# A BAYESIAN APPROACH TO SPATIAL CORRELATIONS IN THE MULTIVARIATE PROBIT MODEL

by

Jervyn Ang

B.Sc, Simon Fraser University, 2008

A Project submitted in partial fulfillment

of the requirements for the degree of

Master of Science

in the Department

of

Statistics and Actuarial Science



SIMON FRASER UNIVERSITY

Fall 2010

# APPROVAL

**Name:** Jervyn Ang

**Degree:** Master of Science

**Title of Project:** A Bayesian Approach to Spatial Correlations in the Multivariate Probit Model

**Examining Committee:** Dr. Richard Lockhart
Chair

_____

Dr. Derek Bingham, Senior Supervisor

_____

Dr. Tim Swartz, Internal Examiner

_____

Dr. Dave Campbell, External Examiner

**Date Approved:** _____

# Abstract

Ordered categorical data arise in many applied settings. For example, many surveys have responses that may be restricted to "Strongly Disagree", "Disagree, "Neutral", "Agree", and "Strongly Agree". Here, the responses are ordinal variables. That is, the agreeability of respondents to questions have relative ranks, but there is no measure of exact magnitude like there is with continuous variables.

In many scenarios, questions may have correlated responses. As well, different respondents may be spatially or otherwise correlated. Probit models are a means to using normal latent variables in modelling ordinal responses. In this project, we take a Bayesian approach and include both "between question" and "between respondent" correlations in a multivariate probit model. We discuss the efficacy of this spatial multivariate probit model.

# Acknowledgments

I would like to thank everyone who has supported me through my educational career. In particular, I would like to thank Dr. Derek Bingham for being an excellent supervisor, advisor, and friend. I would also like to thank Dr. Carl Schwarz for his dedication to teaching, and for showing me how interesting Statistics could be. Thanks also goes to Dr. Dave Campbell and Dr. Tim Swartz for all their help and advice on my work. Special thanks go to my parents, for their love, support, and understanding.

# Contents

**Approval** ii

**Abstract** iii

**Acknowledgments** iv

**Contents** v

**List of Tables** vii

**1 Introduction** 1

**2 Preliminaries and Notation** 4

2.1 Ordinal Data and Latent Variables . . . . . 4

2.2 Regression on Latent Variables . . . . . 5

2.3 Univariate Probit Model . . . . . 6

2.4 Notation and Definitions . . . . . 7

2.4.1 Vectorization . . . . . 7

2.4.2 Kronecker Product . . . . . 7

2.4.3 The Matrix Variate Normal Distribution . . . . . 8

2.5 Multivariate Probit Model . . . . . 9

2.5.1 Inference for the Multivariate Probit Model . . . . . 10

**3 Spatial Correlations, Fitting, and MCMC** **11**

3.1 Spatial Correlations . . . . . 11

3.2 Spatial Probit Models . . . . . 12

3.3 Clipped Gaussian Processes . . . . . 13

3.4 Spatial Multivariate Probit Models . . . . . 13

3.5 Bayesian Inference and Parameter Expansion . . . . . 15

3.5.1 Parameter Expansion in the SMVP Model . . . . . 15

3.6 Priors and Full Conditional Posterior Distributions . . . . . 18

3.7 Markov Chain Monte Carlo Algorithm . . . . . 20

**4 Parameter Estimation and Variable Selection** **24**

4.1 Study Design and Model Fitting . . . . . 24

4.2 Parameter Estimation Results . . . . . 29

4.2.1 Results . . . . . 29

4.3 Variable Selection . . . . . 35

4.3.1 Results . . . . . 35

4.4 Discussion . . . . . 36

**5 Conclusion and Future Work** **38**

**Bibliography** **40**

# List of Tables


4.1 Mean of point estimates (and RMSE) of $R_{12} = 0.5$ . . . . . 29
4.2 Mean of point estimates (and RMSE) of $R_{13} = 0.25$ . . . . . 30
4.3 Mean of point estimates (and RMSE) of $R_{23} = 0.1$ . . . . . 30
4.4 Mean of point estimates (and RMSE) of $\beta_{11} = 2$ . . . . . 30
4.5 Mean of point estimates (and RMSE) of $\beta_{21} = 0$ . . . . . 31
4.6 Mean of point estimates (and RMSE) of $\beta_{31} = 1$ . . . . . 31
4.7 Mean of point estimates (and RMSE) of $\beta_{12} = -3$ . . . . . 31
4.8 Mean of point estimates (and RMSE) of $\beta_{22} = 1$ . . . . . 32
4.9 Mean of point estimates (and RMSE) of $\beta_{32} = 0$ . . . . . 32
4.10 Mean of point estimates (and RMSE) of $\gamma_{21} = 1$ . . . . . 33
4.11 Mean of point estimates (and RMSE) of $\gamma_{22} = 1$ . . . . . 33
4.12 Mean of point estimates (and RMSE) of $\gamma_{23} = 1$ . . . . . 33
4.13 Mean of point estimates (and RMSE) of $\rho$ . . . . . 34
4.14 Proportion of times 0 was in the 95% credible interval for $\beta_{12} = -3$ . 35
4.15 Proportion of times 0 was in the 95% credible interval for $\beta_{22} = 1$ . . 36
4.16 Proportion of times 0 was in the 95% credible interval for $\beta_{32} = 0$ . . 36

# Chapter 1

# Introduction

Ordinal data often arise when using various surveys, polls, and other similar methods to obtain opinions and attitudes toward certain issues. In particular, some methods attempt to gauge the strength of agreeability (or disagreeability) of respondents to statements or questions. One commonly used scale for this purpose is the Likert scale (Likert 1932), where respondents may be presented with a statement or question and possible answers of "Strongly Disagree", "Disagree", "Neutral", "Agree", and "Strongly Agree". It is common to assign numerical scores of 1 (Strongly Disagree) to 5 (Strongly agree) when dealing with such data.

As an example, consider responses the following two statements from a teaching evaluation survey, where possible responses are restricted to those listed above.

- Your course was valuable and informative.
- Your instructor was effective and helpful.

Data obtained from a survey of this form would be ordered categorical data. One temptation would be to treat data of this form as continuous and analyse them as such. However, there are certain fundamental flaws with this technique. For one, the difference between "Neutral" and "Disagree" might be quite dissimilar to the

difference between "Strongly Agree" and "Agree". That is, we have knowledge of the rankings of these responses, but we do not know the exact magnitude of agreeability.

Several approaches to this problem have been proposed and examined. Among them are McCullagh (1980), where the author proposes a latent variable model that uses an explicit measure of distance between each ordinal category. In this approach, the unobserved variable space is partitioned using unknown cutpoints. This approach yields maximum likelihood estimators for various parameters of interest as well as the nuisance cutpoint parameters. Albert and Chib (1993) develop much of the current Bayesian framework for ordered categorical data analysis using univariate probit models. Latent variables in their approach are treated as missing, and data augmentation techniques are used to perform necessary inference.

Chib and Greenberg (1998) provide a similar Bayesian approach to the multivariate probit model. In their case, they propose a method that estimates correlations between questions using Markov chain Monte Carlo (MCMC) techniques. Lawrence et al. (2008) propose a more efficient method of Bayesian inference by using parameter expansion, an algorithm that was first proposed in Liu and Wu (1999) to speed up the convergence of MCMC algorithms.

De Oliveira (2000) develop a method of incorporating spatial dependence in the binary regression model. In this spatial probit model, correlations between respondents were incorporated into a model for binary data using indicator Kriging. Higgs and Hoeting (2010) extend this method to ordered categorical data with more than two categories. In the two papers, predictive algorithms were also implemented and discussed.

In this project, we propose a method to incorporate correlations between questions and correlations between respondents into the probit regression model. This work combines the work of Lawrence et al. (2008), where efficient modelling of "between question" correlations is done, with Higgs and Hoeting (2010), where "between respondent" correlations are modelled using Bayesian spatial methods. In short, we

develop the spatial multivariate probit (SMVP) model, and use Bayesian methods for inference.

# Chapter 2

# Preliminaries and Notation

In this chapter, several models for ordinal responses with various correlation structures are discussed. In particular, the univariate probit model and multivariate probit models are examined.

## 2.1 Ordinal Data and Latent Variables

Ordinal data can arise in many situations. For example, surveys often have questions where responses are limited to "Strongly Disagree", "Disagree", "Neither Disagree Nor Agree", "Agree", and "Strongly Agree". The Likert scale is often used to assign numerical values to the possible responses. In this scenario, the numbers one (Strongly Disagree) to five (Strongly Agree) could be assigned.

One approach to modelling ordinal responses is the use of latent variables. These unobserved random variables can be useful in a mathematical sense, and can be interpreted as follows. In a case of possible responses being restricted to the five options above, the latent variable associated with this response can be thought of as an "amount of agreeability". That is, the higher the value of the latent variable, the more agreeable the respondent is to the statement or question. Latent variables are

an important part of the probit models that we will be discussing.

As an example, students in a class can be asked to answer a survey regarding their course, instructor, and other related issues. Consider again these two statements, with possible responses restricted to those listed above.

- Your course was valuable and informative.
- Your instructor was effective and helpful.

Here, each student is a respondent with a certain attitude toward the statements. Less agreeable students will respond with "Strongly Disagree", slightly less agreeable students will respond with "Disagree", and so on. Here, the "amount of agreeability" is a latent variable where certain values of this variable will be associated with different ordinal values.

In this setting, attitudes towards the first and second questions could be dependent. That is, a student's attitude toward the course might be highly correlated with his attitude toward the instructor. We call this the "between question" correlation. As well, it seems likely that students are not all independent. For example, students that are friendly with each other may have correlated attitudes. We call this the "between respondent" correlation.

## 2.2 Regression on Latent Variables

It is quite likely that certain variables can impact the way respondents answer surveys. One way to account for this is simply to use a regression model on our latent variables. That is, the means of the latent variables are fit based on certain predictors. In our example, the students attitudes may be affected by age, GPA, years of study, etc. For example, one might expect the average 25-year-old student to have a different attitude from the average 20-year-old student.

In traditional regression methods, a model may look something like $Y \sim N(\beta X, \sigma^2)$, where $Y$ is a response and $X$ is a predictor. In the context of latent variable models for ordered categorical data, the latent variable $Z$, not the actual response $Y$, is modelled. That is, one may see something like $Z \sim N(\beta X, \sigma^2)$, where the actual ordered categorical value of each response is based on Z and associated cutpoints.

## 2.3 Univariate Probit Model

Consider possible responses to a single survey question with a 1-5 Likert scale. The possible ordinal responses to this question, $Y$, are said to have associated probabilities $P(Y = k)$, for $k = 1, \ldots, 5$. One method to model these probabilities is the univariate probit model, where a Gaussian latent variable, $Z$, is regressed on predictors. The probabilities $P(Y = k)$, for $k = 1, \ldots, 5$ are determined by the probability that $Z$ lies within certain ranges determined by cutpoints (Johnson and Albert 1999). That is, $Z \sim \mathrm{N}(X\beta, \sigma^2)$, and $P(Y_1 = k) = \int_{\gamma_{k-1}}^{\gamma_k} (2\pi)^{-\frac{1}{2}} (\sigma)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}(Z - \beta X)^2\right\} dZ$.

If the intercept term in $\beta$ and all the cutpoints $\gamma_k$ are estimated, identifiability issues arise. To see this, note that if a constant were added to the intercept of the regression model and to all the cutpoints, the resulting probabilities would be identical. The conventional means to solving this identifiability problem is to set the first cutpoint, $\gamma_1$, to be 0. As well, note that multiplying our variance parameter $\sigma^2$ and all cutpoints by a constant would also yield identical probabilities. This creates another identifiability problem, which can be solved by setting our variance to 1 (see Lawrence et al. 2008, Higgs and Hoeting 2010).

In short, the Gaussian latent variable, $Z$, is regressed on predictors. Associated cutpoint parameters on the latent space are used to determine probabilities of observing each response, the first of these cutpoints being set to zero for identifiability reasons. As well, the variance of the latent variable is set to be 1.

As stated, $Y$ can take on one of $k$ ordered values, with probabilities specified by

a Gaussian random variable. This model is appropriate in the case where questions and respondents in a survey are assumed to be independent. However, in many applications, the assumption of independence may not be valid. When there are several questions, the responses may be dependent. One extension that accounts for this dependency is the multivariate probit model, where correlations among questions are included. In addition, similar subjects may respond in similar ways. To this end, we will propose the spatial multivariate probit model, where spatial correlation is combined with the multivariate probit model. To do so, we must first define some convenient notation.

## 2.4 Notation and Definitions

### 2.4.1 Vectorization

Let vec($A$) denote the vectorization of the matrix $A$ such that the columns of $A$ are stacked to form a vector. If

$$A = \begin{pmatrix} a & b & c \\ d & e & f \end{pmatrix},$$

then

$$\text{vec}(A) = \begin{pmatrix} a \\ d \\ b \\ e \\ c \\ f \end{pmatrix}.$$

### 2.4.2 Kronecker Product

A Kronecker product, denoted $\otimes$ is a matrix multiplication operation. For a $q \times p$ matrix $A$ and an $n \times m$ matrix $B$, $A \otimes B$ yields a $qn \times pm$ matrix as follows (Petersen

and Pedersen 2008):

$$A \otimes B = \begin{bmatrix} a_{11}B & \cdots & a_{1p}B \\ \vdots & \ddots & \vdots \\ a_{q1}B & \cdots & a_{qp}B \end{bmatrix}.$$

### 2.4.3 The Matrix Variate Normal Distribution

A matrix $A$ is said to follow a matrix variate normal distribution with mean matrix $M$ and covariance matrix $B \otimes C$ if and only if $\text{vec}(A')$ follows multivariate normal distribution with mean vector $\text{vec}(M')$ and covariance matrix $B \otimes C$ (Gupta and Nagar 2000). That is,

$$A \sim \text{MatVN}(M, B \otimes C)$$

if and only if

$$\text{vec}(A') \sim \text{MVN}(\text{vec}(M'), B \otimes C),$$

where MatVN denotes the matrix variate normal distribution and MVN denotes the multivariate normal distributiom. Consider a random $q \times n$ matrix $A$ with $q \times n$ mean matrix $M$ and covariance matrix $B \otimes C$, where $B$ is $q \times q$ and $C$ is $n \times n$. The associated matrix variate normal density, $f(A)$, is stated as follows:

$$f(A) = (2\pi)^{-\frac{qn}{2}} |B \otimes C|^{-\frac{1}{2}} \text{etr}\left\{ -\frac{1}{2} B^{-1}(A - M)C^{-1}(A - M)' \right\},$$

where $\text{etr}(x) = \exp(\text{trace}(x))$.

A useful interpretation of the individual matrices $B$ and $C$ is as follows. $B$ represents the "between row" correlation, and C represents the "between column" correlation. Matrix variate normal random variables exhibit a multiplicative correlation structure. In this case, $\text{Cor}(A_{ij}, A_{lk}) = B_{il}C_{jk}$.

## 2.5 Multivariate Probit Model

Multivariate probit models (MVP) are a generalization of univariate probit models. Instead of responses to a single question, consider a survey of $q$ correlated questions and $n$ independent respondents. Let $Y$ be the $q \times n$ matrix of responses to the survey. Let each column, $Y_j$, correspond to data from a single respondent, and let each row correspond to responses to each question. That is, $Y_{ij}$ is the response of person $j$ to question $i$. So,

$$Y = \left[ \begin{array}{cc} Y_1 & Y_2 \cdots Y_n \end{array} \right].$$

We can denote the $j^{th}$ observation and associated predictors by $\{Y_j, X_j\}$, for $j = 1, \ldots, n$. Here, each $Y_{i,j}$ (where i denotes the question index) takes on one of $k_i$ ordered values. That is, the $i^{th}$ question has $k_i$ possible responses. In the case where each question has the same number of possible responses, $k_i$=k. Likewise, $X$ is a $p \times n$ matrix of predictor values, where each row corresponds to one predictor and each column corresponds to one respondent.

Let $Z_j$ be the corresponding $q$-variate normal latent variable for the $j^{th}$ respondent. In the univariate probit model, the latent variable has associated cutpoints that correspond with different values of $Y$ (see Johnson and Albert 1999). Extending this to the multivariate case, each $Y_j$ can be obtained from the latent variable $Z_j$ as follows. Let $\Gamma$ be a set of cutpoints, with each question having $k + 1$ associated cutpoints. For the purposes of identifiability, $\gamma_{i,0} = -\infty$ and $\gamma_{i,k_i} = \infty$ for all $i$. So, if $\gamma_{i,c-1} < Z_{i,j} \leq \gamma_{i,c}$, $Y_{i,j} = c$ is observed. The vector of means of $Z_j$ is determined by the predictors and regression parameters. Specifically, let $\beta_i$ be a $1 \times p$ vector of regression coefficients associated with the $i^{th}$ question. The mean for $Z_{i,j}$ is given by $\beta_i X_j$. If the matrix of regression coefficients is denoted by $\beta = (\beta_1', \beta_2', \ldots, \beta_q')'$, then $Z_j$ is distributed normally with mean $\beta X_j$.

For the model to be identifiable, certain parameters have to be restricted. Similar to the univariate probit model, the first cutpoint for each question is fixed at 0. As

well, the variances are restricted to one, which means that the covariance matrix of $Z$ is restricted to a correlation matrix, denoted $R$. In this scenario, a correlation between questions can be modelled. The off-diagonal entries in the $R$ matrix would be the correlations between questions, such that $R_{ij}$ is the correlation between $Z_{li}$ and $Z_{lj}$. As such, the correlation matrix $R$ contains the "between row" correlations of $Z$. Assuming independence between respondents leads to independence between columns. This yields a convenient expression for the correlation matrix associated with $Z$, $R \otimes I$.

### 2.5.1 Inference for the Multivariate Probit Model

Recall from section 2.4.3 that:
If $\text{vec}(A') \sim \text{MVN}(\text{vec}(M'), B \otimes C)$, then $A \sim \text{MatVN}(\text{vec}(M'), B \otimes C)$.
So, $Z \sim \text{MatVN}(\beta X, R \otimes I)$ since $\text{vec}(Z') \sim \text{MVN}(\text{vec}((\beta X)'), R \otimes I)$.

As such, using the multivariate probit model, the probability of observing Y can be stated in the following manner:

$$\text{P}\{Y_1 = y_1, \ldots, Y_q = y_q\} = \int_{\gamma_{q,y_q-1}}^{\gamma_{q,y_q}} \ldots \int_{\gamma_{1,y_1-1}}^{\gamma_{1,y_1}}$$
$$(2\pi)^{-\frac{qn}{2}} |R \otimes I|^{-\frac{1}{2}} \text{etr}\left\{-\frac{1}{2} R^{-1}(Z - \beta X) I (Z - \beta X)'\right\} dZ_1 \ldots dZ_q,$$

where $Y_i$ is a response to the $i^{th}$ question.

This model is very useful when dealing with survey data with dependent questions and independent respondents; however, it may be necessary in practice to consider the case of spatially dependent respondents. The matrix $I$ in this framework indicates respondents are independent. The case we are specifically interested in has spatial dependence. In the next chapter, we adapt the multivariate probit model to account for this spatial dependence.

# Chapter 3

# Spatial Correlations, Fitting, and MCMC

This chapter deals with modelling multivariate ordered categorical data in the presence of spatial dependence. Gaussian processes are used to include spatial dependence in the multivariate probit model. The aim of this approach is to incorporate both spatial correlations between respondents and correlations between questions.

## 3.1 Spatial Correlations

In addition to regular predictor variables, spatial locations often arise in applied settings. While the method outlined in Lawrence et al. (2008) works well for surveys involving independent respondents, it does not have the direct ability to account for spatially or otherwise dependent respondents. We propose an approach that accounts for this spatial dependency.

Consider again a survey of $q$ questions and $n$ respondents. $Y_{ij}$ is the response of person $j$ to question $i$, and $Z_{ij}$ is the associated latent variable. $\text{Cor}(Z_{i.}, Z_{k.})$ is the correlation between the responses to $i^{th}$ and $k^{th}$ question. Likewise, $\text{Cor}(Z_{.j}, Z_{.l})$ is

the correlation between the $j^{th}$ and $l^{th}$ respondent.

The correlation between questions can be accounted for using the correlation matrix $R$ in the multivariate probit model. $\text{Cor}(Z_{.j}, Z_{.l})$, the spatial correlation, can be modelled using one of many spatial models, such as a Gaussian process model.

**Gaussian Process Models**

One often-used model in spatial statistics is the Gaussian process model (Cressie 1993). Similar to other spatial models, close neighbours are thought of as having a higher correlation than respondents who are further apart. We take a similar view here and model that spatial information using a stationary, isotropic covariance model. If we let $\Psi$ be the matrix of spatial correlations, where $\Psi_{ij} = \text{Cor}(Z_{.i}, Z_{.j})$, this spatial correlation structure can be used to account for the correlation between respondents in our probit model.

## 3.2 Spatial Probit Models

Spatial probit models (SPM) are an extension of univariate probit models, except that respondents are spatially correlated. We have considered the matrix $R$ of between question correlations in some models. Now consider an $n \times n$ matrix $\Psi$ of spatial correlations to account for "between respondent" correlations. In our presentation of the multivariate probit model, the $q \times n$ matrix of $Y$ (or the $qn \times 1$ vector, $\text{vec}(Y')$) responses previously had an associated correlation matrix of $R \otimes I$. Here, the matrix $R$ is a $q \times q$ correlation matrix, where $R_{i,j}$ is the correlation between questions $i$ and $j$. Spatial probit models involve a "between respondent" correlation matrix, denoted $\Psi$, where $\Psi_{lk}$ is the correlation between respondent $l$ and $k$. While other authors have not considered the multivariate probit model with spatial correlation, we will be combining the spatial probit model and multivariate probit model to incorporate both kinds of correlation.

We begin by presenting the univariate probit model with spatial correlations (see Higgs and Hoeting 2010 and De Oliveira 2000), and then propose the SMVP model for multivariate ordinal responses with spatial correlation. Recall in the UVP model, $Z \sim \mathrm{N}(\beta X, 1)$. In the spatial probit model, $Z \sim \mathrm{MVN}(\beta X, \Psi)$.

## 3.3 Clipped Gaussian Processes

Spatial probit models have received attention recently in the literature. De Oliveira (2000) proposed the clipped Gaussian process (CGP) model for an application involving correlated binary data. In Higgs and Hoeting (2010), the clipped Gaussian process model was suggested for a similar problem involving ordinal data with two or more categories.

The idea behind CGP for ordinal data is that a Gaussian process could be fit to the latent variables. The "clipping" is the process of observing certain ordinal values based on latent variables being within two cutpoints. The proposed method in Higgs and Hoeting (2010) for CGP extends the model in De Oliveira (1998) to account for more than two categories. In the next section, we propose the spatial multivariate probit model, where both aforementioned types of correlation are incorporated.

## 3.4 Spatial Multivariate Probit Models

In this section, the spatial multivariate probit model is proposed. The proposed approach is a latent variable model for multivariate ordinal data where both "between question" and "between respondent" correlations are incorporated. In this model, between respondent information is incorporated via spatial methods.

A key assumption made in our approach is the multiplicative correlation function, as seen in McMillian et al. (1999) and Qian et al. (2008). That is, $\mathrm{Cor}(Z_{i,j}, Z_{k,l}) = \mathrm{Cor}(Z_{i.}, Z_{k.})\mathrm{Cor}(Z_{.j}, Z_{.l})$. Using this, we can state the correlation matrix of $Z$, or more

specifically, vec($Z'$), as a Kronecker product, $R \otimes \Psi$. This representation becomes particularly useful as we try to model both kinds of correlations. We now formally develop such a model.

Let $S$ be a $d \times n$ matrix of spatial locations. Here, the $j^{th}$ column of $S$ corresponds to the $j^{th}$ column of $Y$, and $d$ is the number of spatial dimensions. Let $\Psi$ be the matrix of spatial correlations obtained using spatial locations, and let $\rho$ be the corresponding spatial correlation parameter. Similar to Higgs and Hoeting (2010), we use the exponential correlation Gaussian process model for our spatial correlation structure. That is, Cor( $Z_{.i}, Z_{.j}$) $= \Psi_{ij} = \rho^{-dist(S_i,S_j)}$, where $\rho$ is non-negative.

So, if we assume the multiplicative correlation structure such that $\text{Cor}(Z_{ij}, Z_{lk}) = R_{il}\Psi_{jk}$, we can then state the correlation matrix of vec($Z'$) as $R \otimes \Psi$. If the latent variables follow a matrix variate normal distribution, the observed data model would be as follows:

$$\text{P}\{Y_1 = y_1, \ldots, Y_q = y_q\} = \int_{\gamma_{q,y_q-1}}^{\gamma_{q,y_q}} \ldots \int_{\gamma_{1,y_1-1}}^{\gamma_{1,y_1}} (2\pi)^{-\frac{qn}{2}} |R \otimes \Psi|^{-\frac{1}{2}}$$
$$\text{etr}\left\{-\frac{1}{2}R^{-1}(Z - \beta X)\Psi^{-1}(Z - \beta X)'\right\} dZ_1 \ldots dZ_q,$$

where $Y_i$ is the response to the $i^{th}$ question. Here, $Z$ is latent variable that corresponds to $Y$, $R$ is the correlation between questions, and $\Psi$ is the correlation between respondents. Our goal is to perform Bayesian inference using this SMVP model and implement an MCMC algorithm. However, Gibbs sampling of the correlation matrix $R$ is problematic and drawing each individual term in $R$ will lead to long, slowly mixing Markov chains (Lawrence et al. 2008). To get around this issue, we will use the parameter expansion technique, as seen in Liu et al. (1998).

## 3.5 Bayesian Inference and Parameter Expansion

One issue concerning Bayesian inference in this setting is the effective sampling or drawing of valid correlation matrices. While covariance matrices can be easily sampled using the Inverse-Wishart distribution, the restriction of unit diagonal elements creates a problem of slow or inefficient sampling. Chib and Greenberg (1998) use a "one element at a time" approach for sampling individual correlation parameters. To increase efficiency, Lawrence et al. (2008) propose a method to sample an unidentifiable covariance matrix using Gibbs steps and then integrating out the scaling parameters to yield an identifiable correlation matrix. This approach is a special case of parameter expansion as seen in Liu et al. (1998) and Liu and Wu (1999).

Parameter expansion was first created to speed up convergence of MCMC algorithms. However, the technique was adapted in Lawrence et al. (2008) for easy and efficient sampling of correlation matrices in the context of MVP models. In this version of the algorithm, the parameter space is expanded by allowing for latent variables to have non-unit variances. This is done to allow for a Gibbs step as opposed to a Metropolis step. Note that ordinal data are invariant to affine transformations on the underlying latent variables and associated cutpoints. That is, affine transformations on latent values have no impact on the actual observed ordinal data or associated probabilities. So, after drawing from the expanded parameter space, latent values are re-scaled so that a correlation matrix is obtained, thus mapping us back onto the unexpanded parameter space. We adapt the technique as seen in Lawrence et al. (2008) by including the spatial correlation matrix, $\Psi$.

### 3.5.1 Parameter Expansion in the SMVP Model

Recall that $Z$ is the $q \times n$ matrix of latent values. To use the parameter expansion technique, we will be apply a scale transform on the values of $Z$. Let the matrix $V$ be a $q \times q$ diagonal matrix with elements $v_1, \ldots, v_q$, where all elements are positive.

Consider the following transformation on the latent variable $Z$: $\text{vec}(W') = (V^{\frac{1}{2}} \otimes I)\text{vec}(Z')$, where $I$ is the $n \times n$ identity matrix. Note that since $\text{vec}(Z')$ is multivariate normal, $\text{vec}(W') = (V^{\frac{1}{2}} \otimes I)\text{vec}(Z')$ must also be multivariate normal (see Petersen and Pedersen, 2008). So, we derive the expectation and covariance matrix of $\text{vec}(W')$ as follows:

$$\begin{aligned} E(\text{vec}(W')) &= (V^{1/2} \otimes I)\text{vec}((\beta X)') \\ &= \text{vec}(I(\beta X)'V^{1/2}) \\ &= \text{vec}(X'\beta'V^{1/2}) \\ &= \text{vec}(X'(V^{1/2}\beta)'). \end{aligned}$$

$$\begin{aligned} \text{Var}(\text{vec}(W')) &= (V^{1/2} \otimes I)(R \otimes \Psi)(V^{1/2} \otimes I) \\ &= ((V^{1/2}R) \otimes (I\Psi))(V^{1/2} \otimes I) \\ &= ((V^{1/2}RV^{1/2}) \otimes (I\Psi I)) \\ &= (V^{1/2}RV^{1/2}) \otimes \Psi. \end{aligned}$$

Thus, $\text{vec}(W') \sim \text{MVN}(\text{vec}(X'(V^{1/2}\beta)'), (V^{1/2}RV^{1/2}) \otimes \Psi)$. Using the results from section 2.4.3, we see that $W \sim \text{MatVN}(V^{1/2}\beta X, (V^{1/2}RV^{1/2}) \otimes \Psi)$. This transformation and subsequent substitution yields the following probability for $Y$:

$$\text{P}\{Y_1 = y_1, \ldots, Y_q = y_q\} = \int_{\sqrt{v_q}\gamma_{q,y_q-1}}^{\sqrt{v_q}\gamma_{q,y_q}} \cdots \int_{\sqrt{v_1}\gamma_{1,y_1-1}}^{\sqrt{v_1}\gamma_{1,y_1}} (2\pi)^{-\frac{q}{2}}|V|^{-\frac{1}{2}}|R \otimes \Psi|^{-\frac{1}{2}}$$
$$\text{etr}\left\{-\frac{1}{2}(V^{\frac{1}{2}}RV^{\frac{1}{2}})^{-1}(W - V^{\frac{1}{2}}\beta X)\Psi^{-1}(W - V^{\frac{1}{2}}\beta X)'\right\} dW_1 \ldots dW_q,$$

where $Y_i$ is a response to the $i^{th}$ question. This linear transformation on $Z$ is useful due to invariance of the probabilities for $Y$ despite the expansion of the parameter

space. We seek a representation of the probabilities for $Y$ based on the matrix variate normal with unrestricted variances. We get exactly that if we rewrite the parameters in the following way:

- $\alpha = V^{\frac{1}{2}}\beta$
- $\theta_{j,c} = \sqrt{v_j}\gamma_{j,c}$
- $\Sigma = V^{\frac{1}{2}}RV^{\frac{1}{2}}$.

That is, we now have a very convenient form for the probability of Y.

$$\mathrm{P}\{Y_1 = y_1, \ldots, Y_q = y_q\} = \int_{\theta_{q,y_q-1}}^{\theta_{q,y_q}} \ldots \int_{\theta_{1,y_1-1}}^{\theta_{1,y_1}} (2\pi)^{-\frac{q}{2}} |\Sigma \otimes \Psi|^{-\frac{1}{2}}$$
$$\mathrm{etr}\left\{-\frac{1}{2}\Sigma^{-1}(W - \alpha X)\Psi^{-1}(W - \alpha X)'\right\} dW_1 \ldots dW_q.$$

As noted before, the parameter set has been expanded and now includes variance parameters which were fixed at one. As such, this model is called the expanded parameter model. This model has a likelihood as follows:

$$\mathcal{L}(\alpha, \Sigma, \theta) = |\Sigma|^{-\frac{n}{2}} |\Psi|^{-\frac{q}{2}} \mathrm{etr}\left\{-\frac{1}{2}\Sigma^{-1}(W - \alpha X)\Psi^{-1}(W - \alpha X)'\right\}$$
$$\prod_{i=1}^{n}\prod_{j=1}^{q} \mathcal{I}\{\theta_{j,Y_{i,j}-1} < W_{i,j} \leq \theta_{j,Y_{i,j}}\}.$$

In short, $W_{k+1}|\alpha_k, \theta_k, \rho_k, \Sigma_k \sim \mathrm{MatVN}(\alpha_k X, \Sigma_k \otimes \Psi_k)$. We use this expanded parameter likelihood for our MCMC sampling algorithm. To implement the MCMC algorithm, we first state our prior distribution and derive our posterior distribution for each parameter.

## 3.6 Priors and Full Conditional Posterior Distributions

Each parameter has a prior distribution and an associated full conditional posterior distribution. Where applicable, the priors used for our parameters are the same as those used in Lawrence et al. (2008). The authors elected to use reference priors on all parameters, and we take the same approach. The only significant difference is the addition of the prior on $\rho$, where we elect to use a beta(2,2) prior for a bounded but relatively uninformative prior on the spatial correlation parameter. In reality, more informative priors could be used based on available information. We now derive the full conditional posterior distributions for each parameter.

### Between Question Covariance

$\Sigma$ is the "between question" covariance matrix in the expanded parameter space. Let $P(\Sigma) \propto |\Sigma|^{-\frac{q+1}{2}}$, where $P(\Sigma)$ is the prior distribution of $\Sigma$. Let $P(\Sigma|...)$ be the full conditional distribution of $\Sigma$. The full conditional distribution of $\Sigma$ is derived as follows:

$$\begin{aligned} P(\Sigma|...) &\propto \mathcal{L}(\alpha, \Sigma, \theta, \Psi) * P(\Sigma) \\ &\propto |\Sigma|^{-\frac{n+q+1}{2}} \text{etr}\left\{-\frac{1}{2}\Sigma^{-1}(W-\alpha X)\Psi^{-1}(W-\alpha X)'\right\}. \end{aligned}$$

So, given all other parameters, $\Sigma \sim$ Inverse-Wishart$(n+q+1, (W-\alpha X)\Psi^{-1}(W-\alpha X)')$. As such, we have a method for Gibbs sampling of the covariance matrix on the expanded parameter space. This will prove highly useful when rescaling is done to map the parameters back to the original parameter space.

### Regression Parameters

The matrix of regression parameters in the expanded parameter space is denoted $\alpha$. Let $P(\alpha) \propto 1$, $M_\alpha = W\Psi^{-1}X'(X\Psi^{-1}X')^{-1}$, and $H = X\Psi^{-1}X'$. The full conditional

distribution of $\alpha$ is derived as follows:

$$\begin{aligned}
P(\alpha|...) &\propto \mathcal{L}(\alpha, \Sigma, \theta, \Psi) * P(\alpha) \\
&\propto \text{etr}\left\{-\frac{1}{2}\Sigma^{-1}(W-\alpha X)\Psi^{-1}(W-\alpha X)'\right\} \\
&\propto \text{etr}\left\{-\frac{1}{2}\Sigma^{-1}(-2\alpha X\Psi^{-1}W + \alpha X\Psi^{-1}X'\alpha'\right\} \\
&\propto \text{etr}\left\{-\frac{1}{2}\Sigma^{-1}(-2\alpha(X\Psi^{-1}X')(X\Psi^{-1}X')^{-1}X\Psi^{-1}W + \alpha(X\Psi^{-1}X')\alpha'\right\} \\
&\propto \text{etr}\left\{-\frac{1}{2}\Sigma^{-1}(M_\alpha H M_\alpha' - 2\alpha H M_\alpha + \alpha H\alpha')\right\} \\
&\propto \text{etr}\left\{-\frac{1}{2}\Sigma^{-1}((\alpha - M_\alpha)H(\alpha - M_\alpha)')\right\}.
\end{aligned}$$

So, given all other parameters, $\alpha \sim \text{MatVN}(W\Psi^{-1}X'(X\Psi^{-1}X')^{-1}, \Sigma \otimes (X\Psi^{-1}X')^{-1})$. This is also obtained by using a multivariate normal prior and letting the variance of the prior tend to infinity.

### Latent Cutpoints

Recall that $\theta_{k,c}$ denotes the cutpoint for the expanded latent variable $W$ such that any $W$ value below $\theta_{k,c}$ is associated with $Y = c$, and any point above $\theta_{k,c}$ has an associated $Y$ value of $c+1$. The subscript $k$ refers to question $k$. By again using a reference prior, the full conditional distribution of $\theta_{j,c}$ can be derived as follows:

$$\begin{aligned}
P(\theta_{k,c}|...) &\propto \mathcal{L}(\alpha, \Sigma, \theta, \Psi) * P(\theta_{k,c}) \\
&\propto \prod_{i=1}^{n}\prod_{j=1}^{q} \mathcal{I}\{\theta_{j,Y_{i,j}-1} < W_{i,k} \leq \theta_{j,Y_{i,j}}\}. \\
&\propto \prod_{i=1}^{n} \mathcal{I}\{\theta_{k,Y_{i,k}-1} < W_{i,j} \leq \theta_{k,Y_{i,k}}\}\mathcal{I}\{\theta_{k,Y_{i,k}} < W_{i,k} \leq \theta_{k,Y_{i,k}+1}\} \\
&\propto \mathcal{I}\{(\max_i\{W_{i,k}|Y_{i,k}=c\} < \theta_{k,c} \leq \min_i\{W_{i,k}|Y_{i,k}=c+1\})\}.
\end{aligned}$$

Thus, given all other parameters, $\theta_{k,c} \sim \text{Unif}(\max_i\{W_{i,k}|Y_{i,k} = c\}, \min_i\{W_{i,k}|Y_{i,k} = c+1\})$.

The reason for wanting closed forms for the full conditional posterior distributions for $\theta, \alpha, \Sigma$, and $W$ is to do Gibbs sampling. Gibbs steps are easier to implement than the alternative Metropolis steps because Gibbs steps involve direct sampling from the full conditional distribution, as opposed to rejection sampling in the Metropolis step. In fact, the Gibbs algorithm is a special case of the Metropolis-Hastings algorithm where the proposal density is the target density (Carlin and Louis 2000). Unfortunately, we do not have a closed form for the posterior of $\rho$. This leads us to using a Metropolis step for $\rho$, resulting in what is known as a Metropolis-within-Gibbs (MWG) algorithm.

#### Metropolis-within-Gibbs

The Metropolis-within-Gibbs algorithm is an MCMC algorithm that is particularly useful when the full conditional posterior distributions of some parameters have closed forms while the others do not (Ntzoufras 2009). The algorithm is implemented similarly to the Gibbs sampler, except that the parameters with no closed form for the full conditional posterior distribution are sampled using the Metropolis-Hastings algorithm.

## 3.7 Markov Chain Monte Carlo Algorithm

To fit the model, a Metropolis-within-Gibbs algorithm is used. In this particular application, the algorithm is essentially the same as a Gibbs algorithm, except that one of the draws is done by means of proposal-rejection sampling. To run the MCMC algorithm, we first must decide on initial parameter values.

### Initial Parameter Values

For our initial parameter values, we take the same approach as Lawrence et al. (2008). The initial parameter values for the cutpoints, predictive coefficients, and latent variables are chosen by first fitting the model assuming independence between respondents and between questions. This is done using a Gibbs sampling algorithm to ensure that the initial cutpoints, predictive coefficients, and latent variables are plausible. That is, the univariate probit fit allows us to initialize the algorithm at a sensible values for these parameters. In addition, the $R$ matrix of between question correlations is initialized at the identity matrix, and the initial value of $\rho$ is chosen to be 0.4. In our simulated examples, the choice of any sensible starting value for $\rho$ has little to no impact on our inference.

After obtaining initial values, samples of the expanded latent variable $W$, the parameters $\alpha$, $\theta$, and $\Sigma$ are drawn using the full conditional distributions via Gibbs steps. Then, a new $\rho$ is proposed and accepted or rejected based on the Metropolis step. The parameters are then rescaled to map from the expanded parameter space to the original parameter space. The process is then repeated iteratively, yielding MCMC samples of the unexpanded parameters.

### Metropolis Step

As mentioned, a Metropolis step is required to sample the correlation parameter, $\rho$. This Metropolis step is done in the following manner. Propose a candidate correlation parameter, $\rho_c$. This can be done by generating a value from a uniform distribution centred at $\rho_k$. For our purposes, a beta(2,2) prior is used on $\rho$ since values have to be between 0 and 1 and we sought a relatively uninformative prior. This yields a probability of accepting a new candidate $\rho$, which is equal to $\min(t,1)$, where $t$ is

evaluated as follows:

$$\begin{aligned} t &= \frac{f(Z|\alpha,\Sigma,\theta,\Psi_c) * P(\rho_c)}{f(Z|\alpha,\Sigma,\theta,\Psi_{k-1}) * P(\rho_{k-1})} \\ &= \frac{\text{etr}\left\{-\frac{1}{2}\Sigma^{-1}(W-\alpha X)\Psi_c^{-1}(W-\alpha X)'\right\}}{\text{etr}\left\{-\frac{1}{2}\Sigma^{-1}(W-\alpha X)\Psi_{k-1}^{-1}(W-\alpha X)'\right\}} \frac{|\Sigma|^{-n/2}|\Psi_c|^{-q/2}}{|\Sigma|^{-n/2}|\Psi_{k-1}|^{-q/2}} \frac{P(\rho_c)}{P(\rho_{k-1})} \\ &= \frac{\text{etr}\left\{-\frac{1}{2}\Sigma^{-1}(W-\alpha X)\Psi_c^{-1}(W-\alpha X)'\right\}}{\text{etr}\left\{-\frac{1}{2}\Sigma^{-1}(W-\alpha X)\Psi_{k-1}^{-1}(W-\alpha X)'\right\}} \frac{|\Psi_c|^{-q/2}}{|\Psi_{k-1}|^{-q/2}} \frac{(1-\rho_c)\rho_c}{(1-\rho_{k-1})\rho_{k-1}}. \end{aligned}$$

Instead of $t$, $\log(t)$ is calculated in our algorithm because it yields better numerical stability.

$$\begin{aligned} \log(t) &= \log\left\{ \frac{\text{etr}\left\{-\frac{1}{2}\Sigma^{-1}(W-\alpha X)\Psi_c^{-1}(W-\alpha X)'\right\}}{\text{etr}\left\{-\frac{1}{2}\Sigma^{-1}(W-\alpha X)\Psi_{k-1}^{-1}(W-\alpha X)'\right\}} \frac{|\Psi_c|^{-q/2}}{|\Psi_{k-1}|^{-q/2}} \frac{(1-\rho_c)\rho_c}{(1-\rho_{k-1})\rho_{k-1}} \right\} \\ &= -\frac{1}{2}\text{Tr}\left\{\Sigma^{-1}(W-\alpha X)\Psi_c^{-1}(W-\alpha X)' - \Sigma^{-1}(W-\alpha X)\Psi_{k-1}^{-1}(W-\alpha X)'\right\} \\ &\quad -\frac{q}{2}\left\{\log|\Psi_c|) - \log(|\Psi_{k-1}|)\right\} + \left\{\log(1-\rho_c) - \log(1-\rho_{k-1})\right\}. \end{aligned}$$

The MWG algorithm is as follows

1. Draw $W_{k+1}|\alpha_k, \theta_k, \rho_k, \Sigma_k$.
2. Draw $\Sigma_{k+1}|W_{k+1}, \alpha_k, \theta_k, \rho_k$.
3. Draw $\alpha_{k+1}|W_{k+1}, \theta_k, \rho_{k+1}, \Sigma_{k+1}$.
4. Draw $\theta_{k+1}|W_{k+1}, \alpha_{k+1}, \rho_{k+1}, \Sigma_{k+1}$.
5. Draw $\rho_{k+1}$ using a Metropolis step as outlined in the previous section.

Following this, rescale the drawn parameters using the following transformation:

1. $Z_{i,j} = W_{i,j}/\sqrt{\Sigma_{j,j}}$.
2. $R_{i,j} = \Sigma_{i,j}/\sqrt{\Sigma_{i,i} \times \Sigma_{j,j}}$.

3. $\beta_{k,j} = \alpha_{k,j}/\sqrt{\Sigma_{j,j}}$.

4. $\gamma_{j,c} = \theta_{j,c}/\sqrt{\Sigma_{j,j}}$.

Once the rescaling is done, we are effectively drawing samples from our desired posterior distributions using the MWG algorithm.

### Summary of Proposed Algorithm

In summary, the study is designed as follows:

1. Initialize parameters using UVP fit.

2. Draw samples of the expanded parameters $\theta, W, \alpha, \Sigma$ using Gibbs sampling.

3. Draw a sample of $\rho$ using a Metropolis step.

4. Map expanded parameters onto the original parameter space to effectively get samples of $\Gamma, Z, \beta, R$.

5. Repeat steps 2-4 for each MCMC iteration.

In the next chapter, we use this MWG algorithm on some simulated examples for the purposes of parameter estimation. We will be examining the efficacy of the algorithm, as well as some computational issues.

# Chapter 4

# Parameter Estimation and Variable Selection

In this chapter, we will discuss simulated examples aimed at evaluating the proposed methodology. Ideally, our approach would demonstrate good parameter estimation and the ability to identify active or inert predictors and spatial correlations. All simulations and fits are done in Matlab.

## 4.1 Study Design and Model Fitting

To evaluate the proposed approach, we design a set of nine simulation studies to investigate how well the algorithm performs in parameter estimation and variable selection. Ideally, the algorithm would give sensible estimates, as well as the ability to properly identify active or inert predictors and spatial correlations. That is, spatial dependency and the activity of predictors should be properly identified.

In our simulation studies, data are generated using the SMVP model and then fit using the MCMC algorithm as outlined in chapter 3. We consider a three question survey, each with three possible responses using several settings for sample size and

strength of spatial correlation. Ordinal response values will be obtained by first simulating latent variables and using then pre-determined cutpoints to assign actual ordinal values. The model will then be fit via the proposed MWG algorithm using these simulated data.

### Parameter Settings and Study Design

Ordinal data will have to be generated to study the proposed approach. It is relatively easy to simulate the latent variables when given real ordinal values. However, to generate ordinal data from this model, latent variables and cutpoints have to be simulated and used to obtain the actual ordinal values. This process can be difficult. This is because numerical complications can arise if the latent variables yield all or most of the observed ordinal values in one ordered category. No algorithm would be able to give us much information in cases where all responses were identical. In essence, for a particular set of $X$ values, only certain combinations of $\beta$ and $\gamma$ values can be detected.

Consider a case where $\beta$ values were all large, such that $\beta X$ led to large $Z$ means. If the cutpoints were relatively small, any generated set of ordinal values would likely be in one category. That is, the $Y$ values would all be identical. In this scenario, inference would be difficult, if not impossible. As such, the matrix of predictive coefficients $\beta$ and $\Gamma$ has to be chosen carefully to ensure that sensible and plausible ordinal values are actually generated.

In addition to those conditions, we sought slope values that were not identical. This was done to test the algorithm under a setting with some diversity in parameter values. We chose to have have “high”, “medium”, and “low” settings for our regression parameters. In this study, we selected 0, 1, and -3, for our slope values, and 2, 0, and

1 for the associated intercept values. This yielded:

$$\beta = \begin{bmatrix} 2 & -3 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

Given that there were three possible responses to each question, only one cutpoint had to be selected per question. Given that the cutpoints are nuisance parameters, we wanted a simple set of values that would be fairly plausible. For that reason, we set these nuisance cutpoints at one for each question. This yielded the following cutpoint matrix:

$$\Gamma = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 1 & 1 \\ \infty & \infty & \infty \end{bmatrix}.$$

The three "between question" correlations were chosen to be 0.5 (correlation between respones to questions 1 and 2), 0.25 (questions 1 and 3), and 0.1 (questions 2 and 3). These values were chosen to have "high", "medium", and "low" settings like our $\beta$ values. This yielded the correlation matrix:

$$R = \begin{bmatrix} 1.0 & 0.5 & 0.25 \\ 0.5 & 1.0 & 0.1 \\ 0.25 & 0.1 & 1.0 \end{bmatrix}.$$

To investigate the proposed approach under different strengths of spatial correlation, the spatial correlation parameter will also be varied, taking on values of 0, 0.25 and 0.5. Here, we sought scenarios with no spatial correlation, low spatial correlation, and high spatial correlation. To investigate the impact of various sample sizes (low, medium and high), $n$ will take on values of 50, 100, and 200. This yields 9 combinations of sample sizes and spatial correlations, and all 9 will be examined.

Along with parameter values, the design must also be specified. The predictor matrix, $X$, is generated from a latin hypercube design ranging from 0 to 1. The spatial location matrix, $S$, is generated from a latin hypercube design ranging from 0 to 10. In practice, the design could be easily changed, but we elected to use a design with some space-filling properties (Tang 1993). The latent variable matrix, $Z$, is then generated from the matrix-variate normal distribution with a mean of $\beta X$ and a correlation matrix of $R \otimes \Psi$.

For each parameter setting and simulation run, latent variables are simulated from the distribution specified above. Based on our latent values and cutpoints, the ordinal $Y$ responses are then generated as follows: $Y = 1$ if $Z \leq 0$, $Y = 2$ if $0 < Z \leq 1, Y = 3$ if $Z > 1$. These $Y, X$, and $S$ matrices are used as simulated data for model fitting. In essence, we use our design to generate responses at varying parameter settings, and each set of simulated data was fit using the proposed MWG algorithm.

Ideally, the MCMC samples from the posterior distributions of the parameters should be centered somewhere near the true value, the location of the center varying from simulation to simulation. That is, on average, we would want the posterior distribution to be centered at or near the true value. Simulations were repeated 500 times and the medians of parameter samples from the MCMC draws were evaluated. We also wish to examine the efficacy of the algorithm with respect to variable selection. To this end, we first discuss the use of credible intervals.

### Credible Intervals

A 95% credible interval is a range of values such that the posterior probability that a parameter of interest is in that range is 0.95. When sampling using MCMC methods, such an interval can be formed by using the $2.5^{th}$ percentile value as the lower bound and the $97.5^{th}$ percentile value as the upper bound. We take such an approach.

For the purposes of variable selection, 95% credible intervals were evaluated for

each of the slope parameters. This is done based on the MCMC samples to ascertain how often zero is contained in the computed intervals. If zero is contained in the interval for the slope parameter, we deemed the associated predictor to be inert. If zero is not contained in the interval for the slope parameter, the associated predictor was deemed to be active. Ideally, the highly active variable ($\beta_{12} = -3$) will be identified as active almost always, the less active variable ($\beta_{22} = 1$) will be identified as active fairly often, and the inert variable ($\beta_{32} = 0$) will be identified as inert around 95% of the time. In practice, one could adjust the intervals for multiple comparisons. However, for the sake of simplicity, we take the outlined approach.

**Summary of Study Design**

In summary, the study is designed as follows:

1. Parameters are set at values listed above.
2. Samples sizes of $n = 50$, 100, and 200 and spatial correlations $\rho = 0$, 0.25 and 0.5 are used for each of the nine simulation studies.
3. Use a latin hypercube design ranging from 0 to 1 for $X$ and ranging from 0 to 10 for $S$.
4. For each simulation run, draw latent variable matrix $Z$ based listed parameter values and design.
5. Based on the $Z$ values and cutpoints, obtain associated ordinal responses $Y$.
6. Use $Y, X$, and $S$ as data to fit SMVP model using the outlined MWG algorithm.
7. Obtain point estimates for parameters and 95% credible intervals for variable selection.
8. Repeat for each of the 500 simulations in the nine studies.

## 4.2 Parameter Estimation Results

The point estimate from each simulation was taken to be the median of posterior samples from the MCMC algorithm. There were only slight differences between median and mean values, so we elected the measure less affected by potential outliers from our MCMC samples. Ideally, the point estimates would be centered around the truth. The following are the results from our simulation studies.

### 4.2.1 Results

#### Estimation of R

In the nine studies, the 500 simulations each yielded point estimates for all the parameters. The following tables contain means of the 500 point estimates and the associated root mean square errors (RMSE) for each case, where each point estimate is a posterior median. As mentioned, the nine combinations for $n = 50, 100, 200$ and $\rho = 0, 0.25, 0.5$ are examined. Relatively small RMSE values would also suggest good estimation. As well, a decrease in RMSE with increasing sample sizes would indicate improved estimation with higher sample size.

Table 4.1: Mean of point estimates (and RMSE) of $R_{12} = 0.5$

| | $n = 50$ | $n = 100$ | $n = 200$ |
|---|---|---|---|
| $\rho = 0.00$ | 0.54 (0.18) | 0.54 (0.13) | 0.54 (0.09) |
| $\rho = 0.25$ | 0.48 (0.20) | 0.49 (0.16) | 0.48 (0.13) |
| $\rho = 0.50$ | 0.45 (0.23) | 0.46 (0.19) | 0.45 (0.17) |

Table 4.2: Mean of point estimates (and RMSE) of $R_{13} = 0.25$

| | $n = 50$ | $n = 100$ | $n = 200$ |
|---|---|---|---|
| $\rho = 0.00$ | 0.27 (0.21) | 0.27 (0.15) | 0.26 (0.10) |
| $\rho = 0.25$ | 0.24 (0.26) | 0.25 (0.20) | 0.24 (0.15) |
| $\rho = 0.50$ | 0.22 (0.26) | 0.23 (0.22) | 0.23 (0.19) |

Table 4.3: Mean of point estimates (and RMSE) of $R_{23} = 0.1$

| | $n = 50$ | $n = 100$ | $n = 200$ |
|---|---|---|---|
| $\rho = 0.00$ | 0.12 (0.21) | 0.12 (0.15) | 0.10 (0.10) |
| $\rho = 0.25$ | 0.11 (0.23) | 0.11 (0.19) | 0.09 (0.17) |
| $\rho = 0.50$ | 0.09 (0.26) | 0.09 (0.22) | 0.10 (0.19) |

Looking at Table 4.1, we see that the correlation parameter ($R_{12}$=0.50) is being fairly well estimated. For fixed values of the spatial parameter, $\rho$, we see that as $n$ increases, the RMSE decreases. For fixed $\rho$, we see that the RMSE appears to increase slightly as the spatial dependence becomes more severe. These observations are also seen for the other values of R (e.g., Tables 4.2 and 4.3).

**Estimation of Intercept Terms**

Moving on to the intercept terms (Tables 4.4-4.6), we see that the results are somewhat mixed.

Table 4.4: Mean of point estimates (and RMSE) of $\beta_{11} = 2$

| | $n = 50$ | $n = 100$ | $n = 200$ |
|---|---|---|---|
| $\rho = 0.00$ | 1.8 (0.47) | 1.8 (0.34) | 1.9 (0.25) |
| $\rho = 0.25$ | 2.0 (0.61) | 2.2 (0.59) | 2.3 (0.68) |
| $\rho = 0.50$ | 2.4 (0.91) | 2.6 (1.03) | 2.8 (1.19) |

Table 4.5: Mean of point estimates (and RMSE) of $\beta_{21} = 0$

| | $n = 50$ | $n = 100$ | $n = 200$ |
|---|---|---|---|
| $\rho = 0.00$ | -0.04 (0.31) | -0.05(0.24) | 0.00 (0.16) |
| $\rho = 0.25$ | -0.03 (0.44) | -0.02 (0.43) | 0.02 (0.45) |
| $\rho = 0.50$ | -0.02 (0.59) | 0.01 (0.66) | -0.04 (0.75) |

Table 4.6: Mean of point estimates (and RMSE) of $\beta_{31} = 1$

| | $n = 50$ | $n = 100$ | $n = 200$ |
|---|---|---|---|
| $\rho = 0.00$ | 0.9 (0.34) | 0.9 (0.26) | 0.9 (0.19) |
| $\rho = 0.25$ | 1.0 (0.48) | 1.1 (0.52) | 1.2 (0.50) |
| $\rho = 0.50$ | 1.1 (0.70) | 1.3 (0.81) | 1.4 (0.91) |

Notice that when there is no spatial dependence that the intercept is fairly well estimated. This holds particularly true as sample sizes increase, as evidenced by lower RMSE values. When there is a non-zero intercept and strong spatial dependence ($\rho = 0.5$), the methodology has a more difficult time estimating the intercept term. However, point estimates tended to be in the right neighbourhood. That is, larger values tended to yield larger estimates. This is also the case with slope coefficients.

**Estimation of Slope Coefficients**

The following tables show how the slopes were estimated in our study.

Table 4.7: Mean of point estimates (and RMSE) of $\beta_{12} = -3$

| | $n = 50$ | $n = 100$ | $n = 200$ |
|---|---|---|---|
| $\rho = 0.00$ | -2.77 (0.74) | -2.79 (0.54) | -2.84 (0.37) |
| $\rho = 0.25$ | -3.07 (0.80) | -3.34 (0.67) | -3.53 (0.75) |
| $\rho = 0.50$ | -3.58 (1.09) | -3.92 (1.19) | -4.32 (1.51) |

Table 4.8: Mean of point estimates (and RMSE) of $\beta_{22} = 1$

| | $n = 50$ | $n = 100$ | $n = 200$ |
|---|---|---|---|
| $\rho = 0.00$ | 0.90 (0.55) | 0.95 (0.41) | 0.90 (0.29) |
| $\rho = 0.25$ | 1.02 (0.42) | 1.10 (0.30) | 1.17 (0.28) |
| $\rho = 0.50$ | 1.16 (0.47) | 1.32 (0.46) | 1.48 (0.56) |

Table 4.9: Mean of point estimates (and RMSE) of $\beta_{32} = 0$

| | $n = 50$ | $n = 100$ | $n = 200$ |
|---|---|---|---|
| $\rho = 0.00$ | -0.03 (0.57) | 0.03(0.37) | -0.01 (0.29) |
| $\rho = 0.25$ | 0.02 (0.39) | -0.03 (0.24) | 0.00 (0.16) |
| $\rho = 0.50$ | -0.01 (0.36) | -0.01 (0.24) | -0.00 (0.15) |

Again, notice that the algorithm performs well in the absence of spatial correlation. Furthermore, Tables 4.7-4.9 suggest that the algorithm does fairly well in the presence of small spatial correlation. These estimates seem to be helped by increased sample size. However, some difficulties with estimation of non-zero slopes in the presence of high spatial correlation were observed. Still, positive "true values" usually yielded positive estimates. Also, larger values tended to yield larger estimates. Estimation of the nuisance cutpoints yielded a similar situation.

### Estimation of $\Gamma$ Cutpoints

The following tables show how the nuisance cutpoints were estimated in our study.

Table 4.10: Mean of point estimates (and RMSE) of $\gamma_{21} = 1$

| | $n = 50$ | $n = 100$ | $n = 200$ |
|---|---|---|---|
| $\rho = 0.00$ | 0.84 (0.26) | 0.88 (0.20) | 0.89 (0.15) |
| $\rho = 0.25$ | 0.98 (0.27) | 1.08 (0.23) | 1.17 (0.25) |
| $\rho = 0.50$ | 1.19 (0.40) | 1.28 (0.39) | 1.43 (0.50) |

Table 4.11: Mean of point estimates (and RMSE) of $\gamma_{22} = 1$

| | $n = 50$ | $n = 100$ | $n = 200$ |
|---|---|---|---|
| $\rho = 0.00$ | 0.82 (0.25) | 0.87 (0.19) | 0.89 (0.15) |
| $\rho = 0.25$ | 1.00 (0.24) | 1.08 (0.22) | 1.19 (0.26) |
| $\rho = 0.50$ | 1.19 (0.38) | 1.33 (0.45) | 1.48 (0.55) |

Table 4.12: Mean of point estimates (and RMSE) of $\gamma_{23} = 1$

| | $n = 50$ | $n = 100$ | $n = 200$ |
|---|---|---|---|
| $\rho = 0.00$ | 0.84 (0.26) | 0.87 (0.20) | 0.89 (0.14) |
| $\rho = 0.25$ | 1.00 (0.28) | 1.09 (0.26) | 1.18 (0.26) |
| $\rho = 0.50$ | 1.20 (0.44) | 1.33 (0.51) | 1.47 (0.61) |

Once again, we see fairly decent estimation of the parameters in the absence of spatial correlation. As well, Tables 4.10-4.12 demonstrate that the RMSE values tend to decrease with increased sample size in the absence of spatial correlation. However, the algorithm seems to have trouble with estimating the cutpoints in the presence of spatial correlation. Still, sensible estimates tend to be obtained, whereby cutpoints are still estimated in the correct neighbourhood.

### Estimation of $\rho$

The estimation of $\rho$ also yielded mixed results.

Table 4.13: Mean of point estimates (and RMSE) of $\rho$

| | $n = 50$ | $n = 100$ | $n = 200$ |
|---|---|---|---|
| $\rho = 0.00$ | 0.00 (0.00) | 0.00 (0.00) | 0.00 (0.00) |
| $\rho = 0.25$ | 0.20 (0.10) | 0.15 (0.12) | 0.12 (0.14) |
| $\rho = 0.50$ | 0.27 (0.25) | 0.21 (0.30) | 0.15 (0.36) |

It is evident from Table 4.13 that $\rho$ seems to be underestimated when the true value is non-zero. However, the algorithm seems to detect the presence of spatial correlation when there is spatial correlation and seems to detect no spatial correlation when there is none. In terms of properly identifying an active or inert spatial correlation, the algorithm seems to be performing fairly well.

**Summary of Parameter Estimation**

The tables above demonstrate mixed results when it comes to parameter estimates. In the absence of spatial correlation, most parameters are estimated fairly well. As seen, increasing sample size seems to reduce RMSE values in this situation. Some parameters are also estimated well in the presence of low spatial correlation. However, with the exception of the matrix $R$ of "between question" correlations, it seems that there are difficulties with estimating most parameter in the presence of high spatial correlation. Despite this, it seems as if the algorithm tends to yield fairly sensible numbers, whereby estimates are still in the correct neighbourhood. While there are evidently some challenges with parameter estimation, variable selection may still prove fruitful.

## 4.3 Variable Selection

Variable selection is an important activity. Methods that can identify active or inert predictors are highly desirable in applied settings. In this section, we examine the efficacy of the algorithm when it comes to variable selection.

For our purposes, 95% credible intervals were computed for each simulation run, and predictors were deemed active if the interval for the associated regression coefficient did not contain 0. Likewise, a predictor was deemed inert if the associated interval did contain zero. As mentioned, the three slope values examined were -3, 1, and 0.

### 4.3.1 Results

In the nine studies, coverage proportions of zero were computed for the slope values. We would expect to see higher magnitudes of the true slope associated with lower coverage probabilities. That is to say, the higher the true slope, the more likely it is to appear active. For the value zero, would expect the coverage proportion to be around 95%. The following tables show how often 0 is contained in the 95% credible interval for each slope parameter. For $\beta_{12} = -3$ and $\beta_{22} = 1$, these proportions represent how often the predictor was incorrectly identified as inert. For $\beta_{32} = 0$, the proportions represent how often the predictor was correctly identified as inert.

Table 4.14: Proportion of times 0 was in the 95% credible interval for $\beta_{12} = -3$

| | $n = 50$ | $n = 100$ | $n = 200$ |
|---|---|---|---|
| $\rho = 0.00$ | 0.0040 | 0 | 0 |
| $\rho = 0.25$ | 0 | 0 | 0 |
| $\rho = 0.50$ | 0 | 0 | 0 |

Table 4.15: Proportion of times 0 was in the 95% credible interval for $\beta_{22} = 1$

| | $n = 50$ | $n = 100$ | $n = 200$ |
|---|---|---|---|
| $\rho = 0.00$ | 0.5840 | 0.2720 | 0.0580 |
| $\rho = 0.25$ | 0.2460 | 0.0140 | 0 |
| $\rho = 0.50$ | 0.1300 | 0.0040 | 0 |

Table 4.16: Proportion of times 0 was in the 95% credible interval for $\beta_{32} = 0$

| | $n = 50$ | $n = 100$ | $n = 200$ |
|---|---|---|---|
| $\rho = 0.00$ | 0.9460 | 0.9380 | 0.9240 |
| $\rho = 0.25$ | 0.9520 | 0.9660 | 0.9400 |
| $\rho = 0.50$ | 0.9620 | 0.9700 | 0.9740 |

The algorithm seems to be fairly successful at detecting active predictors. As seen in Tables 4.14-4.16, the credible intervals cover 0 most often when the true value is 0, less often when the true value is 1 and least often when the true value is -3. As expected, there are some problems identifying moderately active predictors ($\beta_{22} = 1$) when the sample size is low. However, it is evident that proper identification is helped by increased sample size. As well, the inert factor is being deemed inert about 95% of the time. In short, the algorithm is fairly efficient with respect to distinguishing between active and inert predictors.

## 4.4 Discussion

Our goal in this project was to construct a model in which both "between question" and "between respondent" correlations could be incorporated. We also wanted to propose an approach to allow for Bayesian inference. This was accomplished using the Metropolis-within-Gibbs algorithm to fit the spatial multivariate probit model. Ideally, our fitting algorithm would have yielded perfectly centered point estimates

with small error terms. The algorithm seemed to do fairly well in situations with no spatial correlations and sometimes with small spatial correlations. There were some difficulties with parameter estimation in cases with high spatial correlation. Still, parameter estimation yielded fairly sensible values, with estimates usually in the neighbourhood of the truth.

In the interest of space, 95% credible interval coverage probabilities of the true parameter values were not included. As one might expect, in situations where parameters were well estimated, coverage probabilities were around 95%. Also expected, estimates that were less accurate tended to have associated probabilities that were lower. There was little additional insight to be gained from these coverage probabilities.

Despite facing some challenges with parameter estimation in the presence of high spatial correlation, the algorithm performs fairly well when dealing with variable selection. Active spatial correlations, as well as predictors, were successfully detected. In addition, inert predictors were usually deemed inert, and the absence of spatially correlation was also successfully identified. From a practical standpoint, the ability of the proposed approach in detecting these properties seems highly important.

# Chapter 5

# Conclusion and Future Work

In the proposed approach to correlated ordinal data, we develop the spatial multivariate probit model to incorporate both "between question" and "between respondent" dependency. The proposed method yielded mixed results, with variable selection performing well despite some challenges with parameter estimation in the presence of high spatial correlation.

One feasible way to improve upon this may be to use different prior distributions. For example, One possible change would be the use of an Inverse-Wishart prior for the expanded covariance matrix, $\Sigma$. The improper priors on certain parameters may have skewed results in the model fits and it would be worthwhile investigating this.

Another area for future work would be goodness-of-fit testing (Muthukumarana 2010). Methods that can test the validity of the SMVP model may be very useful in various applied settings.

Prediction is another area for future work. One thing of interest is the ability to predict the ordinal response at a new set of predictors and spatial locations. In Higgs and Hoeting (2010), the authors generate latent values at each iteration of their MCMC algorithm. Clipping based on cutpoints and the latent values, they predict the ordinal response of a new point to be the most commonly observed category. A

similar topic of interest would be the predicton of the exact probabilities associated with different categories at a new set of predictors and locations. This could be done by computing the proportion of times each category is observed based on MCMC generated latent values.

An alternative approach would be to perform Kriging (Cressie 1993) on the latent space. Kriging would allow us to find the conditional distribution of a new point based on our model and existing data. Using this conditional distribution at each MCMC iteration, we could compute probabilities for each categories (or set of categories) and average over each iteration.

# Bibliography


Albert, J. H. and Chib, S. (1993), "Bayesian Analysis of Binary and Polychotomous Response Data," *Journal of the American Statistical Association*, 88, pp. 669–679.

Carlin, B. and Louis, T. (2000), *Bayes and Empirical Bayes methods for data analysis*, Texts in statistical science, Chapman & Hall/CRC.

Chib, S. and Greenberg, E. (1998), "Analysis of Multivariate Probit Models," *Biometrika*, 85, pp. 347–361.

Cressie, N. A. C. (1993), *Statistics for Spatial Data (Wiley Series in Probability and Statistics)*, Wiley-Interscience, rev sub ed.

De Oliveira, V. (2000), "Bayesian Prediction of Clipped Gaussian Random Fields," *Comput. Stat. Data Anal.*, 34, 299–314.

Higgs, M. D. and Hoeting, J. A. (2010), "A Clipped Latent Variable Model for Spatially Correlated Ordered Categorical Data," *Comput. Stat. Data Anal.*, 54, 1999–2011.

Johnson, V. E. and Albert, J. H. (1999), *Ordinal Data Modeliing*, New York: Springer.

Lawrence, E., Bingham, D., Liu, C., and Nair, V. (2008), "Bayesian Inference for Multivariate Ordinal Data Using Parameter Expansion," *Technometrics*, 45, 80–89.

Likert, R. (1932), "A Technique for the Measurement of Attitudes." *Archives of Psychology*, 22, 1–55.

Liu, C., Rubin, D. B., and Wu, Y. N. (1998), "Parameter Expansion to Accelerate EM: The PX-EM Algorithm," *Biometrika*, 85, pp. 755–770.

Liu, J. S. and Wu, Y. N. (1999), "Parameter Expansion for Data Augmentation," *Journal of the American Statistical Association*, 94, pp. 1264–1274.

McCullagh, P. (1980), "Regression Models for Ordinal Data," *Journal of the Royal Statistical Society. Series B (Methodological)*, 42, pp. 109–142.

McMillan, N., Sacks, J., Welch, W., and Gao, F. (1999), "Analysis of Protein Activity Data by Gaussian Stochastic Process Models." *Journal of Biopharmaceutical Statistics*, 9, 145.

Muthukumarana, S. (2010), "Bayesian Methods and Applications using WinBUGS," *Ph.D Thesis, Simon Fraser University*, 67–69.

Ntzoufras, I. (2009), *Bayesian modeling using WinBUGS*, Wiley series in computational statistics, Wiley.

Petersen, K. B. and Pedersen, M. S. (2008), "The Matrix Cookbook," Version 20081110.

Qian, P. Z. G., Wu, H., and Wu, C. F. J. (2008), "Gaussian Process Models for Computer Experiments With Qualitative and Quantitative Factors," *Technometrics*, 50, 383–396.

Tang, B. (1993), "Orthogonal Array-Based Latin Hypercubes," *Journal of the American Statistical Association*, 88, pp. 1392–1397.