RICE UNIVERSITY

Bayesian Inference for Ordinal Data

by

Xian Zhou

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy

APPROVED, THESIS COMMITTEE:

B. Nebiyou Bekele
Assistant Professor of Biostatistics & Applied Mathematics, UT-MDACC

Dennis D. Cox, Committee Chair
Professor of Statistics

K. B. Ensom
Professor and Chair of Statistics

David Lane
Professor of Psychology

Peter Müller, Thesis Adviser
Professor of Biostatistics & Applied Mathematics, UT-MDACC

HOUSTON, TEXAS

NOVEMBER, 2005
INFORMATION TO USERS

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleed-through, substandard margins, and improper alignment can adversely affect reproduction.

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

UMI

UMI Microform 3216811
Copyright 2006 by ProQuest Information and Learning Company. All rights reserved. This microform edition is protected against unauthorized copying under Title 17, United States Code.

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346
Abstract

Bayesian Inference for Ordinal Data

by

Xian Zhou

Albert & Chib proposed a Bayesian ordinal probit regression model using the Gibbs sampler to model ordinal data. Their method defines a relationship between latent variables and ordinal outcomes using cutpoint parameters. However, the convergence of this Gibbs sampler is slow when the sample size is large because the cutpoint parameters are not efficiently sampled. Cowles proposed a Gibbs/Metropolis-Hastings (MH) sampler that would update cutpoint parameters more efficiently. In the context of longitudinal ordinal data, this algorithm potentially require the computation of cumulative probability of a multivariate normal distribution to calculate the acceptance probability for the MH sampler. We propose a probit model where the latent variables follow a mixture of normal distributions. This mixture structure can successfully characterize the ordinality of data while holding the cutpoint parameters constant. Gibbs samplings along with reversible jump MCMC are carried out to es-
timate the size of the mixture. We adopt this idea in modeling ordinal longitudinal data, where the autoregressive error (1) model is proposed to characterize the underlying correlation structure among the repeated measurements. We also propose a Bayesian probabilistic model in estimating the clustering membership using a mixture of Gaussian distributions to tackle the problem of clustering ordinal data. Results are compared with those obtained from K-means method. We further extend the multinomial probit (MNP) model and develop a joint MNP and ordinal probit model to model the cell probabilities for multiple categorical outcomes with ordinal variables nested within each categorical outcome. A hierarchical prior is imposed on the location parameters of the normal kernels in the mixture model associated with the ordinal outcomes. Our model has wide applications in various fields such as clinical trials, marketing research, and social science.
Acknowledgements

I am very grateful to my thesis adviser, Dr. Peter Müller, for the guidance and encouragement he has given me over the past three years. I would like to give special thanks to Dr. Nebiyou Bekele for playing multiple roles in the past five years: committee member, supervisor and friend. I would also like to thank Dr. Dennis Cox for chairing my committee and his advice and insights on my graduate study. Thanks to Drs. Kathy Ensor and Rudy Guerra, for their great help during my graduate study. I would also like to thank Dr. David Lane for being a member of my thesis committee.

Many thanks to the Rice STAT graduate students with whom I took the graduate courses and qualified exams that are part of the graduate student experience. Additionally, I also want to thank my coworkers at the Department of Biostatistics and Applied Mathematics, M.D. Anderson Cancer Center for their generous help and support.

Last, but certainly not the least, I would like to thank my dear husband, Shouchen (Sean) Peng, for his patience, support, strength, and inspirations, and for being with me through good and bad times. Special thanks also go to the most wonderful and amazing parents in the world, without them, I could not be where I am today.
Contents

Abstract ii

Acknowledgements iv

List of Figures viii

List of Tables xi

1 Introduction 1

1.1 Motivation ........................................ 1

1.2 Review of Mixture Models ....................... 5
  1.2.1 Model Selection Using Bayes Factors .......... 8
  1.2.2 Reversible Jump Markov Chain Monte Carlo .... 13
  1.2.3 Pseudoprior (PP) and ‘Metropolized’ PP Method ... 14
  1.2.4 Birth-and-Death Markov Chain Monte Carlo .... 17

2 A Mixture Model for Modeling Ordinal Longitudinal Data 19

2.1 Introduction ..................................... 19
2.2 Model and Prior Specifications ........................................... 21
  2.2.1 Model Specifications ............................................ 21
  2.2.2 Prior Specifications ............................................ 24
2.3 Vague Prior Algorithm by Albert & Chib ............................. 26
2.4 Hybrid of Gibbs/Metropolis-Hastings ............................... 29
2.5 Fixed Size Mixture Normal Distributions ............................ 32
2.6 Random Size Mixture Normal Distributions .......................... 36
2.7 Applications .......................................................... 40
  2.7.1 Simulated Dataset I ............................................ 40
  2.7.2 Simulated Dataset II ........................................... 42
  2.7.3 A Motivational Intervention Study in Houston, Texas Area For
        High Risk Young Smokers ........................................ 45
2.8 Conclusions .......................................................... 47

3 Clustering Analysis for Multivariate Ordinal Data ................... 73
  3.1 Introduction ........................................................ 73
  3.2 Bayesian Clustering Analysis for Multivariate Ordinal Data ....... 76
  3.3 Priors and Posteriors .............................................. 78
    3.3.1 Prior Elicitations ............................................ 78
    3.3.2 Gibbs Sampling: Full Conditional Distributions ............... 80
    3.3.3 Reversible Jump MCMC ...................................... 81
  3.4 Applications ...................................................... 85
3.4.1 A Simulated Dataset .................................. 85
3.4.2 Clustering Analysis of Biomarker Study ............. 87
3.5 Conclusions ............................................. 90

4 A Hierarchical Model for Ordinal Data Nested in Categorical Data 99

4.1 Introduction ............................................. 99
4.2 A Phase III Clinical Trial ................................. 102
4.3 Hierarchical Model for Ordinal Data Nested in Categorical Data .... 104
4.4 Priors and Posteriors .................................... 107
   4.4.1 Prior Elicitations .................................... 107
   4.4.2 Full Conditional Distributions ........................ 108
   4.4.3 Marginal Posterior Probability ........................ 109
4.5 Applications ............................................. 110
   4.5.1 A Simulated Dataset .................................. 110
   4.5.2 A Phase III Clinical Trial of Retinoid Isotretinoin .......... 111
4.6 Conclusions ............................................. 113

5 Summary and Extensions ................................. 121

5.1 Summary ................................................. 121
5.2 Extensions .............................................. 123

Bibliography .............................................. 126
List of Figures

2.1 Directed acyclic graph: the squares represent fixed or observed quantities and the circles are for the unknowns. .......................... 52

2.2 Trace plots of parameters under the Albert-Chib method for the simulated dataset I ......................................................... 63

2.3 Trace plots of parameters under the fixed size mixture model for the simulated dataset I ......................................................... 64

2.4 Trace plots of parameters under the random size mixture model for the simulated dataset I ......................................................... 65

2.5 Marginal posterior density plots for parameters associated with the simulated dataset I ......................................................... 66

2.6 Density plots of marginal posterior distribution of the size of mixture, $G$ 67

2.7 Comparisons of marginal posterior predictive probabilities ............ 68

2.8 Distribution of participant’s characteristics versus the stage of readiness to quit: 0: precontemplation; 1: contemplation; 2: preparation; 3: action ................................................................. 69
2.9 Distribution of participant’s characteristics versus the stage of readiness to quit: 0: precontemplation; 1: contemplation; 2: preparation; 3: action ...................................................... 70

2.10 Posterior distributions of the regression coefficients under the fixed size mixture model (G=3) .......................................................... 71

2.11 Posterior distributions of the regression coefficients under the random size mixture model ......................................................... 72

3.1 Marginal posterior distribution of the size of mixture .................... 93

3.2 Distribution of latent variables represented by clusters under different sizes of mixture of bivariate distributions. The number at the top of each panel indicates the size of mixture. The different types and colors of points stand for different clusters. The size of points reflects the posterior probability of being classified into that cluster. Data is identified to be consisted of 3 (top two panels) or 4 (bottom two panels) clusters. .......................................................... 94

3.3 Convergence plots for the clustering of two biomarkers: ergodic estimates of G, solid line denotes G=4. ............................................. 95

3.4 Clustering of patients based on Caspase and Fus1. Left: clustering using K-means; Right: mixture model of bivariate normal distributions using RJMCMC to identify the size of clusters. ..................... 96
3.5 Convergence plots for the clustering of three biomarkers: ergodic estimates of $G$, solid line denotes $G=3$. .................................................. 97

3.6 Clustering analysis of patients conditional on three biomarker expression levels: Group 1: $n=51$; Group 2: $n=33$; ........................................ 98

4.1 Marginal posterior distributions of the treatment effect on ordinal outcomes ($\beta$s) ......................................................... 120
List of Tables

2.1 Initial values of parameters under each model ........................ 51
2.2 Marginal posterior estimates of the parameters of interest: $\beta_1$: binary covariate; $\beta_2$: continuous covariate; $\rho$: autoregressive coefficient. ... 51
2.3 Marginal posterior distribution of the size of mixture: $G$ ........... 52
2.4 Marginal posterior predicted ordinal response at four time points. The number inside the parenthesis represents the row percentage. ....... 53
2.5 Marginal posterior predicted ordinal response under the fixed size mixture model ......................................................... 55
2.6 Marginal posterior distribution of $G$ ..................................... 57
2.7 Posterior mean, 95% credible interval, and the tail probability of parameters computed using three methods. $\beta_1$: age; $\beta_2$: the number of cigarettes per day at the baseline; $\beta_3$: ever quitting smoking before, yes vs. no; $\beta_4$: treatment assignment, 1: motivational interview, 0: standard ................................................................. 58
2.8 Marginal posterior predicted response (readiness to quit) at four time points: 0: precontemplation; 1: contemplation; 2: preparation; 3: action. The first line in each block is the estimations using the Albert-Chib method. The results in the second line are obtained from the fixed size mixture and those in the last line are from the random size mixture

3.1 Marginal posterior estimation of the number clusters and the corresponding clustering membership. The number of corrected estimated components is bolded under each scenario.

3.2 Distributions of biomarker expression levels. The percentage inside the parenthesis is the column percentage.

4.1 Toxicity frequency for randomized eligible patients by study arms.

4.2 The generated data sorted by the group membership.

4.3 Simulated dataset, marginal posterior cell probabilities with 95% credible intervals. In bold are the empirical cell probabilities.

4.4 Marginal posterior cell probabilities (95% credible intervals) of toxicity.

4.5 Posterior probabilities of particular events: \( p_{i1} \) and \( p_{i2} \) represent the marginal posterior probability of grade 1 toxicity in the isotretinoin group and placebo group, respectively. Likewise, the subscripts 2 and 3 in the other two columns correspond to grade 2 and grade 3 toxicity.
Chapter 1

Introduction

1.1 Motivation

Ordinal data is a frequently encountered data type. Data are often classified into several groups with some intrinsic ordering constraints such as smoking status (non-smoker, former smoker, current smoker). There is a growing interest in modeling ordinal data, especially correlated ordinal data. Most of time, this type of data will be treated as continuous data or reduced to binary outcomes to simplify the presentations and analysis. However, these are not appropriate and could result in underestimation of the variance and considerable loss of information. Many authors have proposed various methods in modeling these types of data both in classical approaches [44] and in a Bayesian framework [2], [22], [17]. One of the most natural ways to model ordinal data is to introduce an underlying latent variable which follows
a continuous distribution. These models assume that the latent variable is mapped to an ordinal variable based on the interval. The observed ordinal outcome is linked with the latent variable through a set of cutoff points. Therefore, the probability of an ordinal outcome is equivalent to the probability of a continuous variable falling into an interval on the real line.

There are several challenges with ordinal data modeling. One of the issues is how to estimate the cutpoints parameters. Albert & Chib [2] and Chib & Greenberg [14] proposed an ordinal probit model in a Bayesian framework, which includes a vague prior on the cutpoint parameters. Proceeding with this approach, we run into the following problem. Convergence of the Gibbs sample implemented by simulating from the full conditional distributions is very slow when the sample size is large. Cowles [16] proposed a hybrid Gibbs/Metropolis-Hastings (MH) sampling schema which update the cutpoint parameters jointly with other parameters in the multivariate distributions. This approach reduces the high auto-correlation and reaches practical convergence within a reasonable number of iterations. In the context of longitudinal ordinal data with the same cutpoints for each ordinal variable, Cowles' method may require the computation of cumulative probabilities of multivariate normal distributions in order to calculate the acceptance probability of Metropolis-Hasting samplers. Therefore, the computation of the cumulative joint distribution becomes intractable. However, the model requires $K_j - 2$ (Let $K_j$ denote the number of categories for each $j^{th}$ ordinal outcome.) cutpoints for each ordinal variable. For a large number of ordinal variables,
or large $K_j$, the computational effort for the estimation of cutpoint parameters can become prohibitive. Given the difficulty in estimating cutpoint parameters jointly with other parameters, we propose a mixture model which can successfully model the ordinal property of data without the need to estimate the cutpoint parameters. Gibbs sampling with reversible jump Markov Chain Monte Carlo (RJMCMC) are carried out to estimate the size of the mixture distribution. We adopt the idea of mixture models for the latent variables in characterizing ordinal longitudinal data, where the autoregressive structure is proposed to model the underlying correlation structure between the repeated measurements (Chapter 2).

The second issue is how to cluster objects based on multiple ordinal outcomes, i.e., how to partition a set of observations into a distinct number of unknown groups or clusters in such a manner that all observations within a group are similar, while observations in different groups are different. Hierarchical and nonhierarchical clustering methods are two broad classes used in the clustering analysis. Ordinal data is often treated as interval-scaled data in the clustering analysis. The raw data is replaced by its rank to compute the dissimilarity using methods for interval-scaled variables. As we have discussed already, treating ordinal data as other data types is not appropriate. So far little research has been done on the clustering analysis of ordinal data in a Bayesian framework. Due to the property of the mixture model, it can be used in the data classification. This research is motivated mainly by a problem of clustering cancer patients based on a set of ordinal valued biomarkers’ outcomes. We propose a
Bayesian probabilistic model to estimate the clustering membership using a mixture of Gaussian distributions to tackle the problem of multivariate clustering (Chapter 3). One additional feature of this model is that there is no need to pre-specify the number of clusters by using RJMCMC.

We further extend our model to fit a set of ordinal data nested within categorical data. Data is structured at two levels. The first stage of data is categorical data. In the second stage, i.e., within each category, data may take different ordinal values. These types of data are observed in many fields such as clinical trials, marketing research, political science, and cost-effective analysis etc. For example, in a phase III study, people are interested not only in the efficacy of the study agent but also in the safety of the treatment. Patients' toxicity profile such as the type of toxicity and the grade of each toxicity is recorded. It is of interest to know how the toxicity profile differs between different treatment groups in terms of both types and grades. We lay out the statistical foundations of the model in the Chapter 4.

A Bayesian framework is used for inference for several reasons. First of all, Bayesian inference allows us to treat classification in a straightforward manner using finite mixtures. This happens through data augmentation technique, whereby group indicators are created and treated as parameters that facilitate the numerical integration of the posterior density. Simulated values of these parameters provide posterior densities of these indicators. Second, mixture models are inherently difficult to estimate, due
to the identification issue. The Bayesian approach helps to identify the model by inputting adequate prior information. A third reason is the need to infer the number of groups. This is conveniently done by computing posterior probabilities on the range of values deemed a priori plausible. The last but not the least is that predictive inference can be easily performed in MCMC without any extra effort.

The remainder of this chapter is devoted to a review of theoretical work pertinent to the model selection of mixture models, from Bayes factor to pseudo-prior [8], reversible jump Markov chain Monte Carlo [30], and birth-death Markov chain Monte Carlo [61].

1.2 Review of Mixture Models

The Bayesian mixture modeling has attracted much attention recently. The standard framework of modeling assumes that the available data is independently and identically distributed, which usually results in a homogenous distribution. However, in most applications such as genetic studies, marketing research, etc, data are sampled from multiple sources, and thus, exhibit heterogeneous distributions. Statistical remedies include nonparametric approaches and elaborated parameterized distributions by acknowledging the complexity of data. However, the nonparametric approach may not be very powerful and require a large sample of data to validate the model [58]. We introduce mixture model to analyze data from heterogeneous resources (and
multinomial data).

Mixture modeling concerns modeling a statistical distribution using a mixture (or weighted sum) of other distributions. One common application of the mixture model is a way of identifying the two separate distributions without knowing which data values comprise the first peak and which data values comprise the second peak. They are able to represent arbitrarily complex probability density functions. The general representation of the mixture model can be written as the follows:

$$y \propto \sum_{g=1}^{G} p_g f(y|\theta_g),$$

where $\theta$ can be a single parameter or a vector of parameters. The $p_g$ is the weight associated with each component. It follows a binomial distribution if $G$ is two. Otherwise, it follows a multinomial distribution, in which case, Dirichlet prior is the conjugate prior for the weight. Some common distributional forms of mixtures include normal distribution, exponential distribution, binomial distribution, etc. Alternatively, the mixture model can be rewritten in a hierarchical form by introducing a latent variable $w$, which identifies the component of the mixture where each $y$ belongs to. The probability of $w$ is the same as the weight of each component.

$$y|w = g \propto f(y|\theta_g), \quad \Pr(w = g) = p_g$$

Special cautions should be taken in mixture modeling. Even though the likelihood of the mixture model can be written down but it is poorly defined and computa-
tionally difficult. There might be the situations where no observations are allocated to the component, i.e., no information from the sample is provided to assist the estimation. Hence, the likelihood becomes unbounded. Another basic feature of the mixture model is that it is invariant under the permutation of components. Hence, the component parameters are not identifiable marginally. We cannot distinguish the component 1 from the component 2 in the likelihood, because they are exchangeable. This is crucial for both Bayesian inference and computation since the maximization and exploration of the posterior surface is harder. Hence, additional precautions of the prior elicitation are required. We cannot use independent improper priors. A common way to overcome label switching is to impose an "identifiability constraint" such as ordering constraint [49], [57], [60]. Post processing MCMC using clustering like tools is another alternative in the setting of high dimension models.

Another fundamental issue in mixture modeling is the selection of the number of components. The usual tradeoff in the model selection arises in two aspects: with too many components, the model may overfit data while a mixture with too few components may not be flexible enough to approximate the true underlying models. Many authors such as Richardson & Green [57], Carline & Chib [8], Stephens [61], have proposed various methods in addressing this issue.
1.2.1 Model Selection Using Bayes Factors

Let $Y$ denote data, $M$ be a set of competing models. There is a set of parameters $\theta_m$ associated with each model $m$. Therefore, the distribution of $Y$ is specified as $f(y|\theta_m, m)$. We are interested in estimating the posterior distribution of $m$, which is defined as

$$f(m|y) = \frac{f(m)f(y|m)}{\sum_{m \in M} f(m)f(y|m)}, \quad m \in M \tag{1.1}$$

where $f(y|m)$ is the marginal likelihood calculated using

$$f(y|m) = \int f(y|m, \theta_m) f(\theta_m|m) d\theta_m \tag{1.2}$$

and $f(\theta_m|m)$ is the conditional prior distribution of $\theta_m$.

One of the most widely used model selection criterion is the Bayes factor (BF) [35], [36], [40]. The Bayes factor is defined as the ratio of posterior odds to prior odds in comparing two models $m_1$ and $m_2$.

$$BF = \frac{f(m_1|y)/f(m_2|y)}{f(m_1)/f(m_2)} = \frac{\{f(y|m_1)f(m_1)/f(y)\}/\{f(y|m_2)f(m_2)/f(y)\}}{f(m_1)/f(m_2)} \tag{1.3}$$

It is a Bayesian analogue of likelihood ratio test. It offers a way of evaluating evidence in favor of one model over another by incorporating external information into the evaluation of evidence about a hypothesis. It is a very general and flexible method, and does not require alternative models to be nested. In other words, the compared models do not have the same set of dependent variables. The computation of a
Bayes factor involves the estimation of the marginal likelihood. However, unlike the posterior distribution or the predictive distribution, which can be easily obtained from MCMC output, it is not straightforward to estimate the BF. As a result, many different approaches have been proposed.

1.2.1.1 Direct Methods

There are several direct methods which offer a convenient way to compute the Bayes factor. One is to sample $K$ values of $\theta$ from the prior distribution of $\theta$ given $m$ and approximate the likelihood:

$$\hat{p}(y|m) = \frac{1}{K} \sum_{k=1}^{K} f(y|\theta^{(k)}, m).$$

However, the conditional likelihood is more informative than the prior distribution. By simply plugging in $\theta$, many terms in the sum will be near 0. Therefore, this is not very efficient. A better approach is to estimate $\theta$ from the posterior distribution and estimate the marginal density using the harmonic mean [51].

$$\hat{p}(y|m) = \left\{ \frac{1}{K} \sum_{k=1}^{K} \frac{1}{f(y|\theta^{(k)}, m)} \right\}^{-1}$$

Although this approach is more efficient compared to the previous one, it is very unstable since some terms will still be close to 0. A compromise between sampling from the prior distribution and the posterior distribution is to define $p(\hat{\theta}) = u_1 p(\theta) + (1 - u_1) p(\theta|y)$ and $\omega(\theta) = p(\theta)/p(\hat{\theta})$. Hence, this approach can eliminate the issues discussed above. Unlike the previous approaches, which only require sampling from
either the prior or the posterior, this methodology has the added complication of requiring that are sampled from both distributions.

\[ \hat{p}(y|m) = \frac{\sum_k^K f(y|\theta^{(k)}, m) \omega(\theta^{(k)})}{\sum_k^K \omega(\theta^{(k)})} \]

### 1.2.1.2 Using Gibbs Sampler Output

Chib proposed a method to estimate the marginal likelihood using the Gibbs sampler outputs [13]. The marginal likelihood can be rewritten as the follows:

\[ p(y) = \frac{f(y|\theta)p(\theta)}{p(\theta|y)} \]

The numerator above is the product of the likelihood and the prior distribution while the denominator is the posterior distribution of the parameters. The latter can be obtained by exploiting on the complete conditional distributions from the Gibbs samplers. There is no overhead for any complex computations. The proposed estimate is

\[ \ln \hat{p}(y) = \ln f(y|\theta^*) + \ln p(\theta^*) - \ln \hat{p}(\theta^*|y) \]

The estimate does not suffer from any problems of instability. For a given number of posterior draws, the point of \( \theta^* \) can be chosen arbitrary. Chib [13] suggested that \( \theta^* \) should be chosen so that the posterior distribution has the highest density. In order to describe the process of the estimation of \( \hat{f}(\theta^*|y) \), we start with two-parameter case.

\[ \hat{f}(\theta^*_1, \theta^*_2|y) = \hat{p}(\theta^*_2|y, \theta^*_1) \hat{p}(\theta^*_1|y) \]
So the first part in the right-hand side of the equation is explicitly available in the Gibbs sampling outputs. The second part can be estimated via 'Rao-Blackwellized' estimator [13].

\[ \hat{p}(\theta^*_1 | y) = \frac{1}{K} \sum_{k=1}^{K} p(\theta^*_1 | y, \theta^{(k)}_2), \]

where \( \theta^{(k)}_2 \) is the \( k^{th} \) draw from \( p(\theta_2 | y) \) for \( k = 1, \ldots, K \). Hence, the estimate of the marginal density becomes

\[ \ln \hat{p}(y) = \ln p(y | \theta^*_1, \theta^*_2) + \ln p(\theta^*_1, \theta^*_2) - \ln p(\theta^*_2 | y, \theta^*_1) - \ln \hat{p}(\theta^*_1 | y) \]

This can be easily expanded to the general case with \( B \) parameters by factoring the joint posterior distribution into \( B \) components. The first \( B - 1 \) parameters can be obtained from the full conditional distributions while the last one can be estimated as the 'Rao-Blackwellized' estimator above. By partitioning the joint posterior distributions into several blocks, the computations of the marginal likelihood can be acquired without any computational burdens.

### 1.2.1.3 Using Metropolis-Hastings Sampler Output

Chib & Jelaizkov [15] showed that a suitable estimate of \( \hat{p}(\theta^* | y) \) can also be obtained on the basis of the Metropolis-Hastings output for sampling \( \theta^* \). The acceptance probability in the MH step for the move from \( \theta \) to \( \theta^* \) is defined as:

\[ \alpha(\theta, \theta^* | y) = \min \left\{ 1, \frac{p(\theta^* | y)q(\theta^*, \theta | y)}{p(\theta | y)q(\theta, \theta^* | y)} \right\} \]
Chib and Jeliazkov proved that

\[
p(\theta^*|y) = \frac{E_1 \{ \alpha(\theta, \theta^*|y)q(\theta, \theta^*|y) \}}{E_2 \{ \alpha(\theta^*, \theta|y) \}}.
\]

Consequently, it can be estimated through

\[
p(\theta^*|y) = \frac{\sum_{k_1=1}^{K_1} \alpha(\theta^{(k1)}, \theta^*|y)q(\theta^{(k1)}, \theta^*|y)/K_1}{\sum_{k_2=1}^{K_2} \alpha(\theta^*, \theta^{(k2)}|y)/K_2},
\]

where \(\theta^{(k1)}\) is drawn from \(p(\theta|y)\) and \(\theta^{(k2)}\) is from \(q(\theta^{(k1)}, \theta^*|y)\). The efficiency of the estimator can also be increased by dividing the parameters into blocks, which are updated separately.

Certainly using Gibbs or MH sampler outputs provides a fairly nice solution in estimating the marginal density. However, for some complicated models or high-dimensional settings, those random effect parameters can not be marginalized out of the likelihood, these methods may not be feasible. Furthermore, Bayes factor via marginal density estimations tends to be very sensitivity with respect to prior distributions. If the prior is not proper, especially in the case of using non-informative prior, the Bayes factor is not well-defined. Several authors have attempted to modify BF to overcome the deficiency such as partial Bayes factor and fractional Bayes factor [53], [54].

Green concludes that the ideal situation for this approach, and generally for the within-model strategies, is when the competing models are very heterogeneous and fully-tested samplers with acceptable performance are already available for each of
them. Han & Carlin [32] considered the method relatively easy to program and tune; it needs a fair amount of 'bookkeeping' but requires little in the way of additional coding beyond what is already required for the parameter estimation of the individual models themselves.

1.2.2 Reversible Jump Markov Chain Monte Carlo

The reversible jump Markov chain Monte Carlo (RJMCMC) proposed by Green [30] is able to account the model uncertainty by imposing a probability distribution on both the set of possible models and the set of parameters. In the setting of mixture models, it allows posterior simulations to jump between the parameter subspaces corresponding to different numbers of components in the mixture. The essence of this method is to incorporate two additional features to Metropolis-Hastings schemas. First of all, additional auxiliary variables are generated to augment parameters in the lower dimensional model to match dimensions across the current and proposed model. Next, the proposal is allowed to include a deterministic mapping to map the augmented vector of parameters and auxiliary variables to a proposed parameter space under the larger model. For any pair of models \((m, m')\), there exists a bijection \((\theta, \theta') = g(\theta, \theta')\) from \(S = \{(\theta, u')\}\) to \(S' = \{(\theta', u)\}\), where \(S\) and \(S'\) have the same dimension. At the current state \((m, \theta)\), a new state, \((m', \theta')\) is proposed by generating \(u'\) from a suitable proposal distribution \(q(u'|\theta)\) with the acceptance probability of \(\alpha\):

\[
\alpha(\theta, u') = \min\{1, A(\theta, u')\}.
\]
\[ A(\theta, u') = \frac{p(y, \theta'|m')p(m')q(m|m')q(u'|\theta')}{p(y, \theta|m)p(m)q(m'|m)q(u'|\theta)} J(\theta, u'), \]

where \( q(m'|m) \) is the probability of proposing model \( m' \) when the current model is \( m \)
and \( J \) is the Jacobian transformation from \((\theta, u')\) to \((\theta')\). After a number of iterations, \( N \), \( p(m|y) \) is estimated as

\[ \hat{p}(m|y) = \frac{n}{N}, \]

where \( n \) is the number of times the Markov chain visited model \( m \). The BF is reflected in the acceptance probability, i.e.:

\[ BF = \frac{p(y, \theta'|m')p(m')q(m|m')}{p(y, \theta|m)p(m)q(m'|m)} \]

One of the main advantages of the RJMCMC, compared to the BF above, is that it offers a single logical and computational framework for the joint inference about the models and the parameters. Potentially the jump between different models can improve mixing. However, it is not straightforward to construct and difficult to tune. When a model is much more likely than the others, the algorithm gets easily stuck in those models, and never visits the other models. We can improve the mixing by changing the priors of the models. A detailed discussion about the RJMCMC is available in the Chapter 2.

### 1.2.3 Pseudoprior (PP) and 'Metropolized' PP Method

In 1995, Carlin and Chib presented a Gibbs sampling method (pseudoprior: PP) which allows the model to search over both the model space and the parameter space.
Corresponding to each model space $m$, the likelihood and the prior are written as $f(y|\theta_m, M = m)$ and $f(\theta_m|M = m)$. Additionally, a set of pseudopriors $f(\theta_m|M \neq m)$ are chosen to complete the Bayesian model specification. Since $M$ is just an indicator of the model, given $M = m$, $Y$ is independent of $\{\theta(-m)\}$, where $(-m)$ denotes a set of models excluding model $m$. Therefore, the conditional independence assumption implies that

$$p(y|M = m) = \int f(y|\theta, M = m)f(\theta|M = m)d\theta$$

$$= \int f(y|\theta_m, M = m)f(\theta_m|M = m)d\theta_m$$

A pseudoprior is not really a prior at all, but a link density for the purpose of convenience. Hence, the joint distribution of $y$ and $\theta$ given $M = m$ is

$$p(y, \theta, M = m) = f(y|\theta_m, M = m) \left\{ \prod_{i=1}^{G} f(\theta_i|M = m) \right\} f(m)$$

Consequently, the full conditional distributions are

$$p(\theta_m|y, \theta(-m), M = m) \propto \begin{cases} f(y|\theta_m, m)p(\theta_m|m) & M = m \\ p(\theta_m|M \neq m) & M \neq m \end{cases}$$

$$p(M = m|\theta, y) = \frac{f(y|\theta_m, M = m) \left\{ \prod_{i=1}^{G} f(\theta_i|M = m) \right\} f(m)}{\sum_{k=1}^{G} f(y|\theta_k, M = k) \left\{ \prod_{i=1}^{G} f(\theta_i|M = m) \right\} f(m)}$$

The model indicator $m$ is generated as a discrete random variable. Therefore, the probability of favoring model $m$ can be estimated from the following ratio

$$\hat{p}(M = m|y) = \frac{\text{the number of } M = m}{\text{the total number of } M}.$$
the sampler must draw all the parameters of all the models, not only those of the one currently being visited by the chain. However, Carlin and Chib mentioned that pseudopriors do not contribute to the marginal likelihood, a simpler form can be chosen to optimize the Gibbs sampling.

A 'Metropolized' PP (MPP) method was presented by Dellaportas, Forster, and Ntzoufras [19], which is a special case of RJMCMC. The model selection is built not based on the full conditional distribution, but a hybrid Gibbs/Metropolis approach, which proposes a move from model $m$ to model $m'$ followed by the acceptance or the rejection of the proposal. The deterministic function $q(u\ldots)$ in RJMCMC are replaced by pseudopriors here. First propose a move from $m$ to $m'$ with the probability $j(m,m')$. Generate $\theta_{m'}$ from the pseudoprior $f(\theta_{m'}|m \neq m')$ and accept the proposed move with the probability of $\alpha = \min(1,A)$, where

$$A = \frac{f(y|\theta_{m'}, m') f(\theta_{m'}|m') f(\theta_{m}|m') f(m') j(m', m)}{f(y|\theta_{m}, m) f(\theta_{m}|m) f(\theta_{m'}|m) f(m) j(m, m')}.$$ 

Since only models $m$ and $m'$ are involved here, therefore, only values of $\theta_m$ and $\theta_{m'}$ are required. The pseudopriors for other $\theta$s are cancelled out. The method consists of the following steps:

This 'Metropolized' method is more efficient with respect to the number of pseudopriors. As we can see here, a simple modification of PP method makes it equivalent to the reversible jump MCMC. The pseudoprior becomes part of the proposal function in the Metropolis-Hasting step. Compared to the Bayes factor via the marginal density
distribution, both the MPP and the RJMCMC have the issue of slower convergence due to the difficulty in finding the pseudoprior or the proposal density. Some random parameters can be marginalized out of the likelihood to speed up the convergence.

### 1.2.4 Birth-and-Death Markov Chain Monte Carlo

In the setting of mixture models, the birth-and-death MCMC (BDMCMC), presented by Stephens in 2000, views each component as a point in the parameter space, and adopts theory from the simulation of point process to construct a Markov chain with the posterior distribution of the parameters as its stationary distribution. Unlike the RJMCMC, it allows a continuous time jump process in different dimensions. The jump process is associated with a birth rate and a death rate and the moves are always accepted. The holding time corresponds to the acceptance probability in RJMCMC move. With a little change of notations here, let $\theta^-$ and $\theta^+$ be the parameter spaces for the lower- and higher-dimension, respectively. A new component is created in continuous time at the rate of $\beta(\theta^-)$. Accordingly, the weight $p$ and the parameter are drawn from a joint density $f(p, \theta^-)$. The weights of old components are scaled down proportionally to make all of the weights sum up to unity, $p_g := p_g/(1 + p)$ for $g = 1, \ldots, G$. The component $g$ is killed at the rate of

$$
\delta_g(\theta^+) = \frac{f(y^{\theta^-}_(-g))f(\theta^+_(-g))}{f(y^{\theta^+}_g)f(\theta^+_g)} \times \frac{1}{G + 1} \times \frac{\beta(\theta^-)f(p, \theta^+_(-g))}{(1 - p_g)^{G-1}},
$$

where $\theta^+_(-g)$ represents the parameters in the higher dimension excluding the elements in the component $g$. Therefore, the total death rate is $\delta(\theta^+) = \sum_{g=1}^G \delta_g(\theta^+)$. The time
to the next jump follows an exponential distribution with the mean of \(1/(\beta(\theta^-) + \delta(\theta^+))\).

Comparing with the RJMCMC, the BDMCMC differs in the following major aspects: 1) the BDMCMC operates in a continuous time space, replacing the accept-rejection schema by allowing events to occur at different rates. However, with respect to the computation time, this replacement does not make a big difference. Sometimes, it might be slower than the RJMCMC since it involves the considerations of the whole range of possible moves and their respective rates after each move. 2) the BDMCMC effectively integrates out the missing data used in RJMCMC when calculating the likelihood. However, this creates problems for the model where calculations of the likelihood require the knowledge of missing data such as hidden Markov models. This issue can be solved by introducing missing data into the MCMC schema and keeping them fixed. However, since the missing data is highly informative for the number of mixtures, it most likely leads to poor mixing. 3) the BDMCMC takes advantage of the natural structure of the models, removing the need to calculate the Jacobian, and making implementation more straightforward; 4) the BDMCMC treats the parameters as a point process, hence, relaxes the constraint on the order of the location parameters in the examples such as mixture normal distributions.
Chapter 2

A Mixture Model for Modeling

Ordinal Longitudinal Data

2.1 Introduction

Ordinal data are categorical data where there is a natural order to the categories. It is frequently encountered in different fields such as biostatistics, social science, and economics. A good example is the smoking status: non-smokers, ex-smokers, light smokers, and heavy smokers. Ordinal data are often analyzed as continuous outcomes or reduced to two categories to simplify analyses and presentations, which may result in a considerable loss of information.

The most natural way to view ordinal data is to postulate the existence of an underlying latent variable associated with each response. Such variables are often assumed
to be drawn from continuous distributions centered at a mean value. A vector of
cutoff points is imposed to categorize latent variables into the corresponding ordinal
groups. In general, the ordinal regression models have a general form such as

\[ p(Y_i = k) = F(X^T \beta), \quad k = 1, \ldots, K \]

\( F \) is called the link function. McCullag [44] proposed the proportional odds or haz-
ARDS models which use a logit or complementary log-log function as the link function
to model data. Many other functions such as the inverse normal distribution and
the Cauchy distribution are the possible functions for the link function. The \( \beta \)s de-
scribe how covariates are correlated with the proposed latent quantities. Many other
authors such as Albert & Chib [2], Cowles et al [17], Newton et al [50], Erkanli et
al [23] illustrated the fitting of ordinal regression models using latent variables in a
with pre-specified mixing locations.

One interesting issue with ordinal data regression is how to estimate the cutpoint pa-
rameters which ties continuous latent quantities and ordinal outcomes. Albert & Chib
[2] proposed Bayesian inference for the ordinal probit model using the Gibbs sampler
method. A vague prior has been imposed on the cutpoint parameters. However, the
convergence of MCMC is delayed due to the strong autocorrelations among param-
eters when the sample size is big. Cowles [16] proposed a hybrid Gibbs/Metropolis-
Hastings sampling scheme that would update the cutpoint parameters in multivariate
distributions with the latent quantities. However, this model is not applicable in the context of ordinal longitudinal data. The details and the limitations of each method will be discussed later. We propose a mixture model which can successfully model the ordinal property of data while keeping cutpoint parameters constant without loss of generality. This model can reduce the number of parameters needed to be estimated. Furthermore, it can also eliminate the issue of convergence of MCMC. The rest of the chapter is organized as the follows. In the Section 2.2, we define an autoregressive (1) error model to capture the correlations of the longitudinal data. A probit function is proposed to link ordinal measurements and continuous latent quantities. Albert & Chib method and Cowles’ method are used to estimate the cutpoint parameters (Sections 2.3 & 2.4). Finally, we introduce our mixture model. We first fix the size of mixture in Section 2.5 and further treat it as a random variable using reversible jump Markov chain Monte Carlo (RJMCMC) algorithm (Section 2.6). Two simulated datasets and a motivational intervention study are presented in Section 2.8 to illustrate and compare each method.

2.2 Model and Prior Specifications

2.2.1 Model Specifications

Let $i = 1, \ldots, n$ be the index for the $i^{th}$ subject and $j = 1, \ldots, J$ be the index for $j^{th}$ time point. The random variable $y_{ij}$ is the ordinal response for $i^{th}$ subject at
time $j$, which takes on the value of $k = 1, \ldots, K$. Associated with each subject $Y_i = (y_{i1}, \ldots, y_{ij})$, there is a vector of latent variables $Z_i = (z_{i1}, \ldots, z_{ij})$ and a set of cutpoint parameters $\theta = (\theta_0, \ldots, \theta_K)$. Let $X_i^T = (x_{i1}, x_{i2}, \ldots, x_{ih})$ be the covariates for each subject $i$. Hence, the probability of $y_{ij} = k$ is modeled through the estimation of the probability of $z_{ij}$ falling into the interval of $[\theta_{k-1}, \theta_k)$, i.e.,

$$y_{ij} = k \quad \text{if} \quad \theta_{k-1} \leq z_{ij} < \theta_k \quad \text{for} \quad k = 1, \ldots, K.$$ 

$$p(y_{ij} = k) = p(\theta_{k-1} \leq z_{ij} < \theta_k),$$

$$-\infty = \theta_0 < \ldots < \theta_K = \infty$$

Therefore, the order of the categories are preserved in the $F$-link for the cumulative probabilities. For the presentation purpose, denote $Y = (Y_1, \ldots, Y_n)$ and $Z = (Z_1, \ldots, Z_n)$.

For different subjects, $Y$s are independent with each other. However, the observations within each subject may be subject to correlations over time. Model of such dataset can be considered as a special case of generalized linear models, with the particular feature that the residual terms are correlated, as the observations at different time points in a longitudinal study are taken on the same subject. Therefore, the autoregressive (1) error (AR(1) error) model is proposed to model the correlations within each subject for the latent variables.

$$z_{ij} = X_i^T \beta + \rho (z_{ij-1} - X_i^T \beta) + \varepsilon_{ij},$$
where,

\[ z_{i0} \sim N(0, \sigma_0^2), \]

\[ \varepsilon_{ij} \sim \sum_{g=1}^{G} p_g N(\mu_g, \sigma^2), \quad \sum p_g = 1 \]

Instead of assuming the latent variables are drawn from a single normal distribution, we propose a mixture of normal distributions. Therefore, the residuals \( \varepsilon_{ij} \)s follow a mixture of \( G \) normal distributions with the location parameter of \( \mu_g \) and the weight of \( p_g \) for each component. Let \( P = (p_1, \ldots, p_G) \). For the purpose of identifiability and consistence with the cumulative distribution of standard normal function, \( \sigma \) is set to be 1 [2]. Alternatively, the mixture model can be represented as a hierarchical structure by introducing another latent variable \( w_{ij} \), which represents the membership of the mixture model. The variable \( W = (w_{11}, \ldots, w_{1J}, \ldots, w_{nJ}) \) take on the value from 1 to \( G \) with the probability of \( p_g \), the weight of each component.

\[ w_{ij} = \begin{cases} 
  g & \text{if } \varepsilon_{ij} \sim N(\mu_g, \sigma^2) \\
  0 & \text{otherwise} 
\end{cases} \]

\[ p(w_{ij} = g) = p_g \]

Therefore, the model can be written as:

\[ \varepsilon_{ij} \sim N(\mu_g, 1), \quad \text{if } w_{ij} = g \]

Thus,

\[ p(y_{ij} = k|w_{ij} = g) = \]

\[ F \left( \theta_k - (X_i^T\beta + \rho(z_{ij-1} - X_i^T\beta) + \mu_g) \right) - F \left( \theta_{k-1} - (X_i^T\beta + \rho(z_{ij-1} - X_i^T\beta) + \mu_g) \right) \]
The likelihood function is written as the follows:

\[
L = \prod_{i=1}^{n} \prod_{j=1}^{J} \prod_{k=1}^{K} \prod_{g=1}^{G} \{F(\theta_k - m_{ij}) - F((\theta_k-1) - m_{ij}))\} I(w_{ij} = g) I(y_{ij} = k) \tag{2.1}
\]

where,

\[
m_{ij} = X_i^T \beta + \rho(z_{ij} - X_i^T \beta) + \mu_g,
\]

\(I(\cdot)\) is the index function.

### 2.2.2 Prior Specifications

Let \(B, DD, ID, NB, N, MVN,\) and \(U\) denote Beta distribution, Dirichlet distribution, identity matrix, negative binomial distribution, normal distribution, multivariate normal distribution, and uniform distribution, respectively. The prior distribution of each parameter is listed below: We assume that \(G\) follows a negative binomial distribution with the mean of \(a(1-b)/b\), where \(a\) is the maximum size of mixtures, \(G_{max}\). The NB prior renders higher probabilities on the smaller value of \(G\) as opposed to assigning a uniform probability on all possible values between 1 and \(G_{max}\). Alternatively, a Poisson distribution is another potential prior distribution for \(G\). A uniform Dirichlet distribution is used as the prior distribution for the weight of each component \((p_g)\) since it follows a multinomial distribution. The Dirichlet prior is a
conjugate prior for the multinomial distributions.

\[ p(G) \sim NB(a, b) \]
\[ p(p_1, p_2, \ldots, p_G | G) \sim D(\gamma, \ldots, \gamma) \]
\[ p(w_{ij} | G, P) \sim I(w_{ij} = g) \]

Diffuse priors and non-informative priors are imposed on \( \beta \) and \( \rho \), respectively. We constrained the autoregressive coefficient \( \rho \) to be between 0 and 1 to ensure the stationarity of the process. However, the restriction on the prior assumption is not necessary in the Bayesian regression. Chib [12] applies the AR(p) model in the error term conditioning on the initial observations and without necessarily assuming the stationarity in the AR processes \textit{a priori}. Different assumptions may be adopted regarding the latent data \( z_{i0} \) [64]. Bayesian analysis allows the latent data to be modeled as extra parameters. Therefore, \( z_{i0} \) is assumed to follow a normal distribution.

\[ f(\beta) \sim MVN(0, \Sigma) \quad \text{where,} \quad \Sigma = 10^6 ID_{h \times h}, \]
\[ f(\rho) \sim U(0, 1) \]
\[ f(z_{i0}) \sim N(0, 1) \]

It is inappropriate to assign a fully non-informative on the location parameters of the mixture in the context of mixture modeling. Furthermore, there is always the possibility that no observations are allocated to one or more components, therefore, data is uninformative about it [59]. Hence, a weakly informative prior is imposed on the location parameter of mixtures. We set \( \mu \) to be the midpoint of the range of the latent variable \( H \) and \( \sigma^2_\mu \) be proportional to \( H^2 \). Moreover, special cautions
should also be taken on the label-switch in the mixture models. Therefore, an order constraint is put on $\mu$s, hence, $\mu_1 < \mu_2 < \ldots < \mu_G$. The details are discussed later.

$$f(\mu_g|G) \sim N(\nu, \sigma_{\mu}^2)$$

As for the cutpoint parameters, in situations where little prior information is available, the simplest way to construct a prior distribution over them is to fix the value of one cutpoint parameter, usually $\theta_1$ at 0 and assume the uniform prior for the remaining cutpoint parameters. However, as we stress several times that the cutpoint parameters can be fixed without loss of generality by fitting the latent variables with a mixture of normal distributions. Therefore, the joint distribution of all variables can be written in general as

$$p(\beta, \rho, \mu, W, G, P, Z, Y)$$

$$= p(G)p(P|G)p(W|G, P)f(\mu|G)f(\beta)f(\rho)f(Z|\beta, \rho, \mu, P, W)p(Y|Z)$$

The complete model is displayed as a directed acyclic graph (DAG) in the Figure 2.1.

### 2.3 Vague Prior Algorithm by Albert & Chib

Gibbs sampling is one of the most widely used approaches in approximating the posterior density. Parameters of interest are iteratively generated from the full conditional distributions. Albert & Chib proposed Bayesian implementations of the ordinal probit model using Gibbs sampler method by imposing a vague prior on the cutpoint parameters. We assume that the error term $\varepsilon_{ij}$ follows a single normal distribution
with the mean of $\mu$ and the variance of 1 for the time being.

**Conditional Distribution of $z_{i0}$:**

$$f(z_{i0}|z_{i1}, \beta, \rho) \sim f(z_{i0})f(z_{i1}|\beta, \rho, y_{i1})$$

$$\sim N\left(\frac{(\rho z_{i1} - (1-\rho)X_i^T \beta)}{\rho^2 + 1/\sigma^2_0}, (\rho^2 + 1/\sigma^2_0)^{-1}\right)$$

**Conditional Distribution of $z_{ij}, j=1, \ldots, J-1$:**

$$f(z_{ij}|z_{i0}, z_{i(-j)}, \beta, \rho, y_{ij}) \sim f(z_{ij}|z_{i(j-1)}, \rho, \beta, \theta)f(z_{i(j+1)}|z_{ij}, \rho, \beta, \theta)I(y_{ij} = k)$$

$$\sim N\left(\frac{(\rho z_{i(j+1)} + z_{i(j-1)} - 2X_i^T \beta)}{1+\rho^2}, \frac{1}{1+\rho^2}\right)I(y_{ij} = k)$$

**Conditional Distribution of $z_{iJ}$:**

$$f(z_{iJ}|z_{i(-J)}, \beta, \rho, y_{iJ}) \sim N(\rho(z_{i(J-1)}) - X_i^T \beta) + X_i^T \beta, 1)I(y_{iJ} = k),$$

**Conditional Distribution of $\beta$:**

$$f(\beta|\rho, Z, Y) \sim N((X^T V^{-1} X + \Sigma^{-1})^{-1}(X^T V^{-1} Z^*), (X^T V^{-1} X + \Sigma^{-1}))$$

$$V = \frac{1}{(1-\rho)^2}ID_{h \times h}, \quad \Sigma = 10^6 ID_{h \times h}, \quad z^*_{ij} = \frac{z_{ij} - \rho z_{i(j-1)}}{1-\rho}$$

**Conditional Distribution of $\rho$:**

$$f(\rho|\beta, Z, Y) \sim f(Z|\beta, \rho, \theta)f(\rho)$$

$$\sim N\left(\frac{\sum_{i=1}^I \sum_{j=1}^J z^*_{ij} z^*_{ij-1} - \sum_{i=1}^I \sum_{j=1}^J z^*_{ij-1}^2}{\sum_{i=1}^I \sum_{j=1}^J \sum_{i=1}^I z^*_{ij-1}^2}, (\sum_{i=1}^I \sum_{j=1}^J z^*_{ij-1}^2)^{-1}\right) I(0 < \rho < 1)$$

$$z^*_{ij} = z_{ij} - X_i^T \beta, \quad z^*_{i(j-1)} = z_{i(j-1)} - X_i^T \beta$$
Conditional Distribution of $\theta$:

$$f(\theta_k|\theta^{(k)}, \beta, \rho, Z, Y) \sim U(\max_{ij}\{z_{ij} : y_{ij} = k\}, \min_{ij}\{z_{ij} : y_{ij} = k+1\})$$

$$\theta^{(k)} = (\theta_1, \theta_2, \ldots, \theta_{k-1}, \theta_{k+1}, \ldots)$$

In order to make the domain of $Z$ be the entire real line, $\theta_0$ and $\theta_K$ are fixed at $-\infty$ and $\infty$ respectively. The model with $K$ cutpoint parameters is overparameterized. One can add a constant to each cutpoint and subtract the same constant of the intercept term from the regression structure while the model remains the same. Therefore, one cutpoint parameter has to be fixed for the purpose of identifiability. Without the loss of generality, we fix $\theta_1$ at 0. An alternative to fix $\theta_1$ would be to omit the intercept from the model. The convergence of the Gibbs sampler implemented by simulating from the above full conditional distributions appears to depend on the size of the sample. The convergence is very slow when the sample size is big because intervals between the cutpoint parameters generated from the full conditional distributions are very narrow. Therefore, the cutpoint values change very little between successive iterations. Hence, the convergence of other parameters are retarded as well. In some cases, the Lag-1 autocorrelation could be as high as 0.99 for the cutpoint parameters, indicating extremely slow mixing of the sample path of parameters. Therefore, in the extremely high-dimensional ordinal probit models, it is not practical to apply a single univariate generation algorithm to the joint posterior distribution of the latent data, regression coefficients, and cutpoints.
2.4 Hybrid of Gibbs/Metropolis-Hastings

Cowles [16] proposed a hybrid Gibbs/Metropolis-Hastings sampling scheme algorithm that would update the cutpoint parameters in the multivariate distributions with the latent quantities. This approach reduces the high auto-correlation and reaches the convergence within a reasonable small number of iterations. Compared to the vague prior, it has the following advantages. First of all, it is relatively easy to implement. Second, it displays good mixing. Furthermore, it can extend to models with arbitrary constraints on the category cutoff points. The algorithm partitions the parameter space into $(Z, \theta)$ and $(\beta, \rho)$. The joint full conditional distribution of $Z$ and $\theta$ can be further partitioned into the followings:

$$f(Z, \theta | \beta, \rho, Y) = f(Z|\theta, \beta, \rho, Y)f(\theta|\beta, \rho, Y)$$

Therefore, $Z$ can be sampled from the conditional distribution and $\theta$ is drawn iteratively using Metropolis-Hasting algorithm. Let $\theta^*$ be the proposed $\theta$. Sample $\theta^*_k$ from a $N(\theta_k, \sigma^2_{MH})$ truncated to the interval of $(\theta^*_{k-1}, \theta_{k+1})$. Due to the correlations within $Z_i$s, the AR(1) error model of $Z$ is re-written as the follows in order to marginalize latent quantities in the above joint full distribution.

$$Z = (1 - \rho)X^T\beta + RZ + \Sigma$$

$$X^T\beta = \left(\begin{array}{c} X_1^T\beta, \ldots, X_1^T\beta, \ldots, X_n^T\beta, \ldots, X_n^T\beta \end{array}\right)^T$$

$$Z = \mu^*_Z + RZ + \Sigma$$
where,

\[ Z = (z_{i0}, z_{i1}, z_{i2}, \ldots, z_{in}, z_{nJ})^T \]

\[ \mu^*_Z = (0, (1 - \rho)X_1^T \beta, \ldots, (1 - \rho)X_{i1}^T \beta, \ldots, 0, (1 - \rho)X_n^T \beta, \ldots, (1 - \rho)X_n^T \beta)^T \]

and R is a block-diagonal matrix with \( N \times (J + 1) \) rows and columns respectively.

\[
R = \begin{pmatrix}
R^* & 0 & \ldots & 0 \\
0 & R^* & 0 & \ldots \\
\vdots & \vdots & \ddots & \vdots \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \ldots & R^*
\end{pmatrix}
\]

where, \( R^* \) is denoted as the following matrix with the dimension of \( (J + 1) \times (J + 1) \).

\[
R^* = \begin{pmatrix}
0 & \ldots & \ldots & \ldots & 0 \\
\rho & 0 & \vdots & \vdots & 0 \\
0 & \rho & \ddots & \vdots & \vdots \\
\vdots & \ddots & \ddots & \ddots & \vdots \\
0 & \ldots & \ldots & \rho & 0
\end{pmatrix}
\]

\( \Sigma \) is a diagonal matrix with the dimension of \( N \times (J + 1) \) rows and columns. The diagonal elements are

\[
\text{diag}(\Sigma) = \left( \sigma_0^2, \frac{\sigma^2}{1 - \rho^2}, \ldots, \frac{\sigma^2}{1 - \rho^2}, \ldots, \sigma_0^2, \frac{\sigma^2}{1 - \rho^2}, \ldots, \frac{\sigma^2}{1 - \rho^2} \right)
\]
Therefore, $Z$ follows a $MVN(\mu_Z, \Sigma_Z)$, where $\mu_Z = (I - R)^{-1} \mu^*_Z$ and $\Sigma_Z = (I - R)^{-1} \Sigma (I - R)^{-T}$. Likewise, $\Sigma_Z$ is a block-diagonal matrix with $N \ast (J + 1)$ rows and columns respectively and $\Sigma^*_Z$ as a block.

$$
\Sigma^*_Z = \begin{pmatrix}
1 & \rho & \rho^2 & \ldots & \ldots & \rho^{(J-1)} \\
\rho & \sum_{i=0}^{1} \rho^{2i} & \rho \sum_{i=0}^{1} \rho^{2i} & \ldots & \ldots & \rho^{(J-2)} \sum_{i=0}^{1} \rho^{2i} \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
\rho^{(J-1)} & \rho^{(J-2)} \sum_{i=0}^{1} \rho^{2i} & \ldots & \rho \sum_{i=0}^{(J-2)} \rho^{2i} & \ldots & \rho^{(J-1)} \sum_{i=0}^{(J-1)} \rho^{2i}
\end{pmatrix}
$$

With this candidate-generating density, the acceptance probability $\alpha$ is equal to $\min(1, A)$, where

$$
A = \prod_{i=1}^{n} \frac{p(\theta^*_{yi1-1} \leq z_{i1} < \theta^*_{yi1}, \ldots, \theta^*_{yiJ-1} \leq z_{iJ} < \theta^*_{yiJ})}{p(\theta^*_{yi1-1} \leq z_{i1} < \theta^*_{yi1}, \ldots, \theta^*_{yiJ-1} \leq z_{iJ} < \theta^*_{yiJ})} \\
\times \prod_{k=2}^{K} \frac{\Phi((\theta^*_{k+1} - \theta^*_k)/\sigma_{MH}) - \Phi((\theta^*_{k+1} - \theta^*_k)/\sigma_{MH})}{\Phi((\theta^*_{k+1} - \theta^*_k)/\sigma_{MH}) - \Phi((\theta^*_{k+1} - \theta^*_k)/\sigma_{MH})}.
$$

First of all, generate candidate values of $\theta$ from truncated normal distributions. Next evaluate the quantity of $A$ to get the acceptance probability $\alpha$. With the probability of $\alpha$, set $\theta$ to the new $\theta^*$. The other variables, $Z$, $\beta$, and $\rho$ are generated from their full conditional distributions. Notice that the computation of the cumulative probability of a multivariate normal distribution is required in order to calculate the acceptance probability. However, given correlations existing among $z_{i1}, \ldots, z_{iJ}$, the
computation becomes intractable. Chin & Dey [11] proved that the computation of the cumulative probability of a multivariate normal distribution could be avoided by computing the conditional distribution of each variate, which follows a univariate normal distribution. However, they assumed various cutpoint parameters for each correlated response variable. In our setting, all correlated response variables \((y_{ij})\) are associated with the latent quantities \((z_{ij})\) through the same set of cutpoints. Hence, it is not straightforward to compute the acceptance probability under the assumption of AR (1) error model, not to mention other higher-order of autoregressive models, or repeated measurements with other correlation structures.

2.5 Fixed Size Mixture Normal Distributions

Given the intractability of Cowles' method in the repeated measurement, an alternative approach is presented. Assume that the latent quantities \(Z\) follow a fixed size of mixture of normal distributions and keep cutpoint parameters constant. We believe this mixture model with the random location parameters and fixed cutpoint parameters is capable to estimate the probability of being in ordinal level \(k\). Due to the complexity of the integration of the probit model and AR (1) error model, the proof will be constructed disregarding the autoregressive structure. In other words, autoregressive coefficient is assumed to be 0 for the purpose of the proof.

We show by a constructive argument that an appropriate choice of \((p_g, \mu_g, g =
1, \ldots, G) can approximate an arbitrary set of desired cell probabilities \((\pi_1^*, \ldots, \pi_K^*)\).

A similar argument was used in Kottas et al. [39] for infinite Dirichlet process mixtures of normal probability.

Consider a mixture of normal distributions with \(G \geq K\) components. Place one component of the mixture into each interval \([\theta_{k-1}, \theta_k]\) by choosing \(\mu_k = \frac{1}{2}(\theta_k + \theta_{k-1})\), and set \(p_k = \pi_k^*\). Specify \(\sigma\), the variance of the latent normal kernels, such that \(1 - \epsilon\) of the probabilities of each kernel is between the adjacent cutoff points. This trivially achieves \(|\pi_k - \pi_k^*| < \epsilon\) for \(k = 1, \ldots, K\). Therefore, the cutoff points \(\theta\) can be fixed without loss of generality.

We recommend as a default choice of cutpoint parameters \((\theta_k)\): \(\theta_0 = -\infty, \theta_K = \infty, \{\theta_1, \ldots, \theta_{K-1}\} = \{0, \pm 4, \pm 8, \ldots\}\). For reasons of identifiability, we suggest fixing \(\sigma = 1\). This choice implicitly restricts cell probabilities \(\pi_2, \ldots, \pi_{K-1}\) to be at most 0.95. If larger or smaller cell probabilities are desired, the widths between \(\theta_{k-1}\) and \(\theta_k\) can be increased or decreased accordingly. In conjunction with the choice of \(\theta_s\), the choice of the size of mixture, \(G\), can be approached by trial and error. Alternatively, some model selection criteria such as reversible jump MCMC can be used to identify \(G\). Therefore, the cutpoint parameters \(\theta\) can be fixed without loss of generality by assuming the latent variable of ordinal outcomes following a mixture of normal distributions. The same proposition holds when \(\rho\) becomes non-zero, i.e., the existence of correlations within subjects.
Mixture models can be thought of as missing data structure. A variety of research has been done in a Bayesian set-up [63], [41], [21], [24]. As stated before, $p_g$ follows a multinomial distribution. The prior distributions for other parameters are specified in Section 2. The full conditional distributions for $p, \mu, W$ are derived as the followings:

**Conditional Distribution of $p_g, g = 1, \ldots, G$:**

$$p(p_g|W, Z, Y) \sim D(\gamma + n_1, \ldots, \gamma + n_G),$$

$$n_g = \sum_{i,j} I(w_{ij} = g), \quad \text{for} \quad g = 1, \ldots, G$$

**Conditional Distribution of $w_{ij}$:**

$$p(w_{ij}|p, \beta, \mu, G, p, Z) = p_{ijg} I(w_{ij} = g),$$

$$p_{ijg} = \frac{p_g f(\varepsilon_{ij}|\mu_g)}{\sum_{g=1}^{G} p_g f(\varepsilon_{ij}|\mu_g)}, \quad \varepsilon_{ij} = z_{ij} - X_i^T \beta - \rho(z_{ij-1} - X_i^T \beta)$$

Generate a random number $u$ from $U(0, 1)$ and assign $w_{ij}$ based on $u$ with the posterior probability of $w_{ij}$.

$$w_{ij} = \begin{cases} 
1 & \text{if} \quad 0 \leq u < p_{ij1} \\
2 & \text{if} \quad p_{ij1} \leq u < p_{ij2} + p_{ij1} \\
\vdots & \vdots \\
G & \text{if} \quad p_{ijG-1} \leq u < 1
\end{cases}$$

**Conditional Distribution of $\mu_g$:**

$$f(\mu_g|p, \rho, \beta, W, Z, Y) \sim N\left(\frac{\sum_{i,j} I(w_{ij} = g) + \mu / \sigma^2_{\mu}}{n_g + 1 / \sigma^2_{\mu}}, \quad (n_g + 1 / (\sigma^2_{\mu}))^{-1}\right)$$

$$\varepsilon_{ij} = z_{ij} - X_i^T \beta - \rho(z_{ij-1} - X_i^T \beta)$$
The likelihood function of the mixture model is invariant to permutation of the labels $g = 1, 2, \ldots, G$. In the case of making inference of the individual element of the mixture such as the cluster analysis, it is important to have a unique label. Although the label switching will not affect the inference of the non-element specific parameters such as $\beta$s and $\rho$ in our model. For the purpose of identifiability, $\mu_g$s are constrained to the following order: $\mu_1 < \mu_2 < \ldots < \mu_G$. Therefore, the samplings of some $\mu$s are indeed sampled from truncated normal distributions. Accordingly, the full conditional distributions of $Z$, $\beta$, $\rho$ are revised as the followings.

$$f(z_{i0} | \ldots) \sim N \left( \frac{(\rho(z_{i1}-(1-\rho)X_i^T \beta) - \mu_g)}{\rho^2 + 1/\sigma_0^2}, (\rho^2 + 1/\sigma_0^2)^{-1} \right)$$

$$f(z_{ij} | \ldots) \sim N \left( \frac{X_i^T \beta + \rho(z_{ij-1} - X_i^T \beta) + \mu_g + \rho(z_{ij+1}-(1-\rho)X_i^T \beta - \mu_{g+1})}{\rho^2 + 1}, \frac{1}{\rho^2 + 1} \right) I(y_{ij} = k)$$

for $j = 1, \ldots, J - 1$

$$f(z_{iJ} | \ldots) \sim N \left( X_i^T \beta + \rho(z_{iJ-1} - X_i^T \beta) + \mu_g, 1 \right) I(y_{iJ} = k)$$

$$f(\beta | \ldots) \sim N \left( (X^TV^{-1}X + \Sigma^{-1})^{-1}(X^TV^{-1}Z^*), (X^TV^{-1}X + \Sigma^{-1}) \right)$$

$$V = \frac{1}{1-\rho}ID_{h \times h}, \quad \Sigma = 10^6 ID_{h \times h}, \quad z_{ij}^* = \frac{z_{ij} - \rho z_{ij-1} - \mu_g}{1-\rho}$$

$$f(\rho | \ldots) \sim N \left( \frac{\sum_{i=1}^I \sum_{j=1}^J z_{ij}^* z_{ij-1}^2}{\sum_{i=1}^I \sum_{j=1}^J z_{ij-1}^2}, \left( \sum_{i=1}^I \sum_{j=1}^J z_{ij-1}^2 \right)^{-1} \right) I(0 < \rho < 1)$$

$$z_{ij}^* = z_{ij} - X_i^T \beta - \mu_g, \quad z_{ij-1}^* = z_{ij-1} - X_i^T \beta$$

The mixture model provides a conceptually simple alternative to estimate the cell probabilities. The size of mixture is pre-determined given no prior information. However, a smaller size of mixture of normal distributions may be not sufficient enough to model cell probabilities. A large size of mixture may result in overfitting data.
Therefore, a more intelligent way in identifying the size of mixture is desirable.

2.6 Random Size Mixture Normal Distributions

The selection of unknown number of components may often be involved with the search over the parameter and the model space. There are many methods such as pseudoprior [8] or metropolized pseudoprior [19], birth and death Markov chain Monte Carlo [61]. To a certain degree, they are different derivatives of the reversible Jump Markov chain Monte Carlo (RJMCMC) [19], [6].

The RJMCMC first introduced by Green in 1995 allows posterior simulations to jump between the parameter subspaces corresponding to different numbers of components in the mixture. The essence of this method is to incorporate two additional features to Metropolis-Hastings schemes. First of all, some additional auxiliary variables are generated to augment parameters in the lower dimensional model to match dimensions across the current and the proposed model. Next, the proposal is allowed to include a deterministic mapping to map the augmented vector of parameters and auxiliary variables to a proposed parameter vector under the larger model.

Let $\Theta, \Theta^*$ denote the parameter spaces before (current state) and after (proposed state) the jump. A complete set of parameters includes $\phi = (g, p, W, \mu, \beta, \rho, \theta, Z)$ in our model. A movement (split or birth) is proposed, from $\phi^* \in \Theta^*$ to a point $\phi \in \Theta$, which is implemented by a vector of continuous random variables $u$, independent of
\( \phi \) and \( \phi^* \). The reverse of the move (combine or death) can be accomplished by using the inverse transformation, so that the proposal is deterministic. The proposal of the split step is accepted with the probability \( \alpha = \min(1, A) \), where \( A \) is defined as:

\[
A = \frac{p(\phi|y)}{p(\phi^*|y)} \frac{r(\phi \rightarrow \phi^*)}{r(\phi^* \rightarrow \phi)q(u)} \frac{\partial \phi}{\partial (\phi^*, u)},
\]

where \( r(\phi^* \rightarrow \phi) \) is defined as the probability of choosing a particular move type \( (\phi) \) when in state \( \phi^* \), and \( q(u) \) is the density function of \( u \). By allowing the size of mixture varying, it will be more flexible in estimating the cell probabilities with fixed cutpoint parameters compared to the fixed size mixture normal distributions proposed in the previous section.

Two matching moves are defined here: split and combine. Other than updating the parameter space of \( G \), the update of other parameters \( p, W, \mu, \beta, \rho, Z \) following the same steps described before.

1: The Selection of the Move: First of all, a random choice between the split and the combine is made with the probabilities of \( b_g \) and \( d_g = 1 - b_g \), respectively, depending on \( G \). If \( G = 1 \), \( b_g = 1 \) and \( d_g = 0 \). If \( G \) reaches the maximum number allowed for \( G \) \( (G_{\text{max}}) \), the probability of the split should be 0. Otherwise, set \( b_g \) and \( d_g \) to be 0.5.

2.1: The Proposal of the Split Step: In the split step, one of the current node \( (g^*) \) is randomly selected with the equal probability of \( 1/G \) over the whole parameter space. Two augmented variables are generated from the following distributions: \( u_1 \sim \)
$Beta(2, 2)$ and $u_2 \sim Beta(2, 2)$ to accommodate the increment of the parameter space. A mapping scheme is proposed below.

$$p_{g_1} = p_g u_1, \quad p_{g_2} = p_g (1 - u_1),$$

$$\mu_{g_1} = \mu_g - c u_2, \quad \mu_{g_2} = \mu_g + c u_2$$

$$c = \min(\mu_{g_{+1}} - \mu_g, \mu_g - \mu_{g_{-1}})$$

The constraint of $c$ in the splitting steps preserves the order of $\mu$ after the split steps. The reallocation of $y_{ij}$ with $g = g^*$ between $g = g_1$ and $g = g_2$ will follow similar steps in updating $w_{ij}$ in the previous section. Therefore, the probability of the reallocation is defined as:

$$p_{\text{real}} = \prod_{i,j} \frac{p_{g_1} f(\varepsilon_{ij} | \mu_{g_1}) I(w_{ij} = g_1) p_{g_2} f(\varepsilon_{ij} | \mu_{g_2}) I(w_{ij} = g_2)}{p_{g_1} f(\varepsilon_{ij} | \mu_{g_1}) + p_{g_2} f(\varepsilon_{ij} | \mu_{g_2})}$$

However, since the index of mixture models ($W$) is a latent quantity used to represent the mixed model alone, it can be marginalized out from the joint distribution with no effect on the posterior inference. Hence, the issue of reallocating elements can be avoided in this step.

2.2: The Proposal of the Merge Step: The combine step follows a similar route. First a random node $g_1$ is selected with the probability of $1/(G + 1)$, and then combined with the nearest element ($g_2$). The new $p^*$ and $\mu^*$ are generated accordingly.

$$p_{g^*} = p_{g_1} + p_{g_2}$$

$$\mu_{g^*} = \frac{1}{2}(\mu_{g_1} + \mu_{g_2})$$
Reallocate all elements with originated in Groups $g_1$ and $g_2$ into the new group $g^*$. The probability of reallocation is 1. Again, this step can be omitted if $W$ has been marginalized out of the joint distribution.

3. The Acceptance Probability: The proposed step is accepted with the probability of $\alpha = \min(1, A)$. Recall that $A$ is a combination of the posterior ratio, the proposal ratio, and the Jacobian transformation. There are two representations of $A$ depends on the existence of $W$. In the split step, if $W$ is kept in the joint distribution, $A$ is defined as:

$$A_{\text{split}} = \text{prior ratio} \times \text{likelihood ratio} \times \text{proposal ratio}$$

\[
\times \text{Jacobian of the transformation}
\]

\[
= \frac{p(\theta) p(y|\theta)}{p(\theta^*) p(y|\theta^*)} \frac{p(\text{merge})}{p(\text{split})} \frac{p_{\text{select}}(g_1, g_2)}{p_{\text{select}}(g^*^*)} \frac{1}{p_{\text{Beta}(2, 2)(u_1)p_{\text{Beta}(2, 2)(u_2)}} \frac{\partial(p_{g_{l_1}, g_{l_2}, g_{l_1}, g_{l_2}})}{\partial(p_{g_{l^*}, g_{l^*}, u_1, u_2})} \\
= \frac{p(G+1)}{p(G)} \frac{1}{\text{Beta}(\alpha, \gamma, \xi, \xi)} \frac{p_{g_1}^{-1} p_{g_2}^{-1}}{p_{g^*}^{-2}} (G + 1) \frac{f(\mu_{g_1}) f(\mu_{g_2})}{f(\mu^*)} \\
\prod_{ij} \frac{f(z_{ij}|\mu_{g_1})I(w_{ij}=g_1)f(z_{ij}|\mu_{g_2})I(w_{ij}=g_2)}{f(z_{ij}|\mu^*)I(w_{ij}\in\{g_1, g_2\})} \frac{p_{g_1}^{n_1} p_{g_2}^{n_2}}{p_{g^*}^{n^2}} \\
\frac{d_{G+1}}{d_{G}} \frac{1/(G+1)}{1/G} \frac{1}{p_{\text{Beta}(2, 2)(u_1)p_{\text{Beta}(2, 2)(u_2)}} \frac{1}{\text{Preall}}}
\]

$$2p_{g^*}c$$

The first line is the ratio of the prior distributions of the parameters of interest. Since $\beta$ and $\rho$ remains the same between the current state and the proposed state, the ratio of their prior distributions are cancelled out. $G + 1$ in the first line comes from the ratio $(G+1)!/G!$ for the order statistics of $\mu$. The second line contains the information of latent quantities during the process of the reversible jump. The likelihood $p(y|\cdots)$ is a function of $Z$ only. Therefore, it remains unchanged during the process of the
jump. The third line includes the proposal ratio, which reflects two steps. The first one is the choice of the move and the others are the decision of which elements to be moved into the proposed state, augmented by the auxiliary elements, including the reallocation probabilities. The last line is the Jacobian transformation of the random parameters. The acceptance probability \( A \) can be also written in the following form if the latent variable \( W \) is marginalized out of the joint distribution, i.e. there is no reallocation of data involved.

\[
A_{\text{split}} = \frac{p_{G+1}}{p_G} \frac{1}{\text{Beta}(\alpha, \alpha)} \frac{p_{G_1}^{a-1} p_{G_2}^{a-1}}{p_G^{a-1}} (G + 1) \frac{f(\mu_G)}{f(\mu^*)} \\
\prod_{ij} \frac{\sum_{g=1}^{G+1} p_{g} f(x_{ij} | \mu_g)}{\sum_{g=1}^{G} p_{g} f(x_{ij} | \mu_g)} \\
\frac{d_{G+1}}{b_G} \frac{1}{1/G} \text{Beta}(2,2)(\nu_1) \text{Beta}(2,2)(\nu_2) \\
2p_g c
\]

Likewise, the merge step will be accepted with the probability of \( \min(1, A_{\text{merge}}) \), where
\[
A_{\text{merge}} = A_{\text{split}}^{-1}.
\]

2.7 Applications

2.7.1 Simulated Dataset I

The performance of the models is assessed through two simulated datasets and a study conducted in Houston. There were 400 subjects in the first simulated dataset. The response variable took the value from 1 to 3 and were measured at four different timepoints. There were two covariates associated with each response variable: \( x_i = \)
\((x_{i1}, x_{i2})\). The simulated responses were generated from the following model:

\[
y_{ij} = \begin{cases} 
1 & \text{if } z_{ij} < \theta_1 \\
2 & \text{if } \theta_1 \leq z_{ij} < \theta_2 \\
3 & \text{if } \theta_2 \leq z_{ij}
\end{cases}
\]

\[
z_{ij} = \beta_0 + x_{i1} \beta_1 + x_{i2} \beta_2 + \rho(z_{ij-1} - \beta_0 - x_{i1} \beta_1 - x_{i2} \beta_2) + \varepsilon_{ij}
\]

\[
\varepsilon_{ij} \sim N(0, 1)
\]

where \(\beta = (2, 1, 0.5)\), \(\theta = (0, 8)\), and \(\rho = 0.2\). The covariates \(x_{i1}\) were generated from a Binomial(400, 0.5) to indicate treatment/group effect, and \(x_{i2}\) followed a uniform distribution \(U(-10, 20)\). Initial values under each method were listed in the table below.

Parameters were estimated using various approaches described above: Albert-Chib method (AC), the fixed size mixture (FSM), and the random size mixture (RSM). For each method, 30,000 iterations of Gibbs sampling or Gibbs sampling/RJMCMC were generated. A mixture of three components was assumed in the FSM method. The \(G_{max}\) was pre-defined at 20 for the method of the RSM. The first 5000 iterations were discarded as the 'burn-in' period, and the MCMC outputs were thinned at every 10\(^{th}\). For both the FSM and the RSM methods, the cutpoint parameters were fixed at the value of (0.6) without loss of generality. A weakly informative prior was imposed on the location parameter since a fully-noninformative prior may not warrant a meaningful posterior distribution. The prior variance of the location parameter was chosen proportional to the variance of the latent variables [57].
The Figure 2.2 presents the trace plots for parameters of interest using AC method. According to the plots, the convergence of $\beta_2$ and cutpoint parameters were very slow while it took much less iterations for chains to stabilize by the other two methods (Figures 2.3 & 2.4). The parameter estimators with the 95% credible intervals for each model are listed in Table 2.2. The marginal posterior density of parameters are presented in the Figure 2.5. The parameter estimators from the FSM and the RSM were very close since both of them assuming a mixture distribution as the underlying distribution. For the RSM, in all runs, the number of components never exceeded 8; hence, the chosen value of $G_{max}$ was inconsequential. The estimated posterior probabilities of $G$ are displayed in Table 2.3. The posterior distribution of $G$ favored 3 components without any doubts. The acceptance rate for the split and the combine steps were 1.9% and 1.4%, respectively.

Although the Albert-Chib method had the property of slow convergence when the sample size is large, it gives a reasonable well estimate of parameters. However, if data is generated from a mixed normal distributions, both the FSM and the RSM methods can capture the multimodality nature of the dataset in addition to fixing the cutpoint parameters.

2.7.2 Simulated Dataset II

A second dataset with a modest change of the underlying distribution of the latent variable was generated to gain an in-depth understanding of each model. Instead of
generating the latent quantities from a standard normal distribution, alternatively they were generated from a mixture of three normal distributions:

\[ f(\varepsilon) \sim .25N(-5, 1) + .5N(0, 1) + .25N(5, 1) \]

In other words, we assume that data was not homogenous. The standard single distribution is insufficient to represent data.

Assuming the same initial values and prior distributions as those in the previous example, the estimated marginal posterior predictive probabilities given each time-point were computed under each method (Table 2.4). Likewise, let \( \phi \) be the set of parameters of interest and \( D \) be the collection of data.

\[
\pi_k(y_{ij} = k) = E_{\phi|D}[\pi_k(y_{ij} = k|\phi)] \\
= E_{\phi|D}[p(\theta_k - 1 \leq z_{ij} < \theta_k|\phi)]
\]

Therefore, the marginal posterior predictive probabilities can be computed as:

\[
\tilde{\pi}_k(y_{ij} = k) = \frac{1}{R} \sum_{r=1}^{R} \pi_k(y_{ij} = k|\phi^{(r)})
\]

where \( R \) is the total number of iterations (Gibbs sampling and RJMCMC) and \( \phi^{(r)} \) is the distinct \( \phi \) at the \( r \)-th iterations. The posterior allocation of sample \( i \) at time \( j \) is based on the following rule:

\[
y_{ij} = \text{argmax}_{1 \leq k \leq K} (\tilde{\pi}_k(y_{ij} = k))
\]

The winner is picked and compared to the real dataset.
The first column lists the observed responses at each timepoint. The marginal posterior predicted ordinal response under each model are compared to the real dataset. At time 1, there were 94, 225, and 81 at each response level. Out of 94 subjects, 49(52.1%), 86(91.4%) and 92(97.9%) subjects were estimated to take on the value of 1 under the AC, the FSM, and the RSM methods, respectively. Due to lack of the mechanism in estimating the multimodality of data, The Albert-Chib method performed poorly in correctly estimating the cell probabilities. The sensitivities of each response level from each model were at least 88% for both the fixed size mixture and the random size mixture models. Two different scenarios are presented in the Table 2.5 to assess how different size of mixture models affects the estimated cell probabilities. As we recall, in the fixed size mixture model, the size of mixture is pre-defined. When a mixture of two normal distributions were assumed for the same set of data, the sensitivity was far lower than that of the size of three. The mixture of the two normal distributions was not flexible enough to capture the multimodality of the dataset. Hence, by adding one more or two more components into the mixture model, the sensitivity was higher than that of size of three. However, extending a mixture by adding more elements may lead to an apparently significant extra coefficient, but without a notable improvement in overall fit and at the loss of precision of the preceding parameters. An overparameterized model may not yield sensible predictions either [34]. Therefore, the RJMCMC algorithm not only provides an estimation of the parameters but also chooses a posterior most likely size of mixtures. This is
conducted through the jump between the parameter spaces. In the random size of mixture model, the acceptance probabilities of the combine step and the split step in RJMCMC were 9.2% and 9.8%, respectively (Table 2.6). The size of components ranged from 2 to 13 with 40.6% chance of being size of five. The distribution of the marginal posterior probabilities of $G$ is presented in the Table 2.4 and the Figure 2.7. The marginal posterior predictive probabilities of response for each subject from the fixed size mixture ($G=3$) and the random size mixture are displayed in the Figure 2.7 at each timepoint. Those two methods give a consistent estimate of cell probabilities while the random size mixture model offers the flexibility in identifying the size of mixture.

2.7.3 A Motivational Intervention Study in Houston, Texas Area For High Risk Young Smokers

This study used the motivational interviewing technique and health status feedback (respiratory symptoms, lung function, heart rate, and CO in expired air) to induce and accelerate progression through the stages of readiness to quit smoking among community college students - a large and rapidly growing population at increased risk for smoking. This project was guided by the transtheoretical model of change and research from a social cognitive perspective. One of the specific aims of the study was to examine the motivation interviewing in progression through the stages of readiness to quit smoking. There were four stages of readiness to quit smoking:
precontemplation, contemplation, preparation, and action. The study employed a
group-randomized, controlled design with repeated measures. Approximately 1,260
students attending the 14 community college campuses in the greater Houston area
were surveyed for normative values on respiratory symptoms and other variables of in-
terest for non-smokers. The treatment will consist of four contacts (baseline, 2-month,
and 4-month assessment, and 10-month final assessment). This study contributed sig-
nificantly to our understanding of behavior changes among an underserved group of
young adults smokers and smoking related cancer risk reduction among these indi-
viduals.

A total of 324 surveys were collected with 166 in the intervention group and the rest
158 in the control group. The average age of participants were 23 year old with an
average of one pack of cigarette per day. Two hundred seventy-eight participants had
tried to quit smoking before. Participants' age and the number of cigarettes per day
at baseline were plotted against the stage of readiness to quit at each time point with
the lowess curves(Figure 2.8). Let's refer to 0, 1, 2, 3 as each stage of readiness to
quit (precontemplation, contemplation, preparation, and action). The lowess curves
were fit to detect the trend between age, cigarettes information and the readiness to
quit smoking. The relationships between the ordinal response and the other binary
responses such as participants' quitting smoking status and treatment assignments
are displayed in the Figure 2.9. The number next to each circle in the figure is the
number of participants falling into each category. The size of each circle is propor-
tional to the number of participants in each category.

A total of 30,000 MCMC iterations have been run and the first 5000 iterations were discarded as the 'burn-in' period. The results were thinned at each 10<sup>th</sup> iteration. The Table 2.7 provides summary statistics for the posterior distribution of parameters using the AC method, the FSM model, the RSM models. Note that the FSM model and the RSM model yield posterior moment values that are similar. However, the parameter estimates obtained from the AC method are different from the other two, particularly the former has narrower 95% credible intervals for the parameters. This is due to the different underlying distributions for the latent variables. The posterior distribution on the regression coefficients under the FSM model and the RSM is plotted in the Figures 2.10 & 2.11. The posterior distributions on $\beta_1$ and $\beta_3$ suggest that participants' age and ever-quit-smoking status at the baseline had a positive effect on their readiness to quit. However, the posterior distributions of $\beta_4$ suggest to reconsider the choice of the motivational interview as opposed to the standard one.

2.8 Conclusions

A mixture of probit models with fixed cutpoint parameters has been proposed to estimate cell probabilities of an ordinal response. After comparing the fixed size mixture model, the random size mixture model, and the Albert-Chib model, the first two are preferable to the last one for several reasons. First of all, we no longer run into
the convergence problem in the setting of large sample since we keep the cutpoints constant in our models. There is no need to update cutpoint parameters jointly with other parameters as suggested by the Albert-Chib method. Second, our model can reduce the dimensions of parameters, especially in the circumstance that there are higher dimension of ordinal outcomes. Furthermore, the nature of mixture models allows us to model data from heterogeneous sources. Therefore, it provides a better estimate of cell probabilities compared to Albert-chib method, which assumes data coming from a single univariate normal distribution.

From the computational perspective, the fixed size mixtures are more efficient than the random size mixtures. However, it requires the investigator to have a good understanding of data and be able to make a sensible judgment of size of mixture. If the pre-specified size of components is less than data supposed to be, the cell probabilities obtained from such a fixed size mixture may be far off the truth. It is not good to overfit the model, either. Hence, the random size mixture using RJMCMC is flexible although it is computationally intensive. The algorithm of the RJMCMC allows the model to move between the possible parameter spaces, and then identify a possible $G$.

Special cautions need to be taken in mixture modeling. For example, a weakly informative prior on the location parameters is recommended in order to warrant a good mixing of Markov chains and a rational posterior inference. Furthermore, the
label-switching is a non-negligible issue in the mixture modeling, especially in the case where component-specific parameter is the primary interest.

In longitudinal studies we do not necessarily require the same number of observations on each subject or those measurements be taken at the same time points. Therefore, this mixture model can also be extended to model data where it is measured repeated over random time points instead of fixed time points in our study. In such as a case, a mixture of linear mixed effect model can be applied on the latent variables, where the residual terms follow a mixture Gaussian distributions. The variability within subjects is modeled by random-effect terms. This model can also handle unequal number of repeated observations over random time points within each subject. Missing data is another issue in modeling longitudinal data. If some responses for some subjects are missing, we may discard incomplete records to produce a balanced dataset, which relies on a very strong assumption that the missing values are missing completely at random (MCAR). However, sometimes, the probability of an observation being missing may depend on observed data for the subject, including covariates and observed responses, which is called missing at random (MAR). For instance, in clinical trials, the follow-up visit may depend on the toxicity of the previous treatment. If the treatment has a moderate side-effect on patients, in consequent, substantial number of patients may quit the study due to the toxicity. The mechanism of handling missing data can be incorporated into our model. Due to the structure of this model, EM algorithm to impute the missing data can easily be implemented in MCMC without
any further efforts.
Table 2.1: Initial values of parameters under each model

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Albert-Chib</th>
<th>Fixed size mixture</th>
<th>Random size mixture</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta$</td>
<td>(5, 2, 1)</td>
<td>(5, 2, 1)</td>
<td>(5, 2, 1)</td>
</tr>
<tr>
<td>$\rho$</td>
<td>0.3</td>
<td>0.3</td>
<td>0.3</td>
</tr>
<tr>
<td>$\theta$</td>
<td>(0.6)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$p$</td>
<td>(0.33, 0.33, 0.33)</td>
<td>(0.33, 0.33, 0.33)</td>
<td></td>
</tr>
<tr>
<td>$\nu$</td>
<td>3</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>$\sigma^2_\mu$</td>
<td>50</td>
<td>50</td>
<td></td>
</tr>
</tbody>
</table>

Table 2.2: Marginal posterior estimates of the parameters of interest: $\beta_1$: binary covariate; $\beta_2$: continuous covariate; $\rho$: autoregressive coefficient.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Albert-Chib</th>
<th>Fixed size mixture</th>
<th>Random size mixture</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_1$</td>
<td>1.13 (0.88, 1.38)</td>
<td>0.93 (0.71, 1.15)</td>
<td>0.92 (0.71, 1.14)</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>0.53 (0.49, 0.59)</td>
<td>0.40 (0.38, 0.41)</td>
<td>0.40 (0.37, 0.41)</td>
</tr>
<tr>
<td>$\rho$</td>
<td>0.19 (0.16, 0.23)</td>
<td>0.20 (0.15, 0.24)</td>
<td>0.19 (0.13, 0.22)</td>
</tr>
</tbody>
</table>
Table 2.3: Marginal posterior distribution of the size of mixture: $G$

| $p(G|\ldots)$ | $p(\text{Split})$ | $p(\text{Combine})$ |
|----------------|-------------------|---------------------|
| $p(1)=0.000$  | $p(2)=0.000$      | $p(3)=0.96$         | $0.019$ | $0.014$ |
| $p(4)=0.03$   | $p(5)=0.005$      | $p(6)=0.002$        |
| $p(7)=0.00045$| $p(8)<0.0001$     |

Figure 2.1: Directed acyclic graph: the squares represent fixed or observed quantities and the circles are for the unknowns.
Table 2.4: Marginal posterior predicted ordinal response at four time points. The number inside the parenthesis represents the row percentage.

<table>
<thead>
<tr>
<th>Time 1</th>
<th>Albert-Chib</th>
<th>FSM(G=3)</th>
<th>RSM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 2 3</td>
<td>1 2 3</td>
<td>1 2 3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>49 45 0</td>
<td>86 8 0</td>
<td>92 2 0</td>
</tr>
<tr>
<td>(n=94)</td>
<td>(52.1) (47.9) (0.0)</td>
<td>(91.4) (8.6) (0.0)</td>
<td>(97.9) (2.1) (0.0)</td>
</tr>
<tr>
<td>2</td>
<td>20 205 0</td>
<td>0 225 0</td>
<td>0 225 0</td>
</tr>
<tr>
<td>(n=225)</td>
<td>(8.9) (91.1) (0.0)</td>
<td>(0.0) (100) (0.0)</td>
<td>(0.0) (100) (0.0)</td>
</tr>
<tr>
<td>3</td>
<td>0 81 0</td>
<td>0 10 71</td>
<td>0 1 80</td>
</tr>
<tr>
<td>(n=81)</td>
<td>(0.0) (100) (0.0)</td>
<td>(0.0) (12.3) (87.7)</td>
<td>(0.0) (1.2) (98.8)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Time 2</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 2 3</td>
<td>1 2 3</td>
<td>1 2 3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>43 47 0</td>
<td>80 10 0</td>
<td>86 4 0</td>
</tr>
<tr>
<td>(n=90)</td>
<td>(47.8) (52.2) (0.0)</td>
<td>(89.9) (11.1) (0.0)</td>
<td>(95.6) (4.4) (0.0)</td>
</tr>
<tr>
<td>2</td>
<td>24 188 0</td>
<td>0 212 0</td>
<td>0 211 1</td>
</tr>
<tr>
<td>(n=212)</td>
<td>(11.3) (88.7) (0.0)</td>
<td>(0.0) (100) (0.0)</td>
<td>(0.0) (99.5) (0.5)</td>
</tr>
<tr>
<td>3</td>
<td>0 98 0</td>
<td>0 6 92</td>
<td>0 3 95</td>
</tr>
<tr>
<td>(n=98)</td>
<td>(0.0) (100) (0.0)</td>
<td>(0.0) (6.1) (93.9)</td>
<td>(0.0) (3.1) (96.9)</td>
</tr>
<tr>
<td>Data (Y)</td>
<td>Albert-Chib</td>
<td>FSM(G=3)</td>
<td>RSM</td>
</tr>
<tr>
<td>---------</td>
<td>-------------</td>
<td>----------</td>
<td>-----------</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Time 3</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>42 45 0</td>
<td>81 6 0</td>
<td>87 0 0</td>
</tr>
<tr>
<td>(n=87)</td>
<td>(48.3 51.7)</td>
<td>(93.0 7.0)</td>
<td>(100.0 0.0)</td>
</tr>
<tr>
<td>2</td>
<td>23 168 0</td>
<td>0 188 3</td>
<td>0 191 0</td>
</tr>
<tr>
<td>(n=191)</td>
<td>(12.0 88.0)</td>
<td>(0.0 98.4)</td>
<td>(0.0 100.0)</td>
</tr>
<tr>
<td>3</td>
<td>0 122 0</td>
<td>0 4 118</td>
<td>0 0 122</td>
</tr>
<tr>
<td>(n=122)</td>
<td>(0.0 100)</td>
<td>(0.0 3.2)</td>
<td>(0.0 100)</td>
</tr>
<tr>
<td><strong>Time 4</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>49 41 0</td>
<td>87 3 0</td>
<td>89 1 0</td>
</tr>
<tr>
<td>(n=87)</td>
<td>(56.3 43.7)</td>
<td>(96.7 3.3)</td>
<td>(98.9 1.1)</td>
</tr>
<tr>
<td>2</td>
<td>18 186 0</td>
<td>0 203 1</td>
<td>0 204 0</td>
</tr>
<tr>
<td>(n=204)</td>
<td>(8.8 91.2)</td>
<td>(0.0 99.5)</td>
<td>(0.0 100)</td>
</tr>
<tr>
<td>3</td>
<td>0 106 0</td>
<td>0 4 102</td>
<td>0 0 106</td>
</tr>
<tr>
<td>(n=106)</td>
<td>(0.0 100)</td>
<td>(0.0 3.8)</td>
<td>(0.0 100)</td>
</tr>
</tbody>
</table>
Table 2.5: Marginal posterior predicted ordinal response under the fixed size mixture model

<table>
<thead>
<tr>
<th>Data (Y)</th>
<th>G=2</th>
<th>G=4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time 1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>82</td>
<td>12</td>
</tr>
<tr>
<td>(n=94)</td>
<td>(87.2%)</td>
<td>(12.3%)</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>225</td>
</tr>
<tr>
<td>(n=225)</td>
<td>(0.0%)</td>
<td>(100.0%)</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>20</td>
</tr>
<tr>
<td>(n=81)</td>
<td>(0.0%)</td>
<td>(25.0%)</td>
</tr>
</tbody>
</table>

<p>| Time 2   |     |     |     |     |     |     |
| 1        | 71  | 19  | 0   | 82  | 8   | 0   |
| (n=90)   | (78.9%) | (21.1%) | (0.0%) | (91.2%) | (8.8%) | (0.0%) |
| 2        | 0   | 210 | 2   | 0   | 212 | 0   |
| (n=212)  | (0.0%) | (99.1%) | (0.9%) | (0.0%) | (100.0%) | (0.0%) |
| 3        | 0   | 9   | 89  | 0   | 2   | 96  |
| (n=98)   | (0.0%) | (9.2%) | (90.8%) | (0.0%) | (2.0%) | (98.0%) |</p>
<table>
<thead>
<tr>
<th>Data (Y)</th>
<th>G=2</th>
<th>G=4</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Time 3</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>76</td>
<td>83</td>
</tr>
<tr>
<td>(n=87)</td>
<td>(87.3%)</td>
<td>(95.4%)</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(n=191)</td>
<td>(0.0%)</td>
<td>(0.0%)</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(n=122)</td>
<td>(0.0%)</td>
<td>(0.0%)</td>
</tr>
<tr>
<td><strong>Time 4</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>84</td>
<td>87</td>
</tr>
<tr>
<td>(n=90)</td>
<td>(93.3%)</td>
<td>(96.7%)</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(n=204)</td>
<td>(0.0%)</td>
<td>(0.0%)</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(n=106)</td>
<td>(0.0%)</td>
<td>(0.0%)</td>
</tr>
</tbody>
</table>
Table 2.6: Marginal posterior distribution of G

|          | p(G|...) | p(Split) | p(Combine) |
|----------|----------|----------|------------|
| p(1)     | 0.000    |          |            |
| p(2)     | 0.072    |          |            |
| p(3)     | 0.002    |          |            |
| p(4)     | 0.145    | 0.092    | 0.098      |
| p(5)     | 0.406    |          |            |
| p(6)     | 0.128    |          |            |
| p(7)     | 0.097    |          |            |
| p(8)     | 0.061    |          |            |
| p(9)     | 0.033    |          |            |
| p(10)    | 0.016    | 0.033    | 0.006      |
| p(11)    | 0.033    |          |            |
| p(12)    | 0.006    |          |            |
| p(13)    | 0.003    |          |            |
Table 2.7: Posterior mean, 95% credible interval, and the tail probability of parameters computed using three methods. $\beta_1$: age; $\beta_2$: the number of cigarettes per day at the baseline; $\beta_3$: ever quitting smoking before, yes vs. no; $\beta_4$: treatment assignment, 1: motivational interview, 0: standard

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Albert-Chib</th>
<th>Fixed size mixture</th>
<th>Random size mixture</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Median (95% C.I.)</td>
<td>Median (95% C.I.)</td>
<td>Median (95% C.I.)</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>0.04 (0.01, 0.08)</td>
<td>0.09 (0.03, 0.14)</td>
<td>0.09 (0.02, 0.14)</td>
</tr>
<tr>
<td></td>
<td>0.997</td>
<td>0.998</td>
<td>0.996</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>-0.04 (-0.06, -0.02)</td>
<td>-0.06 (-0.09, -0.02)</td>
<td>0.05 (-0.09, -0.02)</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>0.003</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>0.59 (0.19, 1.00)</td>
<td>0.89 (0.19, 1.52)</td>
<td>0.89 (0.06, 1.68)</td>
</tr>
<tr>
<td></td>
<td>0.999</td>
<td>0.997</td>
<td>0.98</td>
</tr>
<tr>
<td>$\beta_4$</td>
<td>-0.17 (-0.44, 0.11)</td>
<td>-0.20 (-0.70, 0.23)</td>
<td>-0.37 (-1.01, 0.19)</td>
</tr>
<tr>
<td></td>
<td>0.12</td>
<td>0.20</td>
<td>0.07</td>
</tr>
</tbody>
</table>
Table 2.8: Marginal posterior predicted response (readiness to quit) at four time points: 0: precontemplation; 1: contemplation; 2: preparation; 3: action. The first line in each block is the estimations using the Albert-Chib method. The results in the second line are obtained from the fixed size mixture and those in the last line are from the random size mixture

<table>
<thead>
<tr>
<th>Data (Y)</th>
<th>Albert-Chib</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Fixed size mixture (G=3)</td>
</tr>
<tr>
<td></td>
<td>Random size mixture</td>
</tr>
<tr>
<td>Baseline</td>
<td>0</td>
</tr>
<tr>
<td>0 (n=36)</td>
<td>36 (100%)</td>
</tr>
<tr>
<td></td>
<td>0 (0%)</td>
</tr>
<tr>
<td></td>
<td>0 (0%)</td>
</tr>
<tr>
<td></td>
<td>0 (0%)</td>
</tr>
<tr>
<td>18 (50%)</td>
<td>18 (50%)</td>
</tr>
<tr>
<td>14 (39%)</td>
<td>22 (62%)</td>
</tr>
<tr>
<td>1 (n=209)</td>
<td>209 (100%)</td>
</tr>
<tr>
<td></td>
<td>0 (0%)</td>
</tr>
<tr>
<td></td>
<td>0 (0%)</td>
</tr>
<tr>
<td></td>
<td>0 (0%)</td>
</tr>
<tr>
<td>209 (100%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>0 (0%)</td>
<td>209 (100%)</td>
</tr>
<tr>
<td>0 (0%)</td>
<td>209 (100%)</td>
</tr>
<tr>
<td>0 (0%)</td>
<td>209 (100%)</td>
</tr>
<tr>
<td>2 (n=79)</td>
<td>79 (100%)</td>
</tr>
<tr>
<td></td>
<td>0 (0%)</td>
</tr>
<tr>
<td></td>
<td>0 (0%)</td>
</tr>
<tr>
<td></td>
<td>0 (0%)</td>
</tr>
<tr>
<td>0 (0%)</td>
<td>12 (15%)</td>
</tr>
<tr>
<td>67 (85%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>0 (0%)</td>
<td>5 (6%)</td>
</tr>
<tr>
<td>74 (94%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Month 2</td>
<td>Data (Y)</td>
</tr>
<tr>
<td>---------</td>
<td>----------</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>0 (n=23)</td>
<td>1 (4%)</td>
</tr>
<tr>
<td></td>
<td>14 (61%)</td>
</tr>
<tr>
<td></td>
<td>19 (83%)</td>
</tr>
<tr>
<td>1 (n=195)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td></td>
<td>0 (0%)</td>
</tr>
<tr>
<td></td>
<td>0 (0%)</td>
</tr>
<tr>
<td>2 (n=64)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td></td>
<td>0 (0%)</td>
</tr>
<tr>
<td></td>
<td>0 (0%)</td>
</tr>
<tr>
<td>3 (n=42)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td></td>
<td>0 (0%)</td>
</tr>
<tr>
<td></td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Month 4</td>
<td>Data (Y)</td>
</tr>
<tr>
<td>--------</td>
<td>----------</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>0 (n=23)</td>
<td>2 (9%)</td>
</tr>
<tr>
<td></td>
<td>17 (74%)</td>
</tr>
<tr>
<td></td>
<td>23 (100%)</td>
</tr>
<tr>
<td>1 (n=173)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td></td>
<td>0 (0%)</td>
</tr>
<tr>
<td></td>
<td>0 (0%)</td>
</tr>
<tr>
<td>2 (n=66)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td></td>
<td>0 (0%)</td>
</tr>
<tr>
<td></td>
<td>0 (0%)</td>
</tr>
<tr>
<td>3 (n=62)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td></td>
<td>0 (0%)</td>
</tr>
<tr>
<td></td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Month 10</td>
<td>Data (Y)</td>
</tr>
<tr>
<td>---------</td>
<td>----------</td>
</tr>
<tr>
<td></td>
<td>Fixed size mixture (G=3)</td>
</tr>
<tr>
<td>0 (n=31)</td>
<td>2 (7%)</td>
</tr>
<tr>
<td></td>
<td>10 (32%)</td>
</tr>
<tr>
<td></td>
<td>29 (88%)</td>
</tr>
<tr>
<td>1 (n=132)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td></td>
<td>0 (0%)</td>
</tr>
<tr>
<td></td>
<td>0 (0%)</td>
</tr>
<tr>
<td>2 (n=75)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td></td>
<td>0 (0%)</td>
</tr>
<tr>
<td></td>
<td>0 (0%)</td>
</tr>
<tr>
<td>3 (n=86)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td></td>
<td>0 (0%)</td>
</tr>
<tr>
<td></td>
<td>0 (0%)</td>
</tr>
</tbody>
</table>
Figure 2.2: Trace plots of parameters under the Albert-Chib method for the simulated dataset I
Figure 2.3: Trace plots of parameters under the fixed size mixture model for the simulated dataset I
Figure 2.4: Trace plots of parameters under the random size mixture model for the simulated dataset I
Figure 2.5: Marginal posterior density plots for parameters associated with the simulated dataset I
Figure 2.6: Density plots of marginal posterior distribution of the size of mixture, $G$
Figure 2.7: Comparisons of marginal posterior predictive probabilities
Figure 2.8: Distribution of participant’s characteristics versus the stage of readiness to quit: 0: precontemplation; 1: contemplation; 2: preparation; 3: action
Figure 2.9: Distribution of participant's characteristics versus the stage of readiness to quit: 0: precontemplation; 1: contemplation; 2: preparation; 3: action
Figure 2.10: Posterior distributions of the regression coefficients under the fixed size mixture model (G=3)
Figure 2.11: Posterior distributions of the regression coefficients under the random size mixture model
Chapter 3

Clustering Analysis for
Multivariate Ordinal Data

3.1 Introduction

Clustering is an important technique in many disciplines such as gene selection, medical image analysis, and market segmentation, etc. The goal of cluster analysis is to partition a set of observations into a distinct number of unknown groups or clusters in such a manner that all observations within a group are similar, while observations in different groups are not similar. Comprehensive discussion can be found in many books [3], [33], [25].

There are a variety of clustering algorithms. Among many approaches to clustering, three of the most common ones are hierarchical, partitioning, and overlapping. Hi-
erarchical methods follow either a top-down or a bottom-up approach. In a typical top-down approach, the algorithm begins with each data point represented by a different cluster. Similar data points are then grouped together to form larger groups. This process proceeds until the final step when the entire data set is represented as a single cluster. Bottom-up methods reverse this process, starting with one cluster and then dividing it into smaller and smaller clusters. One advantage of the hierarchical approach is that the results can be presented in a tree-like structure called a dendrogram. In applied work, the researcher will often "cut the tree" at a particular stage that appears useful.

In contrast to hierarchical methods, partitioning methods often start by making an assumption about the number of clusters in data and their center points. These "centers" may represent actual data points or hypothetical ones. Data points are then assigned to the nearest cluster center based on such things as minimum Euclidean distance or some other measure of "dissimilarity" [37]. Depending on the resulting distribution, the investigator may then adjust the assumptions and repeat the process. MacQueen's k-means algorithm begins with k seeds and reassigns the rest of objects using nearest centroid sorting such as mean, median, or Euclidean distance at each step [42]. Reallocations continue until changes in cluster centroid become small and meet some convergence criterion. However, The K-means method is applicable only when mean is defined, which makes it difficult for categorical data. Additionally, the number of clusters needs to be specified in advance. Furthermore, it is sensitive
to noisy data, outliers, and initial seed, especially not suitable to discover clusters with non-convex shapes. Both hierarchical and partitioning methods produce a strict partition of data in which each data point is assigned to a single cluster. With respect to ordinal data, most methods either treat them as interval data or assign ranks to each order to measure the 'dissimilarity'.

Motivated mainly by a problem of clustering cancer patients based on their different biomarkers' expression levels, we proposed a Bayesian probabilistic model in estimating the clustering memebership using mixtures of Gaussian distributions. We regard the task of clustering as a Bayesian model selection problem. Some nice reviews of model-based clustering algorithm can be found in McLachlan & Peel [48], Fraley & Raftery [26], Oh & Raftery [52]. In this framework, a latent discrete variable $W$ is introduced as the clustering membership. The clustering of data is a process of identifying separate models, $P(M|W)$. We select the model with maximum posterior probability. The clustering membership is updated adaptively through Markov chains Monte Carlo approaches. A mixture of probit model was proposed to model the ordinal property of data assuming the existence of underlying latent variables associated with each ordinal response. Reversible jump MCMC is used in estimating the size of mixture, which gives direct information regarding the number of clusters. In summary, the proposed model can intelligently identify the number of clusters first, then, allocate the data point to each cluster according to the maximum posterior probability.
The manuscript is organized as the follows. The model of clustering analysis is presented in Section 3.2. The prior distributions, the full conditional posterior distributions and potential issues in Gibbs sampling and RJMCMC are discussed in Section 3.3. Finally two data sets are presented to illustrate the model and compare it with the K-means method (Section 3.4).

3.2 Bayesian Clustering Analysis for Multivariate Ordinal Data

For each subject, a set of \( J \) number of ordinal outcomes are measured. The variable \( y_{ij} \) reports the \( j^{th} \) ordinal outcome for the \( i^{th} \) subject, which can take the value ranging from 1 to \( K \). In order to model ordinal outcomes, we introduce a set of latent variables, \( Z_i = \{z_{i1}, \ldots, z_{ij}\} \) for the corresponding ordinal outcomes \( Y_i = \{y_{i1}, \ldots, y_{ij}\} \). The probit model is proposed to characterize the ordinal outcomes. I.e., the probability of any ordinal measurement being of \( k \) \( (y_{ij} = k) \) is represented as the probability of latent continuous variable, \( z_{ij} \), falling into the interval of \( [\theta_{j,k-1}, \theta_{j,k}] \), where \( \theta \)s are various cutoff points associated with ordinal outcomes and latent variables. We impose a mixture of multivariate normal distributions on \( Z_i \).

\[
\begin{align*}
    p(y_{ij} = k) &= p(\theta_{j,k-1} \leq z_{ij} < \theta_{j,k}). \\
    Z_i &\sim \sum_{g=1}^{G} p_g MVN_f(\mu_g, \Sigma_g) 
\end{align*}
\] (3.1)
where, $\text{MVN}_J(.)$ denotes a J-dimension multivariate normal distribution. The variables $G$ and $p_g$ are used to represent the size and the weight of mixture. Each component of the mixture distributions has a distinct location parameter $\mu_g$ and a scale parameter $\Sigma_g$. The mixture model can be alternatively rewritten as a hierarchical model by introducing another latent variable $w_i$:

$$Z_i|w_i = g \sim \text{MVN}_J(\mu_g, \Sigma_g)$$

$$p(w_i = g) = p_g.$$  \hspace{1cm} (3.2)

The random variable $w_i$ is defined as the clustering membership to group $n$ subjects into $G$ clusters. Therefore, the goals are to estimate the size of mixture ($G$) and the probability of an observed data falling into a specific cluster. Generally speaking, there are $K - 1$ various cutpoint parameters associated with each pair of $z_{ij}$ and $y_{ij}$. In the conventional Bayesian ordinal probit model, the cutpoints are considered as random parameters and sampled from the full complete conditional distributions along with other parameters. Hence, a total of $(K - 1)J$ cutpoint parameters need to be estimated at each MCMC iteration. However, the property of mixture model here allows us to capture the cell probabilities of ordinal outcome by fixing the cutpoints without loss of generality.

Such a model gives us an insight into the biological connections between biomarkers. The independence assumptions we make are conditional ones. We assume that the expression level of the same biomarker in different condition (subject) are independent given the cluster the subject belongs to. This assumption states that the cluster
captures the 'first order' description of the biomarker's behavior, and treats all other fluctuations within the cluster as noise that is independent in each measurement. The model is attractive for several reasons. First, from estimation point of view we need to estimate relatively few parameters: the clustering membership index, and parameters associated with the conditional distributions. There is no need to estimate the cutpoint parameters since they are fixed without loss of generality. Second, the estimated model can be interpreted as modeling data by $G$ clusters. Thus, dependencies between the observed variables are represented by the cluster variable. Finally, by using RJMCMC, this model allows us to estimate the number of clusters without pre-defining it.

### 3.3 Priors and Posteriors

#### 3.3.1 Prior Elicitations

We use conjugate priors for the parameters, centering the prior to represent the prior judgment about the marginal prevalence of the outcomes and the effects of the covariates. Assuming $G$ follows a negative-binomial distribution, the prior size of $G$ is set to be $a(1 - b)/b = 3$. The prior choice of $\alpha$ is 1. The prior probability of the
clustering membership is the weight of each component.

\[ p(G) \sim NB(a, b) \]

\[ p(p_1, \ldots, p_G | G) \sim D(\alpha, \ldots, \alpha) \]

\[ p(w_i | G) \sim p_g, \]

The conjugate prior for a multivariate normal distribution is a normal-Wishart distribution.

\[ p(\mu_g | G) \sim MVN_J(\xi_g, \tau_g \Sigma_g) \]

\[ p(\Sigma_g^{-1} | G) \sim W_J(\gamma_g, W_g) \]

where \( \xi_g^T = (\xi_{g1}, \ldots, \xi_{gJ}) \). In the setting of mixture distributions, improper priors may lead to improper posterior distributions. Therefore, empirical-Bayes approximation may be used to get over the problem of the absence of prior information in one of the component. Following the choice of priors in Richardson and Green [57], we could choose \( \xi_{gj} = \min(Z_j) + gH_j/(G + 1) \) for \( j = 1, \ldots, J \), where \( Z_j^T = (z_{1j}, \ldots, z_{nj}) \) denotes the latent values for \( j^{th} \) ordinal measurement and \( H_j \) denotes the range of data in the measurement \( j \). Instead, we follow the prior specifications in Cappe et al. [7] and set \( \xi_{gj} = \bar{Z}_j \), the sample mean of \( Z_j \), for all components, and \( \tau_g = 1 \). \( W_J \) denotes a \( J \)-dimensional Wishart distribution with \( \gamma \) degrees of freedom and \( W_g \) is a component-specific scale matrix. When \( J \) equals to 1, the Wishart distribution is reduced to a Gamma distribution \( Ga(\gamma_g/2, W_g^{-1}/2) \). We choose \( \gamma_g = 3 + J \) for all components and \( W_g = 3S_g^{-1} \), where \( S_g \) is the empirical variance-covariance matrix of the latent variables. Therefore, the expected scale of the prior distributions is \( S_g/2 \) when \( J = 3 \).
3.3.2 Gibbs Sampling: Full Conditional Distributions

For parameters within the same parameter space, Gibbs sampling is used to make posterior inference. The complete posterior conditional distribution of the weight component is a Dirichlet distribution with the parameters of \((\alpha_1 + n_1, \ldots, \alpha_G + n_G)\), where \(n_g\) is the number of data points in each cluster. The full conditional posterior distributions of \(\mu_g\) and \(\Sigma_g^{-1}\) follow multivariate normal and Wishart distributions given the specified conjugate priors.

\[
f(\mu_g | G, \Sigma_g, W, Z) \propto MVN_J(\xi_g^*, \tau_g^* \Sigma_g)
\]

\[
f(\Sigma_g^{-1} | G, W, Z) \propto W_J(\gamma_g^*, W_g^*)
\]

\[
\xi_g^* = (\xi_{g1}^*, \ldots, \xi_{gJ}^*, \ldots, \xi_{gJ}^*)^T
\]

\[\xi_{gj}^* = \frac{\sum_{i=1}^n z_{ij} I(w_i = g) + \xi_{gj} \tau_g^{-1}}{n_g + \tau_g^{-1}}, \text{ for } j = 1, \ldots, J\]

\[\tau_g^* = \frac{\tau_g}{\tau_g n_g + 1}\]

\[\gamma_g^* = \gamma_g + n_g\]

\[\bar{Z}_g^T = (z_{g1}, \ldots, z_{gJ})\]

\[z_{gj} = \frac{\sum_{i=1}^n z_{ij} I(w_i = g)}{n_g}\]

\[W_g^* = \left( W_g^{-1} + S_g + \frac{n_g}{\tau_g n_g + 1} (\bar{Z}_g - \xi_g)(\bar{Z}_g - \xi_g)^T \right)^{-1}\]

\[S_g = \sum_{i:w_i=g} (Z_i - \bar{Z}_g)(Z_i - \bar{Z}_g)^T, \quad g = 1, \ldots, G\]

Instead of sampling \(z_{ij}\)s from a J-dimensional truncated multivariate normal distribution, we draw \(z_{ij}\) from a truncated univariate normal distribution. Let \(i(-j)\) be the index for a vector of \(J - 1\) elements excluding \(ij^\text{th}\) component for \(i^\text{th}\) individual.
Conditional on the other \( z_{i(-j)} \)s, \( z_{ij} \) follows a truncated univariate normal distribution constrained by the fixed cutpoint parameters \( \theta_s \) [29].

\[
f(z_{ij} | z_{i(-j)}, y_{ij}, \mu_g, \Sigma_g) \propto N(\mu_g(-j) + m_{ij}, \Sigma_{gj,gj} - m_{ij}^T I(\theta_{j,k-1} \leq y_{ij} \leq \theta_{j,k})
\]

where,

\[
m_{ij} = \Sigma_{gj,g(-j)}^{-1} \Sigma_{g(-j),g(-j)}^{-1} (z_{i(-j)} - \mu_g(-j))
\]

\[
m_{ij}' = \Sigma_{gj,g(-j)}^{-1} \Sigma_{g(-j),g(-j)}^{-1} \Sigma_{gj,g(-j)}
\]

### 3.3.3 Reversible Jump MCMC

The size of mixture, i.e., the possible number of clusters is determined by reversible jump MCMC (RJMCMC) algorithm due to the possible change of parameter spaces. The change of parameter space involves the following parameters: \( p_g, \mu_g, \Sigma_g \). Green and Richardson [57] proposed the movement between the lower dimension and the higher dimension by matching the first, second, and third moments of the full conditional distributions of the weight, location, and scale parameter of mixture distribution. We could adopt the same strategy here by matching the above parameters in different dimension spaces. However, one of the major problems arising in the the split move is that the scale matrix of two new components may not be positive definitive. There are two essential requirements using RJMCMC: deterministic and reversible. Therefore, this approach violates one of the principals of RJMCMC. Yet, Dellaportas and Papageorgious [20] proposed to use the spectral decomposition of the current scale matrix and then operate the reversible jump steps on the resulting eigenvalues and eigenvectors. Alternatively, the scale parameters can be marginalized out of the
likelihood in the RJMCMC step so that the jump is performed on the weight and
the location parameters only. Therefore, the multivariate normal distributions of $Z_i$
are replaced by the multivariate $T$ distributions in the likelihood function. The joint
distribution of the likelihood ($L$) and the prior distributions can be rewritten as the
follows:

$$L \times p(G) \times p(P|G) \times p(W|G,P) \times \prod_{g=1}^{G} f'(\mu_g|G,W) \times f'(Z|W,g,G)$$

where, $P = (p_1, \ldots, p_G)$, $W = (w_1, \ldots, w_n)$

$$f'(\mu_g|G) \sim MVT_{\nu_g}(\xi_g, \Sigma_g)$$

$$\nu_g = \gamma_g + 1 - J, \quad \Sigma_g = \tau_g W_g^{-1}/\nu_g$$

$$f'(Z|\mu_g,W,G) \sim MVT_{\nu_g}(\mu_g, \Sigma_g)$$

$$\nu_g = \gamma_g + 2 - J, \quad \Sigma_g = (\tau_g^{-1}(\mu_g - \xi_g)^T(\mu_g - \xi_g) + W_g^{-1}) / \nu_g$$

Let the superscript $*$ and $t$ denote the variables in the lower and higher dimension
space and the subscripts $I$, $2$ denote the elements to be combined in the higher
dimension space. Let $\phi^*$ and $\phi^t$ denote the set of all variables in the respective space.
Let $G_{\text{max}}$ be the predefined maximum size of the mixture.

**A:** The split or the merge step will be chosen with the equal probabilities when
$1 < G < G_{\text{max}}$. Only the merge step will be performed if $G_{\text{max}}$ has been reached and
only the split step will be conducted when $G$ is 1. 

\[
p(split) = \begin{cases} 
1 & \text{if } G = 1 \\
0.5 & \text{if } 1 < G < G_{\text{max}} \\
0 & \text{if } G = G_{\text{max}} 
\end{cases}, \quad \text{p(merge)} = \begin{cases} 
0 & \text{if } G = 1 \\
0.5 & \text{if } 1 < G < G_{\text{max}} \\
1 & \text{if } G = G_{\text{max}} 
\end{cases}
\]

**B1:** At the split step, elements are randomly chosen with the probability of $1/G$ and the subjects are reallocated into two new adjacent elements, which satisfy the following conditions:

\[
p_{g_1} = p_g u_1, \quad p_{g_2} = p_g (1 - u_1)
\]

\[
\mu_{g_1} = \mu_g - u_2, \quad \mu_{g_2} = \mu_g + u_2
\]

\[
u_1 \sim \text{Beta}(1.1, 1.1), \quad u_2 \sim \text{MVN}_J(0, \Sigma_g^*)
\]

where, $u_2 = (u_{21}, \ldots, u_{2J})$ and $\Sigma_g^*$ is the empirical scale matrix. Due to the nature of multi-dimension, it is not feasible to implement the ordering constraint for the adjacent elements as the one defined in the one-dimension space. The adjacent elements are defined as any pair of elements with the smallest Mahalanobis distance. At each time, the Mahalanobis distance (see below) between two generated components is computed and compared with the distances between any other pairs of elements. The new ones will be kept to compute the selection probability if they are the nearest ones. Otherwise, the selection probability of this step is assumed to be zero.

**B2:** A pair of adjacent elements $(g_i, g_j)$ was chosen with the probability of

\[
p_{\text{select}}(g_i, g_j) = 1/d(g_i, g_j) / \sum_{i \in \mathcal{K}} \sum_{j \in \mathcal{K}} 1/d(g_i, g_j),
\]
where \( d \) is the Mahalanobis distance between \( g_i, g_j \), and defined as

\[
d(g_i, g_j) = (\mu_{g_i} - \mu_{g_j})\Sigma_{g_i}^{-1}(\mu_{g_i} - \mu_{g_j}) + (\mu_{g_j} - \mu_{g_i})\Sigma_{g_j}^{-1}(\mu_{g_j} - \mu_{g_i}).
\]

The combine step is determined based on the following steps:

\[
p_{g^*} = p_{g_1} + p_{g_2}, \quad \mu_{g^*} = (\mu_{g_1} + \mu_{g_2})/2
\]

\[
u_1 = p_{g_1}/p_{g^*}, \quad \nu_2 = \mu_{g^*} - \mu_{g_1}
\]

C: A split move is accepted with the probability of

\[
\alpha_{\text{split}} = \min(1, A)
\]

The full form of \( A \), given the target distributions and split/combine transformation,

is defined as

\[
A = \frac{p(\phi' | Y) p_{\text{merge}}(G + 1) p_{\text{merge}}^{\text{select}}(g_1, g_2)}{p(\phi^* | Y) p_{\text{split}}(G) p_{\text{split}}^{\text{select}}(g^*)} \times \\
\frac{\frac{1}{p_{\text{Beta}(1,1,1,1)}(u_1)\prod_{j=1}^J p_{N_j(0, \Sigma_{g^*})}(u_2j)} \times \left| \frac{\partial \phi'}{\partial (\phi^*, u)} \right|}{p(\mu_{g_1}) p(\mu_{g_2}) p_{g_1}^{-1} p_{g_2}^{-1} 1}{\frac{1}{p_{\text{Beta}(\gamma, G\gamma)}(u_2j)} \times \\
(p_{g_1, \text{MVT}}(Z_i | \mu_{g_1}, W_{g_1}) + p_{g_2, \text{MVT}}(Z_i | \mu_{g_2}, W_{g_2}^*))}{p_{g^*, \text{MVT}}(Z_i | \mu_{g^*}, W_{g^*})} \times \\
\frac{p_{\text{merge}}(G + 1) 1/(G + 1) p(G + 1)}{p_{\text{split}}(G) 1/G p(G)} \times \\
\frac{1}{p_{\text{Beta}(1,1,1,1)}(u_1)\prod_{j=1}^J p_{N_j(0, \Sigma_{g^*})}(u_2j)} \times 2 \times p_{g^*}^J}
\]

The combined step is accepted with the probability of

\[
\alpha_{\text{combine}} = \min(1, A^{-1}).
\]
3.4 Applications

3.4.1 A Simulated Dataset

One simulated dataset and one real dataset have been used to illustrate the model specifications. We first start with a simulated dataset. There were a total of 200 subjects. Each subject had two ordinal measurements: \( y_{i1} \) and \( y_{i2} \). Hence, there were a set of continuous latent variables: \( z_{i1} \) and \( z_{i2} \), which were jointly drawn from a mixture of bivariate normal distributions.

\[
f(Z_i) \sim p_1 MV N_2(\mu_1, \Sigma_1) + p_2 MV N_2(\mu_2, \Sigma_2) + p_3 MV N_2(\mu_3, \Sigma_3)
\]

\[
p_1 = p_2 = p_3 = \frac{1}{3}
\]

\[
\mu_1 = (-1.5, -1.5)^T, \quad \mu_2 = (1, 1)^T, \quad \mu_3 = (-2, 2)^T
\]

\[
\Sigma_1 = \begin{pmatrix} .25 & -1.75 \\ -1.75 & .25 \end{pmatrix}, \quad \Sigma_2 = \begin{pmatrix} .25 & .05 \\ .05 & .25 \end{pmatrix}, \quad \Sigma_3 = \begin{pmatrix} .25 & .1 \\ .1 & .25 \end{pmatrix}
\]

A set of cutoff points \( \theta = (-2.5, -1, 0, 1, 2.5)^T \)'s were used to categorize the continuous variables into six groups, from 0 to 5. Therefore, data could be viewed as being composed of three clusters, which represented by each component of the mixture bivariate distributions. The goal is to identify the three clusters based on the two ordinal outcomes. The prior distributions were specified as described in the subsection of prior specifications.

A total of 30,000 iterations of Gibbs sampling and RJMCMC were run and the first
10,000 burn-in period. The marginal posterior distributions of the number of clusters is presented in the Figure 3.1. Out of 20,000 iterations, almost ninety-six percent of time, the model was identified as a mixture of four to six components (G=4: 6744 (33.7%); G=5: 9099 (45.5%); G=6: 3320 (16.6%)). The selection probabilities of the split and the merge step were 0.042 and 0.040, respectively.

A straightforward classification rule was proposed. Data belongs to the group to which it belongs most frequently a posteriori. For instance, one of the 200 subjects in our example never belongs to the first group. There are 16 times it belonging to the second group and the rest times belonging to the third group. As a consequence, this subject is said to belong to the third group. Formally it is represented by the averaged posterior probability of $w_i$ given $G$.

$$\frac{1}{R} \sum_{r=1}^{R} p(w_i^{(r)}|G^{(r)}, P^{(r)}, Z^{(r)}, \mu_g^{(r)}, \Sigma_g^{(r)})$$

where $R$ is the total number of iterations. We applied this rule to all 200 subjects and summarized the classification results in Table 3.1. The distributions of the latent quantities at the last iteration are displayed in Figure 3.2, with different colors and types of points for different clusters. The size of each point symbolizes the average posterior probability of each subject being classified in the corresponding cluster.

We draw attentions to the following several points. Let black 'o', blue 'Δ', red '+', and green 'x' stand for the Clusters 1, 2, 3, and 4 respectively. When the size of mixture was 3, the number of clusters was 3 (upper left panel). However, when the
size of mixture was 4, the size of clusters was identified to be 3 indeed. In the bottom
two panels of the Figure 3.2, the sizes of mixture were 5 and 6, respectively. However,
the number of clusters was 4 for both scenarios. Therefore, the number of clusters
might be different from the actual size of mixture. It is because that the number of
clusters is determined after averaging out the posterior probabilities of \( W \). Secondly,
The Cluster 4 identified in the lower bottom two panels was on the borders of the
Clusters 1 and 2. This group of data could be either classified into the same group
as other data in the Clusters 1/2 or a separate group (Cluster 4). Therefore, the
maximum value of the posterior probability of being in the Cluster 4 was mediated
by the relative high probabilities of being in the Clusters 1 & 2, which was reflected in
the relatively smaller size of points. In consequence, in the upper two panels, instead
of classifying those data into a separate cluster, they were considered to reside in
the Cluster 1. Furthermore, the percentage of correct classifications were 99%, 99%,
94.5%, and 93.5% when \( G \) was 3, 4, 5, and 6, respectively.

3.4.2 Clustering Analysis of Biomarker Study

A dataset of 84 patients with non-small-cell lung cancer has been analyzed using the
proposed model. There were three apoptosis-related biomarkers (Caspase, FAS, and
FUS1) associated with non-small-cell lung cancer patients. The induction of apo-
ptosis is a highly regulated process that can be initiated by a variety of extracellular
ligand-directed or intracellular stress-induced stimuli. Caspases act in a central role
to both initiate and execute the intracellular cascade of events that result in protein and nucleic acid cleavage and ultimate cell death. FAS (Fatty acid synthase) has been established as a biomarker and prognostic indicator for lung cancer. Elevated FAS levels are associated with reduced level of cell apoptosis. FUS1 is a tumor suppressor gene. Lack of FUS1 activity is associated with the early onset of lung cancer. All biomarker expression levels were quantified at a scale from 0 to 3. A higher score represented a higher expression level. The distributions of biomarkers are listed in the Table 3.2. The goal was to cluster patients based on their biomarker levels, eventually relate biomarker expressions to their clinical outcomes such as overall survival and time to progression. We used the cutpoints of (-2, 0, 2) for all ordinal outcomes and set the $G_{\text{max}}$ as 10. The choice of priors follows the discussions ahead.

We first clustered patients based on the two biomarkers: Caspase and FUS1. A total of 30,000 iterations of MCMC were conducted with the first 10,000 treated as the burn-in period. The RJMCMC visited 6 models with the posterior probabilities of $(0.04, 0.15, 0.44, 0.28, 0.09, 0.004)$ for $G = 2, \ldots, 7$. The probabilities of the split and the merge steps were 3.0% and 2.5%, respectively. The estimates of the marginal posterior model probabilities are presented in the Figure 3.4. There were more almost 86% chance that the model is favoring the mixture of 3, 4, and 5 components. Therefore, the inference was made conditional on $G = 4$, which had the highest posterior probability. K-means methods treating the ordinal outcomes as the continuous data were applied on the dataset and results were compared with those from the
proposed models. In the Figure 3.5, the last representations of the latent variable of each biomarker in MCMC iterations were plotted while the dashed lines represent the cutpoint parameters. The group membership are presented using different colors and types of points. The average posterior probability of each patient being identified in certain group is reflected in the size of each point. The mixture model using RJMCMC successfully partitioned patients into four distinct groups.

We further applied our model to cluster three biomarkers by including FAS. A total of 30,000 Gibbs/RJMCMC iterations were run. The possible size of mixtures included 2, 3, 4, 5, 6, and 7 with the marginal posterior probabilities of 0.045, 0.399, 0.438, 0.095, 0.02, and 0.002, respectively (Figure 3.6). The split and the merge steps occurred at 2.8% and 2.3% of the iterations. In the Figure 3.7, the distributions of data is displayed in a 3-D plot, with different types and colors of points for different clusters. The size of the points is proportional to the number of patients at each coordinate. There were 17 patients whose expression levels were 0 for all three biomarkers. When $G = 3$ and $G = 4$, there were 51 and 33 patients in each cluster. According to the clustering analysis results, patients with lower expression levels tended to be classified to be in one group.
3.5 Conclusions

Compared to the K-means method, there are several advantages of this model. First of all, it is a probability-based classification algorithm, which allows data to be classified into clusters based upon the posterior membership probabilities. On the contrary, the K-means method uses an ad-hoc distance measure for classifications. Secondly, the K-Means method provides no assistance in determining the number of clusters, in other words, it needs to be pre-specified. In contrast, our clustering schema can determine the number of clusters by treating it as a random variable. Furthermore, K-means method is limited to interval scale quantitative variables, for which Euclidean distance measures can be calculated. I.e., the ordinal data is assumed to be interval data in use of K-means method. Instead, our model can be generalized to be performed on variables of mixed metrics. Variables may be continuous, categorical (nominal or ordinal), or counts or any combination of these. Prior to performing the K-means clustering, variables must be standardized to have equal variance to avoid obtaining clusters that are dominated by variables having the largest amounts of variations. However, our solution is invariant of linear transformations on the variables; thus, the standardization of variables is not necessary. Lastly, a common practice following a K-means clustering is to use discriminant analysis to describe differences among the clusters on one or more exogenous variables. In contrast, our model can be easily extended to include exogenous variables (covariates). This allows both classification and cluster description to be performed simultaneously.
So far, we have concentrated our work on the clustering of ordinal outcomes. Further expansions of our current work can include the following aspects. First of all, our model assumes that there is only one measurement for each subject, and hence, does not take into account the variability in repeated measurements. However, in some biomarker study, each individual may be subject to several measurements for the biomarker intensity. In addition to the multivariate normal distribution for each cluster, another layer of multivariate normal distributions is assumed to model the repeated measurements. Secondly, there are two directions of clustering analysis. One is to cluster the study subjects and the other is to cluster the ordinal measurements. For example, one might be interested in knowing which biomarkers are closer to each other as opposed to which subjects are close to each other. Our model focuses on the latter. Further research can be pursued in the clustering of biomarkers. Third, we can also incorporate the clustering analysis with principal component analysis or variable selections. Under some circumstances that there are many ordinal outcomes, there is no clear sign regarding which one is the important factor. The investigator may be like to reduce the dimension of the variables prior to performing any clustering analyses.
Table 3.1: Marginal posterior estimation of the number clusters and the corresponding clustering membership. The number of corrected estimated components is bolded under each scenario.

<table>
<thead>
<tr>
<th>True Cluster</th>
<th>G=3</th>
<th>G=4</th>
<th>G=5</th>
<th>G=6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (n=63)</td>
<td>63</td>
<td>0</td>
<td>63</td>
<td>0</td>
</tr>
<tr>
<td>2 (n=73)</td>
<td>0</td>
<td>71</td>
<td>0</td>
<td>71</td>
</tr>
<tr>
<td>3 (n=64)</td>
<td>0</td>
<td>0</td>
<td>64</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 3.2: Distributions of biomarker expression levels. The percentage inside the parenthesis is the column percentage.

<table>
<thead>
<tr>
<th>Expression</th>
<th>Caspase</th>
<th>FUS1</th>
<th>FAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>45 (53.6%)</td>
<td>33 (39.3%)</td>
<td>46 (54.8%)</td>
</tr>
<tr>
<td>1</td>
<td>16 (19.0%)</td>
<td>8 (9.5%)</td>
<td>12 (14.3%)</td>
</tr>
<tr>
<td>2</td>
<td>13 (15.5%)</td>
<td>13 (15.5%)</td>
<td>9 (10.7%)</td>
</tr>
<tr>
<td>3</td>
<td>10 (11.9%)</td>
<td>30 (35.7%)</td>
<td>17 (20.3%)</td>
</tr>
</tbody>
</table>
Figure 3.1: Marginal posterior distribution of the size of mixture
Figure 3.2: Distribution of latent variables represented by clusters under different sizes of mixture of bivariate distributions. The number at the top of each panel indicates the size of mixture. The different types and colors of points stand for different clusters. The size of points reflects the posterior probability of being classified into that cluster. Data is identified to be consisted of 3 (top two panels) or 4 (bottom two panels) clusters.
Figure 3.3: Convergence plots for the clustering of two biomarkers: ergodic estimates of $G$, solid line denotes $G=4$. 
Figure 3.4: Clustering of patients based on Caspase and Fus1. Left: clustering using K-means; Right: mixture model of bivariate normal distributions using RJMCMC to identify the size of clusters.
Figure 3.5: Convergence plots for the clustering of three biomarkers: ergodic estimates of $G$, solid line denotes $G=3$. 
Figure 3.6: Clustering analysis of patients conditional on three biomarker expression levels: Group 1: n=51; Group 2: n=33;
Chapter 4

A Hierarchical Model for Ordinal Data Nested in Categorical Data

4.1 Introduction

We discuss modeling and inference for data that includes categorical outcomes and ordinal outcomes nested within each level of the categorical variable. The motivating application is to model adverse event (toxicity) data in clinical trials. Toxicity type and severity are usually recorded as categorical and ordinal outcome, respectively. In a randomized phase III study, in addition to the efficacy of the study agent, investigators and regulators are also interested in learning about the toxicity profile of the study agent. Traditionally, simple descriptive statistics such as cross-tabulations are provided. However, this purely descriptive approach fails to offer an in-depth under-
standing of how the treatment affects the toxicity type and the severity associated with a specific type of toxicity.

The multinomial probit (MNP) model [1] and the multinomial logit model are popular model choices for implementing regression for categorical outcomes. The computational burden associated with implementing full posterior inference hinders the routine application of this model in applied work. In recent years, there have been some advances in computationally practical methods, using classical and Bayesian approaches. For example, the method of simulated moments by McFadden [47], the method of simulated scores by Hajivassiliou and McFadden [31], and Gibbs sampling with data augmentation as discussed in Albert and Chib [2] and McCulloch and Rossi [45], have made the required computations in the multinomial probit model practical.

For inference with ordinal data, many authors have proposed methods using both a classical approach [44] and a Bayesian framework [2], [22], [17]. A natural way to model ordinal data is to introduce an underlying continuous latent variable. The ordinal outcome is linked with the latent variable through a set of cutpoints. The probability of an ordinal outcome is represented as the probability of the latent continuous variable falling into an interval on the real line defined by the cutpoints. A normal latent variable leads to the probit model.

The estimation of cutpoint parameters associated with ordinal data and latent vari-
ables is a critical challenge. Albert and Chib [2] proposed an ordinal probit model in a Bayesian framework, including a vague prior on the cutpoint parameters. However, poor convergence of MCMC posterior simulation using the Gibbs sampler, i.e., iterative draws from the full conditional posterior distributions, can be a concern when the sample size is large. The MCMC output suffers from high autocorrelation. To mitigate this problem, Cowles [16] proposed a hybrid Gibbs/Metropolis-Hastings sampling scheme which updates the cutpoint parameters jointly with the other parameters by using a multivariate proposal. This approach reduces the high auto-correlation and achieves practical convergence within a reasonable number of iterations of the MCMC simulation. However, for a multivariate ordinal response, excessively many cutpoint parameters are required to be specified. For example, let $K_j$ denote the number of categories for the $j^{th}$ ordinal outcome. The probit model requires $K_j - 2$ cutpoints for each ordinal variable. In addition, the computation of the acceptance probability in the Metropolis-Hasting step requires the computation of multivariate normal quantiles (proved in the second chapter), leading to a prohibitive computational effort.

We propose a mixture model which can model ordinal data without the need to estimate cutpoint parameters. We show that in the proposed mixture model, the cutoff points can be fixed without loss of generality. An important feature of the approach is the joint modeling of categorical data with nested ordinal outcome. Such data formats frequently arise in clinical trial and health outcomes research. The rest of the paper is organized as follows. In Section 4.2, we introduce a phase III clinical trial. In
Section 4.3, we present a joint multinomial and ordinal probit model to estimate the cell probabilities of multiple categorical outcomes with different ordinal levels nested in each categorical outcome. The prior specifications and posterior inference are discussed in Section 4.4. We illustrate properties of the model by applying the model to a simulated dataset and a phase III clinical trial. The results are presented in Section 4.5. A summary and discussion of possible extensions are presented in Section 4.6.

4.2 A Phase III Clinical Trial

Studies have suggested that retinoid chemoprevention may help control second primary tumor, recurrence, and mortality for stage I non-small-cell lung cancer (NSCLC) patients. A National Cancer Institute (NCI) intergroup phase III trial of 1166 patients with pathologic stage I NSCLC was conducted to validate the efficacy of isotretinoin, a retinoid hypothesized to have chemopreventive properties. Patients were randomly assigned to receive either placebo or isotretinoin (30 mg/day) for 3 years in a double-blinded study. Patients were stratified at randomization by tumor stage, histology, and smoking status. A total of 589 patients received isotretinoin.

One of the objectives of this study is to assess the treatment effect on different toxicities and the grade associated with each of them. In this paper we focus on inference related to this objective only. The treatment-related toxicities include: cheilitis, conjunctivitis, arthralgia, hypertriglyceridemia, headache, and abnormal vision. Cheilitis
is dryness, usually associated with an uncomfortable sensation of the lips with scaling and cracking and accompanied by a characteristic burning sensation. Conjunctivitis is one of the most common nontraumatic eye complaints, involving the inflammation of conjunctiva. Arthralgia is pain in the joints. With the exception of hypertriglyceridemia, toxicity was graded by use of the Common Toxicity Criteria (NCI). Triglyceride toxicity was graded as follows: grade 1 toxicity was defined as more than 2.5 times but less than or equal to five times the normal level; grade 2 toxicity was defined as more than five times but less than or equal to 10 times the normal level; and grade 3 toxicity was defined as more than 10 times the normal triglyceride level or if a patient experienced complications (e.g., pancreatitis) at any grade of triglyceride toxicity. If patients experienced multiple incidents of the same toxicity, only one incident at the highest grade was counted. If multiple different toxicities are reported on one patient, they are recorded as separate data points. Since equal numbers of patients were randomized to isotretinoin and placebo groups, it remains meaningful to compare incident rates across treatment arms. A summary of the recorded toxicities is provided in Table 4.1.
4.3 Hierarchical Model for Ordinal Data Nested in Categorical Data

For each recorded adverse event, the data reports two variables. The first variable is a multinomial outcome, \( y_i \in \{1, \ldots, J\} \), indicating the type of toxicity. The second variable is an ordinal outcome, \( z_{ij} \), which reports the grade at which the \( j^{th} \) categorical outcome was observed. The quantity \( z_{ij} \) can take on values \( k = 1, \ldots, K_j \). Thus, the observation \((y_i = j, z_{ij} = k)\), means that the \( i^{th} \) toxicity was of type \( j \) and grade \( k \). Let \( X \) be a \((N \times H)\) matrix of possible regressors, with the \( i^{th} \) row, \( x_i \), recording \( H \) covariates for the \( i^{th} \) adverse event. In our example, focusing on assessing the treatment effect, we assume \( x_{i1} \) is an indicator for isotretinoin, i.e., for the \( i^{th} \) observed toxicity being recorded for a patient who was assigned experimental therapy. With an additional intercept, we have \( H = 2 \) and \( x_i^T = (1, x_{i1}) \).

To model the multinomial outcomes \( y_i \), we define a vector of latent variables, \( U_i^T = (u_{i1}, u_{i2}, \ldots, u_{iJ}) \).

\[
U_i = \gamma_0 + x_{i1} \gamma_1 + \epsilon_i, \quad \epsilon_i \sim MVN(0, I_J),
\]

where \( \gamma_0 \) and \( \gamma_1 \) are \( J \)-dimensional parameter vectors. The vector \( \gamma_0 = (\gamma_{01}, \ldots, \gamma_{0J}) \) is a vector of category specific parameters, \( \gamma_1 = (\gamma_{11}, \ldots, \gamma_{1J}) \) is the set of regression parameters. Denote \( \gamma_j = (\gamma_{0j}, \gamma_{1j}) \). The residuals \( \epsilon_i^T = (\epsilon_1, \ldots, \epsilon_J) \) follow a \( J \)-dimension multivariate normal distribution. The probability of \( y_i \) taking the value \( j \)
is defined as the probability of the $j^{th}$ latent variable being the maximum.

$$y_i = j \quad \text{if} \quad u_{ij} \geq u_{ij'} \quad \text{for} \quad j \neq j', \quad j, j' \in 1, \ldots, J.$$  \hspace{1cm} (4.1)

Conditional on $y_i = j$, the relationship between covariates and the nested ordinal outcome $z_{ij}$ is modeled through a latent variable $v_{ij}$. We set up an ordinal probit regression for $z_{ij}$ on covariates $x_i$. The probability $p(z_{ij} = k)$ is represented as the probability that the continuous latent variable $v_{ij}$ falls into the interval $[\theta_{k-1,j}, \theta_{k,j})$.

Multiple cutpoints are required for the $K_j$ ordinal outcomes:

$$z_{ij} = k \quad \text{if} \quad \theta_{k-1,j} \leq v_{ij} < \theta_{k,j} \quad \text{for} \quad k = 1, \ldots, K_j.$$  \hspace{1cm} (4.2)

$$v_{ij} = \beta_0 + x_i \beta_1 + \xi_{ij}, \quad \xi_{ij} \sim \sum_{g=1}^{G_j} p_{jg} \mathcal{N}(\mu_{jg}, \sigma_\xi^2)$$

Here, $\beta_j = (\beta_{0j}, \beta_{1j})$ parameterize the ordinal probit model for $z_{ij}$. Instead of assuming a normal latent variable $v_{ij}$, we impose a mixture of normal distributions with distinct location parameters $(\beta_{0j} + x_i \beta_{1j} + \mu_{jg})$, fixed number of components ($G_j$) and fixed weights ($p_{jg}$). It can be shown that without loss of generality, we can fix the cutpoints $\theta$s. See also the discussion below.

The mixture model can alternatively be written as a hierarchical model by introducing a latent indicator variable $w_{ij}$. Specifically, conditional on $w_{ij} = g$, the latent variable $v_{ij}$ follows a normal distribution $f(v_{ij} \mid w_{ij} = g) = \mathcal{N}(\beta_{0j} + x_i \beta_{1j} + \mu_{jg}, \sigma_\xi^2)$. The prior probability of $w_{ij} = g$ is the weight for each component, $p(w_{ij} = g) = p_{jg}$.

Let $\Phi(.)$ denote the standard normal cdf. Marginalizing the latent variable $v_{ij}$, we
have:

\[
P(z_{ij} = k \mid w_{ij} = g) = \Phi \left( \frac{\theta_{k,j} - x_i^T \beta_j - \mu_{ij}}{\sigma_\xi} \right) - \Phi \left( \frac{\theta_{k-1,j} - x_i^T \beta_j - \mu_{ij}}{\sigma_\xi} \right)
\]

(4.3)

\[
P(w_{ij} = g) = p_{jg}
\]

In summary, the probability of being in the level \( k \) of the category \( j \) is the joint probability of two continuous variables falling into a defined region in the parameter space:

\[
\pi_{jk} \equiv p(y_i = j, z_{ij} = k) = p(u_{ij} \geq u_{ij'}, j \neq j', \theta_{k-1,j} \leq \nu_{ij} < \theta_{k,j})
\]

For reasons of identifiability, we restrict the variance of the normal kernels in (4.2) to \( \sigma_\xi^2 = 1 \).

Recall that \( \{y_i = j\} = \{u_{ij} \geq u_{ij'}, j \neq j', j, j' = 1, \ldots, J\} \) which in turn can be written as \( \{-\infty < u_{ij} < b_{ij}\} \) with

\[
b_{ij} = \begin{cases} +\infty & \text{if } j' = j \\ u_{ij} & \text{if } j' \neq j \end{cases}
\]

Let \( N(x; m, s) \) indicate a normal pdf with parameters \((m, s)\) for the random variable, \( x \). We denote \( L_i \) as the likelihood for the \( i^{th} \) recorded adverse event. Assuming independence across events, the overall likelihood is a product, \( L = \Pi_{i=1}^n L_i \). Let \( \omega \) denote the vector of all model parameters.

\[
L_i = p(y_i = j \mid \omega)p(z_{ij} = k \mid \omega, y_i = j)
\]

\[
= \left\{ \int_{-\infty}^{b_{ij}} \cdots \int_{-\infty}^{b_{ij}} f(u_{i1}, \ldots, u_{ij}) du_{i1} \cdots du_{ij} \right\} \times
\]

\[
\left\{ \sum_{g=1}^{G_j} p_{jg} (\Phi (\theta_{k,j} - \beta_{0j} - x_i \beta_{ij} - \mu_{jg}) - \Phi (\theta_{k-1,j} - \beta_{0j} - x_i \beta_{ij} - \mu_{jg})) \right\}
\]
where,

\[ f(u_{i1}, \ldots, u_{iJ}) \sim f(u_{i1})f(u_{i2}) \cdots f(u_{iJ}) \]
\[ \sim N(u_{i1}; x_i^T \gamma_1, 1)N(u_{i2}; x_i^T \gamma_2, 1) \cdots N(u_{iJ}; x_i^T \gamma_J, 1) \]

4.4 Priors and Posteriors

4.4.1 Prior Elicitations

We use conjugate priors for the probit regression parameters, centering the prior to represent the prior judgment about the marginal prevalence of the outcomes and the effects of the covariates. See Section 5 for specific choices in an example. We use

\[ f(\gamma_j) \sim MVN(m_{\gamma_j}, \Sigma_{\gamma}) \]
\[ f(\beta_j) \sim MVN(m_{\beta_j}, \Sigma_{\beta}). \]

As default choice for \( G_j \), size of the mixture model in (2), we recommend \( G_j = K_j - 1 \). The prior probability of \( w_{ij} \) is the weight for each component in the mixture model. With respect to the location parameter \( (\mu_{jg}) \) for the components of the mixture of normals in (2), we use independent normal priors \( f(\mu_{jg}) \sim N(\phi, \sigma_{\mu}^2) \) with a conjugate hyper prior \( f(\phi) \sim N(0, 10^6) \).

In the dose-toxicity study introduced in Section 2, the investigators are interested in knowing how different dose levels affect the toxicity grade. For a cytotoxic agent, it is usually assumed that a higher dose incurs worse toxicity. The parameter \( \beta_{i1} \), the dose effect on the toxicity grade, may be restricted to be positive when there
are only two dose levels and the lower dose group is the reference group. In the case
where there are \(M\) dose groups, \(\beta_{ij}\) becomes an \((M-1)\)-dimensional vector. To enforce
monotonicity of toxicity with increasing dose, we introduce an ordering constraint on
\(\beta_j\) as follows. Assuming that the lowest dose is the reference group and the highest
dose group is group \(M\), monotonicity can be represented as \(\beta_{(M-1)j} \geq \beta_{(M-2)j} \geq \ldots \geq \beta_{ij}\). This assumption guarantees \textit{a priori} that a higher dose incurs worse toxicity.
The hierarchical prior can accommodate the intrinsic dependence structures across
different dose groups.

### 4.4.2 Full Conditional Distributions

Since we adopt conjugate priors for all parameters, all full conditional posterior dis-
tributions have tractable close forms. Random variable generation of parameters in
the Gibbs sampler is straightforward. Define \(u_{i(-j)}\) to be \(U_i\) excluding \(u_{ij}\).

\[
f(u_{ij} \mid u_{i(-j)}, \gamma_j, y_i) \propto \begin{cases} 
N(x_i^T \gamma_j, 1)I((-\infty, u_{ij}'), & \text{if } y_i = j' \neq j \\
N(x_i^T \gamma_j, 1)I(\max(u_{i(-j)}), \infty), & \text{if } y_i = j 
\end{cases}
\]

\[
f(v_{ij} \mid \beta_j, \mu_{jg}, z_{ij} = k, y_i = j) \propto N(x_i^T \beta_j + \mu_{jg}, 1)I(\theta_{k-1,j} \leq z_{ij} < \theta_{k,j})
\]

The complete conditional posterior distributions for \(\gamma_j\) and \(\beta_j\) conditional on, in
particular, the imputed latent variable, \(u_{ij}\) and \(v_{ij}\), are the posterior distributions
for the linear regressions, \(u_{ij} = x_i^T \gamma_j + \epsilon_{ij}\) and \(v_{ij} = x_i^T \beta_j + \xi_{ij}\), respectively. The
latent indicator variable \(w_{ij}\) in equation (4.3) is sampled from the following complete
conditional posterior distribution.

\[ p(w_{ij} = g \mid \beta_j, v_{ij}, \mu_{jg}, p_{jg}, g = 1, \ldots, G_j) \propto p_{ijg}, \]

with \( p_{ijg} = p_{jg}N(\xi_{ij}; \mu_{jg}, 1)/\sum_{g=1}^{G_j} p_{jg}N(\xi_{ij}; \mu_{jg}, 1) \). The complete conditional posterior distributions for \( \mu_{jg} \) and \( \phi \) are:

\[
\begin{align*}
  f(\mu_{jg} \mid \phi, w_{ij}, v_{ij}) & \propto N \left( \frac{\phi/\sigma_{\phi}^2 + \sum_{g=1}^{G_j} w_{ij}I(w_{ij} = g)}{1/\sigma_{\mu}^2 + \sum_{i=1}^{n} I(w_{ij} = g)}, \left(1/\sigma_{\mu}^2 + \sum_{i=1}^{n} I(w_{ij} = g)\right)^{-1} \right) \\
  f(\phi \mid \mu_{jg}, G_j) & \propto N \left( \frac{\sum_{i=1}^{n} \mu_{ij}}{1/\sigma_{\phi}^2 + \sum_{i=1}^{G_j} \sum_{j=1}^{n} I(w_{ij} = g)^{-1}}, \left(1/\sigma_{\phi}^2 + \sum_{i=1}^{G_j} \sum_{j=1}^{n} G_j \right)^{-1} \right)
\end{align*}
\]

### 4.4.3 Marginal Posterior Probability

Let \( D \) be the data. The marginal posterior cell probabilities are estimated from Gibbs sampling outputs. Let \( \omega^{(r)} \) denote the imputed parameter vector after \( r \) iterations of the Gibbs sampler. We report

\[
\begin{align*}
  \pi_{jk}(y_i = j, z_{ij} = k) &= E[(p(y_i = j, z_{ij} = k \mid \omega)) \mid D] \\
  &= E[p(u_{ij} \max \{u_{ij} \mid \omega\}) \mid \omega] \\
  &\approx \frac{1}{R} \sum_{r=1}^{R} p(y_i = j, z_{ij} = k \mid \omega^{(r)}) \\
  &= \frac{1}{R} \sum_{r=1}^{R} I(u_{ij} \max \{u_{ij} \mid \omega^{(r)}\}) \\
  &\times p(\theta_{k-1,j} \leq \theta_{k,j} \mid \omega^{(r)})
\end{align*}
\]

where \( R \) is the total number of MCMC iterations retained after an initial burn-in.
4.5 Applications

4.5.1 A Simulated Dataset

A simulated dataset was generated to illustrate the model. A total of 400 subjects were assigned into groups A or B with equal probability. There were four \((J = 4)\) possible categorical outcomes. Associated with each categorical outcome, there was an ordinal outcome with three possible levels \(K_j = 3\) for \(j = 1, \ldots, 4\). In the simulation study, the latent variable \(U\), related to the categorical outcome, was generated from the following normal distributions:

\[
u_{ij} \sim N(\gamma_{0j} + x_{i1} \gamma_{1j}, 1), \quad \text{for } i = 1, \ldots, 400 \text{ and } j = 1, \ldots, 4
\]

\[
\gamma_0^T = (0 \ 0 \ .5 \ .5), \quad \gamma_1^T = (1 \ 1 \ .5 \ 1.5),
\]

where \(x_{i1}\) represented the group membership and was generated from a Bernoulli distribution with success probability .5. Group A was used as the reference group, i.e., \(x_{i1} = 0\) for the \(i^{th}\) observation from the group A. The parameter \(\gamma_{ij}\) represented the group effect for the categorical outcome \(j\). The categorical outcome for each subject was selected as described in the equation (4.1). The latent continuous variables \(v_{ij}\) related to the ordinal levels nested within the categorical outcome \(j\) were drawn from a mixture of two Gaussian distributions:

\[
v_{ij} \sim .5N(x_{i1} \beta_{1j}, 1) + .5N(x_{i1} \beta_{1j} + 2, 1), \quad \text{for } i = 1, \ldots, 400 \text{ and } j = 1, \ldots, 4
\]

\[
\beta_1^T = (1 \ .5 \ .8 \ 1.2).
\]
The latent continuous variables were classified into three groups within each categorical group with cutpoints \( \{0, 1.5\} \).

\[
\begin{align*}
z_{ij} = & \quad 0 \text{ if } v_{ij} < 0 \\
1 \text{ if } 0 \leq v_{ij} < 1.5 \\
2 \text{ if } v_{ij} \geq 1.5
\end{align*}
\]

The simulated data is summarized in Table 4.2 below.

The prior means for the \( \gamma_0 \) and \( \gamma_1 \) were set to be 0 with the prior variance of 100. To investigate conjugate prior sensitivity, we considered several alternative prior choices. We found that the posterior estimations were quite robust to the prior specifications.

The ordinal level outcome within each category group was modeled through a probit model with a mixture of \( G=2 \) normal distributions with equal weights. The cutpoints were fixed at \( \theta_1 = 0 \) and \( \theta_2 = 4 \). Following Gelman, Carlin, Stern and Rubin [28], a diffuse hyper-prior was imposed on the \( \beta_j \)'s. A total of 100,000 iterations of MCMC were simulated after a 10,000 burn-in period. The Gibbs sampling outputs were saved at each 10th iteration. The marginal posterior probabilities of each cell were computed and compared with the true cell probabilities (Table 4.3). The model yielded reasonable estimates of the cell probabilities.

### 4.5.2 A Phase III Clinical Trial of Retinoid Isotretinoin

We applied our model to the phase III clinical trial introduced in Section 2. We chose an informative prior with \( m_{\gamma_0} = (0.1 \ 0.1 \ 0.1 \ 0.1 \ 0.1 \ 0.1) \) and \( m_{\gamma} = (0 \ 0.3 \ 0.8 \ 0.7 \ 0 \ 0) \)
0) to reflect a prior knowledge of the incidence rate of each type of toxicity. The prior implies that the incidence rates are .05 for all types of toxicity in the placebo group. In the treatment group, the toxicity rates for abnormal vision, arthralgia, cheilitis, conjunctivitis, fatigue, headache, and hypertriglyceride were a priori estimated as .05, .1, .3, .2, .05, .05, .05. The size of the mixture was pre-defined as \( G = 3 \) with equal weight associated with each component. A diffuse hyperprior centered at 0 with the variance of \( 10^6 \) was imposed on \( \phi \). The cutoff points were chosen following the default choices. A total of 100,000 Gibbs sampling iterations were run to estimate marginal posterior cell probabilities. Table 4.4 displays the estimated cell probabilities together with 95% credible intervals. The estimates formally confirm what is expected from inspection of the data. Only one incidence of grade 4 abnormal vision was observed in the placebo group. There were more incidences of cheilitis and conjunctivitis observed in the isotretinoin group than in the placebo group. Elevated triglyceride levels also were found more frequently in the isotretinoin group than in the placebo group. However, more patients experienced headache in the placebo group.

The posterior probabilities of observing more incidences in the isotretinoin group compared to the placebo group were computed for each toxicity-grade (Table 4.5). There was almost a 100% chance that the isotretinoin group caused more grade 2 or 3 cheilitis than the placebo group. Moreover, the isotretinoin group also caused more grade 2/3 conjunctivitis than the placebo group did. However, there was only a .2% chance to observe more grade 1 conjunctivitis in the isotretinoin group compared to
that in the placebo group. More incidences of headache were observed in the placebo
group as the probabilities of observing more grade 1 headache in the isotretinoin
group than the placebo group were 0. The marginal posterior distributions of $\beta$s are
presented in Figure 4.1. There was over a 99% chance that the treatment (isotretinoin)
had a positive effect on arthralgia, cheilitis, and conjunctivitis.

4.6 Conclusions

We have proposed a Bayesian hierarchical model to analyze ordinal data nested within
categorical data. Our model characterizes the ordinal/categorical data structure by
a joint multinomial and ordinal probit model. We provide posterior summaries to
assess treatment effects. In the phase III clinical trial example, traditional analysis
might simply group the toxicities into two levels: 1+2 and 3+4, and then, applying
a Chi-squared test or Fisher's exact test to compare the two treatment groups for a
particular toxicity type instead. We first model different types of toxicities jointly in
the framework of multinomial probit model. Next, our model treats the toxicity grade
as ordinal data in compliance with the natural structure of the data. Our approach
has several advantages compared to Albert and Chib's uniform prior and Cowles'hybrid Gibbs/Metropolis sampling scheme. First, we no longer have to worry about
the convergence of the cutpoint parameters within the MCMC context. Updating the
latent outcomes is simplified. Second, our model reduces the number of parameters
that must be estimated. Lastly, the nature of the mixture model allow us to model
data from heterogeneous sources. Therefore, it might provide a better estimate of cell probabilities in many biomedical datasets.

One critical issue is the choice of the size of the mixture in modeling ordinal outcomes. One could use reversible jump MCMC to decide the size of the mixture. Alternatively, a simple rule of thumb is that the size of mixture, $G$, is heterogeneity of data, hence, fail to model the cell probabilities. However, large $G$ may over-fit the model.

This model also has interesting applications in other areas such as health outcomes research and clinical trial design. For example, some studies have shown that even when treatments are known to be effective, many patients who could benefit from them are not getting them. Beta blocker medication, given after heart attacks, can reduce mortality; blood-thinning medication can prevent stroke; and thrombolytic therapy given immediately after a heart attack can reduce the damage from the attack.

The outcome instrument has focused on assessing the overall level of functioning after receiving the treatment conditional on patients' prognostic characteristics. The overall level of functioning is a quantified variable on an ordinal scale. Therefore, by assessing the ordinal outcomes within each category, the outcome researchers will be able to identify and address the barriers to better care, eventually; translate the findings into practical strategies to improve care.

One of the basic assumptions of our model is the experimental unit. The occurrence of a single toxicity as opposed to each individual patient in the isotretinoin study is
viewed as the experimental unit. This is a reasonable assumption in a clinical trial study since this is the way data is recorded. However, there might be within-subject correlations for the cases where patients experienced multiple types of toxicities. The within-subject correlations can be easily handled by imposing a random effect in the model if patient identifiers are recorded.
Table 4.1: Toxicity frequency for randomized eligible patients by study arms.

<table>
<thead>
<tr>
<th>Toxic effect</th>
<th>Placebo</th>
<th>Isotretinoin</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>G1  G2  G3  G4</td>
<td>G1  G2  G3  G4</td>
</tr>
<tr>
<td>Abnormal vision</td>
<td>10  1  2  1</td>
<td>8  2  1  0</td>
</tr>
<tr>
<td>Arthralgia</td>
<td>26  10 0  0</td>
<td>36 12 5  0</td>
</tr>
<tr>
<td>Cheilitis</td>
<td>83  8  0  0</td>
<td>341 125 10 0</td>
</tr>
<tr>
<td>Conjunctivitis</td>
<td>45  4  1  0</td>
<td>122 34 11 0</td>
</tr>
<tr>
<td>Fatigue</td>
<td>14  6  2  0</td>
<td>14 3 0 0</td>
</tr>
<tr>
<td>Headache</td>
<td>16  3  4  0</td>
<td>9 0 0 0</td>
</tr>
<tr>
<td>Hyper-triglyceride</td>
<td>25  4  0  0</td>
<td>74 10 1 0</td>
</tr>
</tbody>
</table>

Table 4.2: The generated data sorted by the group membership.

<table>
<thead>
<tr>
<th>Treatment A Category</th>
<th>Treatment B Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>ordinal level</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1  2  3  4</td>
</tr>
<tr>
<td>2</td>
<td>17 14 7 14</td>
</tr>
<tr>
<td>3</td>
<td>11 22 21 27</td>
</tr>
<tr>
<td></td>
<td>1  2  3  4</td>
</tr>
<tr>
<td></td>
<td>3  19 2 2</td>
</tr>
<tr>
<td></td>
<td>8  30 9 8</td>
</tr>
<tr>
<td></td>
<td>25 49 20 33</td>
</tr>
</tbody>
</table>
Table 4.3: Simulated dataset, marginal posterior cell probabilities with 95% credible intervals. In bold are the empirical cell probabilities.

<table>
<thead>
<tr>
<th>ordinal level</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.078(0.084)</td>
<td>0.081(0.069)</td>
<td>0.036(0.034)</td>
<td>0.071(0.069)</td>
</tr>
<tr>
<td></td>
<td>(0.048, 0.112)</td>
<td>(0.051, 0.117)</td>
<td>(0.016, 0.065)</td>
<td>(0.041, 0.110)</td>
</tr>
<tr>
<td>2</td>
<td>0.052(0.054)</td>
<td>0.097(0.11)</td>
<td>0.107(0.10)</td>
<td>0.127(0.13)</td>
</tr>
<tr>
<td></td>
<td>(0.026, 0.085)</td>
<td>(0.062, 0.138)</td>
<td>(0.068, 0.154)</td>
<td>(0.079, 0.178)</td>
</tr>
<tr>
<td>3</td>
<td>0.050(0.054)</td>
<td>0.107(0.099)</td>
<td>0.095(0.094)</td>
<td>0.098(0.099)</td>
</tr>
<tr>
<td></td>
<td>(0.025, 0.085)</td>
<td>(0.074, 0.146)</td>
<td>(0.063, 0.130)</td>
<td>(0.068, 0.133)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ordinal level</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.014(0.015)</td>
<td>0.089(0.096)</td>
<td>0.007(0.010)</td>
<td>0.008(0.010)</td>
</tr>
<tr>
<td></td>
<td>(0.003, 0.033)</td>
<td>(0.058, 0.123)</td>
<td>(0.001, 0.02)</td>
<td>(0.001, 0.019)</td>
</tr>
<tr>
<td>2</td>
<td>0.036(0.041)</td>
<td>0.17(0.15)</td>
<td>0.049(0.046)</td>
<td>0.044(0.041)</td>
</tr>
<tr>
<td></td>
<td>(0.016, 0.062)</td>
<td>(0.127, 0.220)</td>
<td>(0.028, 0.077)</td>
<td>(0.026, 0.068)</td>
</tr>
<tr>
<td>3</td>
<td>0.063(0.071)</td>
<td>0.25(0.25)</td>
<td>0.103(0.10)</td>
<td>0.16(0.17)</td>
</tr>
<tr>
<td></td>
<td>(0.031, 0.102)</td>
<td>(0.197, 0.308)</td>
<td>(0.062, 0.154)</td>
<td>(0.105, 0.233)</td>
</tr>
</tbody>
</table>
Table 4.4: Marginal posterior cell probabilities (95% credible intervals) of toxicity.

<table>
<thead>
<tr>
<th>Toxic effect</th>
<th>G1</th>
<th>G2</th>
<th>G3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abnormal vision</td>
<td>.01 (.004, .017)</td>
<td>.001 (.000, .003)</td>
<td>.002 (001, .004)</td>
</tr>
<tr>
<td>Arthralgia</td>
<td>.02 (.01, .04)</td>
<td>.009 (.05, .01)</td>
<td>.000 (.000, .002)</td>
</tr>
<tr>
<td>Cheilitis</td>
<td>.08 (.06, .09)</td>
<td>.008 (.004, .01)</td>
<td>.000 (.000, .000)</td>
</tr>
<tr>
<td>Conjunctivitis</td>
<td>.04 (.03, .05)</td>
<td>.007 (.003, .01)</td>
<td>.000 (.000, .000)</td>
</tr>
<tr>
<td>Fatigue</td>
<td>.01 (.007, .02)</td>
<td>.006 (.003, .01)</td>
<td>.002 (.000, .005)</td>
</tr>
<tr>
<td>Headache</td>
<td>.01 (.008, .02)</td>
<td>.003 (.001, .007)</td>
<td>.003 (.001, .007)</td>
</tr>
<tr>
<td>Hyper-triglyceride</td>
<td>.02 (.01, .04)</td>
<td>.004 (.001, .008)</td>
<td>.000 (.000, .000)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Isotretinoin</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Abnormal vision</td>
<td>.01 (.002, .014)</td>
<td>.002 (.0002, .004)</td>
<td>.001 (.000, .003)</td>
</tr>
<tr>
<td>Arthralgia</td>
<td>.03 (.02, .04)</td>
<td>.01 (.006, .02)</td>
<td>.005 (.001, .01)</td>
</tr>
<tr>
<td>Cheilitis</td>
<td>.30 (.27, .32)</td>
<td>.14 (.13, .16)</td>
<td>.000 (.000, .0002)</td>
</tr>
<tr>
<td>Conjunctivitis</td>
<td>.11 (.09, .13)</td>
<td>.03 (.02, .04)</td>
<td>.01 (.005, .02)</td>
</tr>
<tr>
<td>Fatigue</td>
<td>.01 (.005, .02)</td>
<td>.003 (.001, .006)</td>
<td>.000 (.000, .000)</td>
</tr>
<tr>
<td>Headache</td>
<td>.01 (.003, .02)</td>
<td>.000 (.000, .002)</td>
<td>.000 (.000, .000)</td>
</tr>
<tr>
<td>Hyper-triglyceride</td>
<td>.07 (.05, .09)</td>
<td>.01 (.01, .02)</td>
<td>.000 (.000, .000)</td>
</tr>
</tbody>
</table>
Table 4.5: Posterior probabilities of particular events: $p_{1i}$ and $p_{1p}$ represent the marginal posterior probability of grade 1 toxicity in the isotretinoin group and placebo group, respectively. Likewise, the subscripts 2 and 3 in the other two columns correspond to grade 2 and grade 3 toxicity.

<table>
<thead>
<tr>
<th>Toxic effect</th>
<th>$p(p_{1i} &gt; p_{1p})$</th>
<th>$p(p_{2i} &gt; p_{2p})$</th>
<th>$p(p_{3i} &gt; p_{3p})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abnormal vision</td>
<td>0.71</td>
<td>0.71</td>
<td>0.083</td>
</tr>
<tr>
<td>Arthralgia</td>
<td>0.17</td>
<td>0.25</td>
<td>1.00</td>
</tr>
<tr>
<td>Cheilitis</td>
<td>0.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Conjunctivitis</td>
<td>0.002</td>
<td>0.93</td>
<td>1.00</td>
</tr>
<tr>
<td>Fatigue</td>
<td>0.997</td>
<td>0.13</td>
<td>0.00</td>
</tr>
<tr>
<td>Headache</td>
<td>1.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Hyper-triglyceride</td>
<td>0.51</td>
<td>0.49</td>
<td>0.46</td>
</tr>
</tbody>
</table>
Figure 4.1: Marginal posterior distributions of the treatment effect on ordinal outcomes ($\beta$s)
Chapter 5

Summary and Extensions

5.1 Summary

The current challenge of ordinal data modeling is how to estimate the cutpoint parameters which link the continuous latent variables and ordinal outcomes. As we have discussed before, the uniform prior algorithm proposed by Albert & Chib has the problem of the slow convergence of MCMC. The hybrid Gibbs/Metropolis-Hasting algorithm proposed by Cowles might be intractable due to the complexity of the acceptance of probability of Metropolis-Hasting (MH) in the setting of repeated measurements. Therefore, the goal of the research is to find an efficient alternative in estimating the cutpoint parameters. This dissertation achieves this goal by proposing a mixture model for the underlying latent variables, where the size of mixture is considered as random variable. Therefore, the cutpoint parameters can be fixed
without loss of generality.

There are two key contributions of this research. First of all, we propose a mixture model which can capture the ordinality of data while keeping the cutpoint parameters constant. The results in Chapter 2 have shown that the convergence of MCMC using Albert & Chib algorithm is slow when the sample size is large. The computation of cumulative probabilities of multivariate normal distributions is required and prohibitive in calculating the acceptance probabilities of MH using Cowles' method. Therefore, neither of methods is applicable in modeling ordinal longitudinal data. Our mixture model can not only eliminate the slow convergence issue, but also model data from heterogeneous resources. Furthermore, the use of reversible jump MCMC in estimating the size of mixture also provides an automatic process in finding the most appropriate model. This unsupervised process allows the model to characterize data structure without the concerns of underestimating or over-fitting data.

Second, we provide a model-based algorithm in clustering ordinal data. We assume that all the component distributions conditional on the cluster membership belong to the same parametric family. With each cluster, the interval group variability may be expected to behave as random and non-correlated error. The number of cluster can be estimated through reversible jump MCMC. The model-based approach has several advantages over heuristic clustering methods such as K-means method. First, it clusters ordinal outcomes and estimates component parameters simultaneously. Second,
it provides clustering uncertainties which is important especially for objects close to
cluster boundaries. Third, the problems of determining the number of components
and the component probability distributions can be recast as statistical model selec-
tion problems. The results presented in Chapter 3 have shown that the model-based
algorithm does have appealing features.

5.2 Extensions

In longitudinal studies we do not necessarily require the same number of observations
on each subject or those measurements to be taken at the same time points. Therefore,
this mixture model can also be extended to model data where it is measured repeated
over random time points instead of fixed time points in our study. In such a case, a
mixture of linear mixed effect model can be applied on the latent variables, where the
residual terms follow a mixture Gaussian distributions. The variability within sub-
jects is modeled by random-effect terms. This model can also handle unequal number
of repeated observations over random time points within each subject. Missing data
is another issue in modeling longitudinal data. If some responses for some subjects
are missing, we may discard incomplete records to produce a balanced dataset, which
relies on a very strong assumption that the missing values are missing completely at
random (MCAR). However, sometimes, the probability of an observation being miss-
ing may depend on observed data for the subject, including covariates and observed
responses, which is called missing at random (MAR). For instance, in the clinical tri-
als, the follow-up visit may depend on the toxicity of the previous treatment. If the treatment has a moderate side-effect on patients, in consequent, substantial number of patients may quit the study due to the toxicity. The mechanism of handling missing data can be incorporated into our model. Due to the structure of this model, EM algorithm to impute the missing data can easily be implemented in MCMC without any further efforts.

With respect to clustering analysis, there are two aspects of clustering. One is to cluster the study subjects and the other is to cluster the ordinal measurements. For example, one might be interested in knowing which biomarkers are closer to each other as opposed to which subjects are close to each other. Our model focuses on the latter. Further research can be pursued in the clustering of biomarkers. We can also incorporate the clustering analysis with principal component analysis or variable selections. Under some circumstances that there are many ordinal outcomes, there is no clear sign regarding which one is the important factor. The investigator may be like to reduce the dimension of the variables prior to performing any clustering analyses. Additionally it is also a challenge to cluster various types of data such as a mix of ordinal data, binary data, and continuous data. Our model can be easily extended to accommodate different data structure as well.

Another aspect of the extension of our research is the hierarchical model structure. Take an example of a clinical trial studying response status of four subtype of disease
after receiving the treatment. The response status is categorized as complete response (CR), partial response (PR), stable disease (SD), and progression (PR). Before the trial starts, little knowledge is known about the response rate of each subtype of disease group. Therefore, the assumption of exchangeability may hold. We can put a hyper prior on the location parameters for different subgroups. This structure allows us to borrow strength in estimating response rate across different groups. However, as more patients are enrolled into study, more information about the response rate is obtained. The exchangeability assumption may be no longer valid for certain subgroups. For example, two groups have very high response rate and the other groups have very low groups. A hyper prior 'borrowing strength' across all groups is no longer appropriate in such a case. A Dirichlet process on the location parameters of latent variables may behave better in terms of identifying 'exchangeable' group. It will identify the groups which behave alike and allow them to share more information.
Bibliography


