covariance of  $\mathbf{H}$  was fixed at  $\phi_1 = 1$  and  $\phi_2 = 15.76$ . Note that  $\phi_1$  is the sill parameter of the covariance and  $\phi_2$  is the range parameter of the exponential correlation function. The effective range of the exponential correlation function is  $3/\phi_2 = 0.19$ . In

this simulation, the spatial correlation is only large at locations at close proximity. The true latent spatial random variable  $\mathbf{H}$  was a random drawn from a multivariate normal distribution with mean  $\mathbf{X}\boldsymbol{\beta}$  and covariance  $\Sigma_H(\boldsymbol{\phi}) = \phi_1 \exp^{-\mathbf{d}\boldsymbol{\phi}_2}$  where  $\mathbf{d}$  is the  $n \times n$  matrix of distances between sample locations. The true fixed effects  $\boldsymbol{\theta}$  and  $\boldsymbol{\omega}$  were randomly drawn from independent uniform and normal distributions, respectively where  $\boldsymbol{\theta} = \{1.73, 3.80, 3.62\}$  and  $\boldsymbol{\omega} = \{1.00, 1.37, -0.77\}$ . The true values of  $\theta_1$ ,  $\theta_2$ , and  $\theta_3$  were all chosen to be positive to ensure that the observed ordinal response values spanned each of the  $K = 5$  categories. For  $j = 1, 2, 3$ , the true continuous latent random vector  $\mathbf{Z}_j$  was drawn from its multivariate normal distribution with mean  $\theta_j \mathbf{1} + \omega_j \mathbf{H}$  and variance  $\sigma_j^2 \mathbf{I}_n$ . The length-6 vector of threshold values  $\boldsymbol{\lambda}$  was fixed such that  $\lambda_0 = -\infty$ ,  $\lambda_1 = 0$ , and  $\lambda_5 = \infty$ . The other thresholds were drawn from a multivariate normal distribution on the transformed scale and then were back-transformed. This was done to preserve the constraint that  $\lambda_k \leq \lambda_{k+1}$ . The resulting true threshold was set to

$$\boldsymbol{\lambda} = \{-\infty, 0, 1.81, 3.26, 4.71, \infty\}.$$

The observed ordinal response data  $\mathbf{Y}_j$  for metrics  $j = 1, \dots, 3$  are in the set  $\{1, \dots, 5\}$  based on the values of  $\mathbf{Z}_j$  and the true threshold vector  $\boldsymbol{\lambda}$ .

We ran the MCMC algorithm for 100,000 iterations and discarded the first 10,000 as burn-in. The true parameter values as well as the posterior median and 95% credible intervals are given in Table A.8. The results show that all but two parameters,  $\omega_2$  and  $\sigma_2^2$  are captured their respective credible interval.

Using the posterior draws of the model parameters, we make predictions using the Bayesian posterior prediction distributions  $p(\tilde{\mathbf{Y}}|\mathbf{y})$  and  $p(\tilde{\mathbf{H}}|\mathbf{y})$ . We evaluated the predictive ability of the model by comparing the mode of the posterior prediction distribution to the true metric value at each site for each metric (Table A.9). Of the 300 predicted metric scores, the truth was captured 57% of the time, whereas the predicted metric value was within 1 of the truth 93% of the time.

Table A.8: Simulated parameter values and posterior median estimates and 95% credible interval from model output.

Parameter	True Value	Estimate	95 % CI
$\beta_1$	0.22	0.21	(0.04, 0.39)
$\beta_2$	0.95	1.15	(0.90, 1.41)
$3/\phi_2$	0.19	0.31	(0.17, 0.66)
$\omega_1$	1.00	1.00	
$\omega_2$	1.37	0.73	(0.47, 1.12)
$\omega_3$	-0.77	-0.80	(-1.01, -0.61)
$\theta_1$	1.73	1.44	(0.73, 2.13)
$\theta_2$	3.80	4.16	(3.30, 4.93)
$\theta_3$	3.62	4.01	(3.28, 4.83)
$\sigma_1^2$	1.00	1.00	
$\sigma_2^2$	2.75	0.96	(0.48, 1.97)
$\sigma_3^2$	1.41	1.52	(1.03, 2.18)

Table A.9: The posterior modes and true discrete metric response values at the  $m = 100$  new locations for all metrics. In bold are the number of correct predictions of metric response values.

Posterior Median	True Value				
	1	2	3	4	5
1	<b>1</b>	1	0	0	0
2	10	<b>15</b>	8	3	0
3	5	29	<b>36</b>	25	3
4	0	2	4	<b>15</b>	10
5	1	2	5	22	<b>103</b>

Capturing the latent random field  $\mathbf{H}$  is of primary focus in this work. We make predictions of  $\widetilde{\mathbf{H}}$  at  $m = 100$  new locations by taking draws from the Bayesian posterior prediction distribution  $p(\widetilde{\mathbf{H}}|\mathbf{y})$ . Ninety-five percent posterior prediction intervals of  $\widetilde{\mathbf{H}}$  at  $m = 100$  new locations indicate that only one interval failed to capture the true value (Figure A.6). This indicates that our method achieves appropriate predictive coverage.

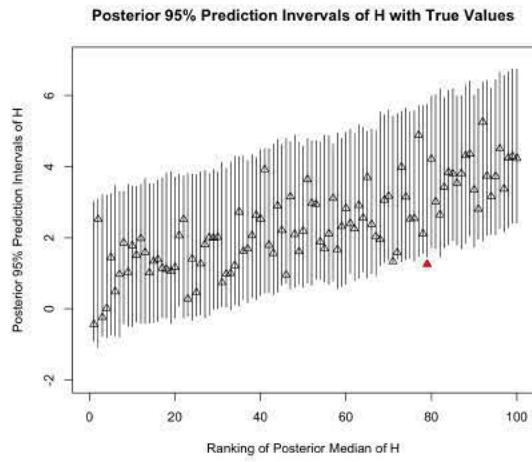


Figure A.6: Ninety-five percent posterior prediction intervals for latent spatial field  $\mathbf{H}$  at  $m = 100$  new locations. The true value is captured in 99 of the intervals.