Nonparametric Bayesian Methods for Multiple Imputation of Large Scale Incomplete Categorical Data in Panel Studies

by

Yajuan Si

Department of Statistical Science
Duke University

Date: __________________________

Approved:

______________________________
Jerome P Reiter, Supervisor

______________________________
Alan E Gelfand

______________________________
Fan Li

______________________________
Vincent Joseph Hotz

Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistical Science in the Graduate School of Duke University 2012
ABSTRACT

Nonparametric Bayesian Methods for Multiple Imputation of Large Scale Incomplete Categorical Data in Panel Studies

by

Yajuan Si

Department of Statistical Science
Duke University

Date: __________________________

Approved:

______________________________
Jerome P Reiter, Supervisor

______________________________
Alan E Gelfand

______________________________
Fan Li

______________________________
Vincent Joseph Hotz

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistical Science in the Graduate School of Duke University 2012
Abstract

The thesis develops nonparametric Bayesian models to handle incomplete categorical variables in data sets with high dimension using the framework of multiple imputation. It presents methods for ignorable missing data in cross-sectional studies, and potentially non-ignorable missing data in panel studies with refreshment samples.

The first contribution is a fully Bayesian, joint modeling approach of multiple imputation for categorical data based on Dirichlet process mixtures of multinomial distributions. The approach automatically models complex dependencies while being computationally expedient. I illustrate repeated sampling properties of the approach using simulated data. This approach offers better performance than default chained equations methods, which are often used in such settings. I apply the methodology to impute missing background data in the 2007 Trends in International Mathematics and Science Study.

For the second contribution, I extend the nonparametric Bayesian imputation engine to consider a mix of potentially non-ignorable attrition and ignorable item nonresponse in multiple wave panel studies. Ignoring the attrition in models for panel data can result in biased inference if the reason for attrition is systematic and related to the missing values. Panel data alone cannot estimate the attrition effect without untestable assumptions about the missing data mechanism. Refreshment samples offer an extra data source that can be utilized to estimate the attrition effect while reducing reliance on strong assumptions of the missing data mechanism.
I consider two novel Bayesian approaches to handle the attrition and item non-response simultaneously under multiple imputation in a two wave panel with one refreshment sample when the variables involved are categorical and high dimensional.

First, I present a semi-parametric selection model that includes an additive non-ignorable attrition model with main effects of all variables, including demographic variables and outcome measures in wave 1 and wave 2. The survey variables are modeled jointly using Bayesian mixture of multinomial distributions. I develop the posterior computation algorithms for the semi-parametric selection model under different prior choices for the regression coefficients in the attrition model.

Second, I propose two Bayesian pattern mixture models for this scenario that use latent classes to model the dependency among the variables and the attrition. I develop a dependent Bayesian latent pattern mixture model for which variables are modeled via latent classes and attrition is treated as a covariate in the class allocation weights. And, I develop a joint Bayesian latent pattern mixture model, for which attrition and variables are modeled jointly via latent classes. I show via simulation studies that the pattern mixture models can recover true parameter estimates, even when inferences based on the panel alone are biased from attrition.

I apply both the selection and pattern mixture models to data from the 2007-2008 Associated Press/Yahoo News election panel study.
to my mom
Contents

Abstract iv
List of Tables xi
List of Figures xiii
List of Abbreviations and Symbols xv
Acknowledgements xvi

1 Introduction 1

1.1 Missing Data Mechanisms . . . . . . . . . . . . . . . . . . . . . . . . 3
1.2 Selection and Pattern Mixture Models . . . . . . . . . . . . . . . . . 4
1.3 Multiple Imputation . . . . . . . . . . . . . . . . . . . . . . . . . . . 7
1.4 Nonparametric Bayesian Methods . . . . . . . . . . . . . . . . . . . . 10
1.5 Outline . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 13

2 Nonparametric Bayesian Multiple Imputation for Incomplete Categorical Variables in Large-Scale Surveys 15

2.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 15
2.2 Review of Missing Categorical Data Methods . . . . . . . . . . . . . . 17
2.3 Mixture Models for Multiple Imputation . . . . . . . . . . . . . . . . . 20
  2.3.1 Finite Mixture of Products of Multinomial Distributions . . . . . . . 21
  2.3.2 Infinite Mixture of Products of Multinomial Distributions . . . . . . 23
2.4 Posterior Computation and Inference . . . . . . . . . . . . . . . . . . . 25
2.4.1 Blocked Gibbs Sampler ...................... 25
2.4.2 Slice Sampling Algorithm ..................... 27
2.5 Simulation Studies of Frequentist Performance ............. 29
    2.5.1 Case 1: Small p and Log-linear Model Data Generation .... 29
    2.5.2 Case 2: Large p and Mixture Model Data Generation ....... 33
2.6 Imputation of TIMSS Background Variables .................. 34
2.7 Conclusions ..................................... 40

3 Semi-parametric Selection Models for Potentially Non-ignorable Attrition in Panel Study with Refreshment Sample 41
    3.1 Introduction ..................................... 41
    3.2 Methods ......................................... 44
        3.2.1 Attrition Model ................................. 46
        3.2.2 Prior on $\beta$ ............................... 47
        3.2.3 Model for Survey Variables ..................... 50
    3.3 Posterior Computations ........................... 51
    3.4 Simulation Study .................................. 59
        3.4.1 Case 1: Logistic Regression .................... 60
        3.4.2 Case 2: Bayesian Probit Auxiliary Variable Models ...... 63
        3.4.3 Case 3: Regularized Probit Auxiliary Variable Models .... 64
    3.5 AP Yahoo News Panel Data Analysis .................. 66
        3.5.1 Discussions on Semi-parametric Models for the APYN Data 82
    3.6 Conclusions ..................................... 86

4 Bayesian Latent Pattern Mixture Models in Panel Study with Refreshment Samples 91
    4.1 Introduction ..................................... 91
    4.2 Pattern Mixture Models and Data Description ................. 92
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.2.1</td>
<td>Pattern Mixture Models without Refreshment Samples</td>
<td>92</td>
</tr>
<tr>
<td>4.2.2</td>
<td>Pattern Mixture Models in Two Wave Panel with One Refreshment Sample</td>
<td>94</td>
</tr>
<tr>
<td>4.2.3</td>
<td>AP Yahoo News Data</td>
<td>96</td>
</tr>
<tr>
<td>4.3</td>
<td>Latent Class Models</td>
<td>98</td>
</tr>
<tr>
<td>4.4</td>
<td>Bayesian Latent Pattern Mixture Models</td>
<td>100</td>
</tr>
<tr>
<td>4.4.1</td>
<td>Intuitive Compare of Dependent BLPM and Joint BLPM</td>
<td>101</td>
</tr>
<tr>
<td>4.5</td>
<td>BLPM under CI Assumption</td>
<td>103</td>
</tr>
<tr>
<td>4.5.1</td>
<td>Dependent BLPM</td>
<td>103</td>
</tr>
<tr>
<td>4.5.2</td>
<td>Joint BLPM</td>
<td>106</td>
</tr>
<tr>
<td>4.5.3</td>
<td>Simulation Study</td>
<td>107</td>
</tr>
<tr>
<td>4.5.4</td>
<td>APYN Data Analysis</td>
<td>113</td>
</tr>
<tr>
<td>4.6</td>
<td>Release the CI Assumption for BLPM</td>
<td>125</td>
</tr>
<tr>
<td>4.6.1</td>
<td>BLPM Model Specification</td>
<td>125</td>
</tr>
<tr>
<td>4.6.2</td>
<td>Simulation Study</td>
<td>128</td>
</tr>
<tr>
<td>4.6.3</td>
<td>APYN Data Analysis</td>
<td>129</td>
</tr>
<tr>
<td>4.7</td>
<td>Conclusions</td>
<td>136</td>
</tr>
<tr>
<td>5</td>
<td>Current research and future directions</td>
<td>140</td>
</tr>
<tr>
<td>A</td>
<td>Appendix to Chapter 2 - Hierarchical DP Mixture of Products of Multinomial Distributions</td>
<td>143</td>
</tr>
<tr>
<td>A.0.1</td>
<td>Model Specification</td>
<td>143</td>
</tr>
<tr>
<td>A.0.2</td>
<td>Posterior Computation for HDPMPM</td>
<td>145</td>
</tr>
<tr>
<td>B</td>
<td>Appendix to Chapter 4 - Proof for Theorem 1</td>
<td>148</td>
</tr>
<tr>
<td>C</td>
<td>Appendix to Chapter 4 - Posterior Computation Algorithm for Dependent BLPM under CI Assumption</td>
<td>150</td>
</tr>
<tr>
<td>D</td>
<td>Appendix to Chapter 4 - Posterior Computation Algorithm for Joint BLPM under CI Assumption</td>
<td>152</td>
</tr>
<tr>
<td>Section</td>
<td>Title</td>
<td>Page</td>
</tr>
<tr>
<td>---------</td>
<td>----------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>E</td>
<td>Appendix to Chapter 4 - Posterior Computation Algorithm for Dependent BLPM</td>
<td>154</td>
</tr>
<tr>
<td>F</td>
<td>Appendix to Chapter 4 - Posterior Computation Algorithm for Joint BLPM</td>
<td>157</td>
</tr>
<tr>
<td></td>
<td>Bibliography</td>
<td>159</td>
</tr>
<tr>
<td></td>
<td>Biography</td>
<td>169</td>
</tr>
</tbody>
</table>
## List of Tables

3.1 Graphical representation for panel and refreshment sample........... 46
3.2 The marginal and conditional probability for semi-parametric model simulation study ............................................. 64
3.3 The joint probability for semi-parametric model simulation study .. 65
3.4 Collected measures for selection models ............................... 69
3.5 Collected demographic variables for selection models ............... 70
3.6 Estimates of coefficients on demographic variables for attrition model 73
3.7 Estimates of coefficients on outcome variables for attrition model .. 74
3.8 The marginal and conditional probability for the APYN study in selection models .................................................. 76
3.9 The joint probability for the APYN study in selection models ...... 76
3.10 Estimation for model W~LV31 in selection models ................. 82
3.11 Estimation for model W~LV3 in selection models ..................... 83
3.12 Estimation for model W~CND1 in selection models ................. 83
3.13 Estimation for model LV1~ X₁ in selection models .................... 84
3.14 Estimation for model LV3~ X₁ in selection models .................... 85
3.15 Estimation for model LV31~ X₁ in selection models ................. 86
3.16 Estimation for model CND1~ X₁ in selection models ............... 87
3.17 Estimation for model FAV1-4~ X₂ in selection models ............... 88
3.18 Estimation for model FAV1-4~ X₂ in selection models - cont. ....... 89
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1 A simple two wave panel</td>
<td>94</td>
</tr>
<tr>
<td>4.2 Directly estimated terms for the simple example</td>
<td>94</td>
</tr>
<tr>
<td>4.3 Collected measures for BLPM models</td>
<td>97</td>
</tr>
<tr>
<td>4.4 Collected demographic variables for BLPM models</td>
<td>98</td>
</tr>
<tr>
<td>4.5 The marginal and conditional probability using joint BLPM under CI for the APYN study</td>
<td>124</td>
</tr>
<tr>
<td>4.6 The marginal and conditional probability using joint BLPM for the APYN study</td>
<td>136</td>
</tr>
<tr>
<td>4.7 The joint probability using joint BLPM for the APYN study</td>
<td>137</td>
</tr>
</tbody>
</table>
List of Figures

2.1 Small $p$ simulation results ........................................ 32
2.2 Large $p$ simulation results ........................................ 35
2.3 Frequency distributions of $ppp$ for the TIMSS imputation. ....... 39
3.1 Simulation outputs with logit link .................................. 62
3.2 Simulation outputs with probit link ................................ 63
3.3 Simulation outputs for probit under Bayesian Lasso ............. 67
3.4 The posterior mean estimates for $\beta$ in the probit attrition model ... 72
3.5 Probability comparison between panel data and refreshment sample of selection models .......................... 77
3.6 Comparison on marginal probability for $Y_2$ between stayed and dropout samples in panel of selection models .......... 78
3.7 Comparison on regression coefficients between completed panel and refreshment sample after MI ........................................ 81
3.8 Posterior mean estimates for $\beta$ under Bayesian Lasso for APYN study 90
4.1 Simulation outputs in Case 1 of dependent BLPM under CI ....... 109
4.2 Simulation outputs in Case 2 of dependent BLPM under CI ....... 111
4.3 Simulation outputs in Case 1 of joint BLPM under CI .......... 114
4.4 Simulation outputs in Case 2 of joint BLPM under CI .......... 115
4.5 Comparison on marginal probability for $Y_2$ between stayed and dropout samples in panel using dependent BLPM under CI .......... 118
4.6 Analysis comparison between completed panel and refreshment sample after MI using dependent BLPM under CI .......... 119
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.7</td>
<td>Comparison on marginal probability for $Y_2$ between stayed and dropout samples in completed panel using joint BLPM under CI</td>
<td>121</td>
</tr>
<tr>
<td>4.8</td>
<td>Analysis comparison between completed panel and refreshment sample after MI using joint BLPM under CI</td>
<td>122</td>
</tr>
<tr>
<td>4.9</td>
<td>Probability comparison between completed panel data and refreshment sample after MI using joint BLPM under CI</td>
<td>123</td>
</tr>
<tr>
<td>4.10</td>
<td>Simulation outputs of joint BLPM</td>
<td>130</td>
</tr>
<tr>
<td>4.11</td>
<td>Analysis comparison between completed panel and refreshment sample after MI using joint BLPM for APYN study</td>
<td>133</td>
</tr>
<tr>
<td>4.12</td>
<td>Comparison on marginal probability for $Y_2$ between stayed and dropout samples in completed panel using joint BLPM for APYN study</td>
<td>134</td>
</tr>
<tr>
<td>4.13</td>
<td>Probability comparison between completed panel data and refreshment sample using joint BLPM for APYN study</td>
<td>135</td>
</tr>
<tr>
<td>4.14</td>
<td>Persistency check between wave 1 and wave 2 measures</td>
<td>138</td>
</tr>
<tr>
<td>4.15</td>
<td>Estimates for selection models under MI</td>
<td>139</td>
</tr>
</tbody>
</table>
List of Abbreviations and Symbols

Symbols

- $Z$ Collected data set.
- $Y_1$ Collected outcome measures in wave 1.
- $Y_2$ Collected outcome measures in wave 2.
- $X$ Collected background variables.
- $W$ Attrition indicator.

Abbreviations

- MI Multiple Imputation.
- MCAR Missing completely at random.
- MAR Missing at random.
- MCMC Markov Chain Monte Carlo.
- DP Dirichlet process.
- AN Additive non-ignorable.
- DPM Dirichlet process mixture.
- DDP Dependent Dirichlet process.
- CI Conditional independence.
- BLM Bayesian latent pattern mixture.
Acknowledgements

I have been enjoying the experience at Duke for the past four years, and thank to all who share the wonderful memories with me in this friendly environment.

First of all, I would like to express my deepest gratitudes to my advisor Dr. Jerry Reiter for being incredibly supportive through my doctoral studies. Thank you Jerry, you have been a knowledge advisor and a terrific mentor to me. Without Jerry, I would have had a hard time developing from a student to a researcher. I appreciate and admire his patience and enthusiasm in nurturing step by step. Throughout the countless inspirations I learn from Jerry, his tremendous work ethics, rigorous attitudes, extraordinary passion and profound ingenuity towards research always impress me. Also as an encouraging colleague and a trusted friend, Jerry has been remarkably generous and caring. I am very honored to have Jerry as my role model for high standards of excellence to be a bright and honest academician.

Special thanks to my dissertation committee members, Dr. Alan Gelfand, Dr. Fan Li and Dr. V. Joseph Hotz for their continuous support and encouragement. Alan’s thoughts on my research help me understand everything better, and his advice on my career development will benefit me for the whole life. Thanks to Fan, for providing much help on research directions and beyond. I appreciate the encouragement and time from Joe for interesting discussions and extensions of my work. I also appreciate the help from my colleague Dr. Sunshine Hillygus for our enjoyable meetings on the panel data.
Further thanks to Dr. Alan Gelfand, Dr. Jim Berger, Dr. Robert Wolpert, Dr. Michael Reed, Dr. David Banks, Dr. David Dunson, Dr. Merlise Clyde, Dr. Scott Schmidler, and Dr. Fan Li, who teach me helpful courses at Duke and provide wonderful advise on research along the way. Dr. Dalene Stangle is my first year advisor, and I sincerely am grateful for her mentorship. Thanks a lot for the help on my job application materials, communications of academic experience and career advice from Dr. Li Ma and Dr. Kari Lock.

I also want to thank the helpful stuff in the department, Karen, Nikki, Anne and Lance, whose work with lots of efforts and kindness is sincerely appreciable.

It has been my privilege to be funded by the National Science Foundation for my research work.

My fellow graduate students make my life at Duke colorful. Thanks to all my wonderful friends, who help me through study and life, and share my sweets and bitters. I will cherish the fond memories with them.

I am greatly fortunate and indebted to my terrific husband, Wenshuo Liu, who has been there for me all along the time. We go through everything together, and his love and support make my life much sweeter.

Lastly but most importantly, to my parents, and my two little sisters, thank you for all the support and the love. No words can describe how proud I am of you. Without you, it would not even have been possible for today. I dedicate this work to my dearest mom, who could not wait on my filial piety.
Across many disciplines, large scale panel surveys are popular tools to study individuals’ attitudes and behaviors over time. While panel surveys offer rich information, they can suffer from missing data. Item nonresponse occurs when sampled individuals do not provide answers for some questions in the survey. Item nonresponse is a common problem in large scale surveys that include an extensive set of questions. Survey participants tend to ignore questions especially when the survey has low stakes, e.g., there is no direct benefit or punishment for the individuals to answer or omit the questions. Unit nonresponse occurs when the survey fails to collect any information from the sampled individuals, unit nonresponse occurs. In panel studies, a particularly relevant form of unit nonresponse is attrition, also called dropout. This occurs when the same individuals are interviewed repeatedly in panel studies, but some drop out for the follow-up waves after previous participation. Both the item nonresponse and unit nonresponse reduce effective sample sizes and may lead to inference bias if the reason for the missing data is systematic.

In large scale surveys, typical solutions for handling missing values have shortcomings. Treating missing data as another category for multiple choice questions
distorts the original data structure. Using complete cases can lead to only a few available cases that may be unrepresentative. Using available-case analysis with re-weighting is difficult since construction of reasonable weights is challenging. Plugging in means or modes underestimates uncertainty and can lead to biased inference. Related problems exist when using the last observation carried forward in longitudinal studies. Such single imputation methods ignore the correlation among items and fail to capture the uncertainty induced by missing data.

An alternative is model-based imputation, which is attractive with the development of advanced computation techniques. Analysts estimate models directly from the data using data augmentation to generate plausible imputed values. Bayesian imputation procedures are especially useful here. They iteratively update model parameters and missing values, for example, via Gibbs sampling, thus generating posterior samples that approximately are draws from the true conditional distributions on observed data.

In panel studies with attrition, when only the incomplete panel data are available, untestable assumptions have to be made on the missing data mechanism or the attrition process. To reduce reliance on such assumptions, panel data can be supplemented by new units randomly sampled from the same population, called refreshment samples. A refreshment sample includes new, randomly sampled respondents who are given the same questionnaire at the same time as a second or subsequent wave of the panel. Hirano et al. (2001) show that refreshment samples provide information that can be used to assess the effects of panel attrition and to correct for biases via statistical modeling. However, refreshment samples have not been fully utilized for statistical modeling. Typically analysts only use the external source for exploratory checks or comparisons.

The main contribution of this thesis is to develop flexible and efficient imputation approaches featuring nonparametric Bayesian models for large scale incomplete cat-
egorical data and handle the non-ignorable attrition in multiple wave panel studies with refreshment samples. In the remainder of this chapter, I review missing data methodologies and introduce the motivation and main research questions of this thesis. Section 1.1 reviews the classification of missing data mechanisms. Section 1.2 describes two typical models for non-ignorable likelihood. Section 1.3 introduces the multiple imputation approach and its correspondence to Bayesian data augmentation under ignorable likelihood. Section 1.4 describes the attractive features of nonparametric Bayesian methods for imputation. Section 1.5 provides the thesis framework.

1.1 Missing Data Mechanisms

Rubin (1976) formalizes the concepts of missing mechanism, which I briefly review here. Suppose the collected dataset is $Z$ with $N$ individuals and $p$ variables. Similar to Little and Rubin (2002), denote the missing-data indicator matrix as $M = (M_{ij})$ for individual $i$ and variable $j$, for $i = 1, \ldots, N$ and $j = 1, \ldots, p$. Here, $M_{ij} = 1$ if the individual $i$ did not provide response for the $j$th variable, and $M_{ij} = 0$ if this value is observed. Denote the observed part of $Z$ as $Z_{obs} = \{Z_{ij}, (i, j) : M_{ij} = 0\}$ and the missing values $Z_{mis} = \{Z_{ij}, (i, j) : M_{ij} = 1\}$. If the missingness does not depend on the values of the data $Z$, no matter missing or observed, it is called missing completely at random (MCAR); that is, $f(M|Z, \phi) = f(M|\phi)$ for all $Z$ and $\phi$, where $\phi$ is the corresponding parameter for the missingness model. The case missing at random (MAR) arises when missingness depends only on the components that are observed, $f(M|Z, \phi) = f(M|Z_{obs}, \phi)$ for all $Z_{mis}$ and $\phi$. The mechanism is called not missing at random (NMAR) if missingness also depends on the missing values $f(M|Z, \phi) = f(M|Z_{obs}, Z_{mis}, \phi)$.

MCAR is too ideal for unintentional missing data, and MAR is a common assumption for the missing data mechanism. However, sometimes NMAR is the case,
especially in panel studies. The collected data $Z$ typically include demographic information $X$ and outcome measures $Y$. It is often assumed as MAR if the individuals refuse to provide the demographic information or omit part of the outcome measures. As the panel moves on, it is often believed as NMAR when some individuals drop out of panel; that is, the reason for the attrition is related to their nonresponse in later waves. Hence, in practice and in this thesis, item nonresponse is treated as MAR and unit nonresponse is assumed as NMAR.

Missing data mechanisms can be extended. Roy (2003) discusses conditional MAR but with marginal NMAR: given a latent class, the missing data mechanism only depends on the observed data. However, after integrating the condition out, the missing data mechanism also depends on the missing values. Lin et al. (2004) describe a similar construction called conditional independence but with marginal dependence: given a latent class, the missing data mechanism is independent of the missing values. After integrating the condition out, the missing data mechanism is dependent on the missing values. In this thesis, I investigate both the traditional missing data mechanism assumptions (Rubin, 1976; Little and Rubin, 2002) and the more general missing data mechanisms (Roy, 2003; Lin et al., 2004).

1.2 Selection and Pattern Mixture Models

Rubin (1976) shows that if the missing data mechanism is MAR or MCAR, and the parameters for missingness and the parameters for the data are distinct, then the missing data mechanism is ignorable. The likelihood function only depends on the models of data. However, when the data are NMAR, we have to model the missing data mechanism for likelihood inference. Selection models and pattern mixture models are the two common choices for modeling non-ignorable nonresponse. Take a two wave panel study with dropout occurring for the second wave as an example. Let $X$ denote the demographic variables that do not change with time, e.g.,
race and sex; let $Y_1$ denote the variables collected from the first wave of study; and let $Y_2$ denote variables collected from the second wave of study. Here, $Y_2$ are partially observed only for the individuals who still stay in the panel after participating in wave 1. Let $W = 1$ if the individual stays in the panel and $W = 0$ otherwise. Let $\theta$ represent model parameters for the data.

Selection models (Hausman and Wise, 1979; Kenward, 1998; Scharfstein et al., 1999) specify the joint distribution of $W$ and $(X,Y_1,Y_2)$ through models for the marginal distributions of $(X,Y_1,Y_2)$ and conditional distribution of $W$ given $(X,Y_1,Y_2)$. We have

$$f(X,Y_1,Y_2,W|\theta,\phi) = f(X,Y_1,Y_2|\theta)f(W|X,Y_1,Y_2,\phi).$$

Under NMAR, most literature on longitudinal data studies assumes $f(W|X,Y_1,Y_2,\phi) = f(W|X,Y_2,\phi)$. Under MAR, it assumes $f(W|X,Y_1,Y_2,\phi) = f(W|X,Y_1,\phi)$. It is impossible to have both $Y_2$ and $Y_1$ in the model if using non-informative priors because of identifiability problems.

Pattern mixture models (Glynn et al., 1986; Little, 1993; Roy and Daniels, 2008; Lin et al., 2004; Kenward et al., 2003) specify the joint distribution through the marginal distribution for $W$ and the conditional distribution of $(X,Y_1,Y_2)$ given $W$. We have

$$f(X,Y_1,Y_2,W|\theta,\phi) = f(X,Y_1,Y_2|W,\theta)f(W|\phi).$$

The analyst specifies two models: $f(X,Y_1,Y_2|W=1)$ and $f(X,Y_1,Y_2|W=0)$. However, there is no data to estimate $f(X,Y_1,Y_2|W=0)$. Extra constraints or analyst-determined parameters are necessary.

Hedeker and Gibbons (2006, Chapter 18) list the attractive features for selection models and pattern mixture models. Let $Z = (X,Y_1,Y_2)$. Selection models are a natural way of factoring the model, with $f_Z$ the model for completed data and
\(f_{W|Z}\) the model for the missing-data mechanism that determines what parts of \(Z\) are observed. Substantively, it seems more natural to consider relationships between \(Y = (Y_1, Y_2)\) and \(X\) in the full target population of interest rather than in subpopulation defined by missing-data pattern. Inferences for the parameters in the distribution \(f_Z\) are available directly from the selection model analysis. If the MAR assumption is plausible, the selection model formulation leads directly to the ignorable likelihood, which can be based solely on \(f_Z\). Under reasonable MAR assumptions and required inferences for the population aggregated over the missing-data pattern, the selection model is compelling.

Pattern mixture models have some desirable features when NMAR is suspected. The pattern mixture model formulation targets the conditional distribution of substantive interest. From an imputation perspective, \(Z_{mis}\) should be imputed from their predictive distribution given the observed data including \(W\), that is, \(f(Z_{mis}|Z_{obs}, W)\). If data are not MAR, the predictive distribution of \(Z_{mis}\) given \(Z_{obs}\) and \(W\) is modeled directly in the pattern mixture formulation, but it is related to the components of the selection model by the complex expression using Bayes rule to transfer the conditional distributions.

The more direct relationship between the pattern mixture formulation and the predictive distribution for imputations yields gains in transparency and computational simplicity in some situations, which is quite useful for variable selection and dimension reduction. The selection model factorization does not require full specification of the model for the missing-data mechanism when the data are MAR, but it does if the data are NMAR, in order for identification. Some pattern mixture models avoid specification of the model for the missing data mechanism in NMAR situations by using assumptions about the mechanism to yield restrictions on the model parameters.

In this thesis I utilize both selection and pattern mixture models for handling
attrition under the framework of multiple imputation, which I now summarize.

1.3 Multiple Imputation

Multiple imputation (MI) is a popular tool to deal with missing data problems (Rubin, 1976, 1986, 1987, 1996; Reiter and Raghunathan, 2007). The basic idea is to simulate values for the missing data repeatedly by sampling from their predictive distributions \( f(Z_{mis}|Z_{obs}) \). Usually \( f(Z_{mis}|Z_{obs}) \) is not in closed form and is difficult to sample directly. Data augmentation (Tanner and Wong, 1987) is able to generate the samples from the joint posterior distribution \( f(\theta, Z_{mis}|Z_{obs}) \) and hence from \( f(Z_{mis}|Z_{obs}) \).

In data augmentation, we iteratively sample \( \theta \) from the conditional distributions given \( Z_{com} = (Z_{obs}, Z_{mis}) \), and simulate \( Z_{mis} \) from the full conditional distributions given \( (\theta, Z_{obs}) \). Given generated values \( \theta^{(t)} \) and \( Z_{mis}^{(t)} \) at iteration \( t \), we have

\[
\text{I Step: } Z_{mis}^{(t+1)} \sim f(Z_{mis}|Z_{obs}, \theta^{(t)}) \quad (1.1)
\]

\[
\text{P Step: } \theta^{(t+1)} \sim f(\theta|Z_{obs}, Z_{mis}^{(t+1)}) \quad (1.2)
\]

This Markov Chain Monte Carlo (MCMC) chain can be viewed as a Gibbs sampler. After convergence, the posterior samples of \( \theta \) and \( Z_{mis} \) are treated as draws from their marginal posterior distributions \( f(\theta|Z_{obs}) \) and \( f(Z_{mis}|Z_{obs}) \).

To create Bayesian proper multiple imputation (Schafer, 1997), we have to subsample from the chain with large enough lag to collect \( m \) approximately independent samples. If the lag value is too small, the multiple imputations will be correlated and the missingness uncertainty will be underestimated. Running \( m \) parallel chains with different starting values can avoid this problem. However, if the starting values are over dispersed and the lengths of parallel chains are too small, the missingness uncertainty will be overestimated. To reduce computational burden, in this thesis, I use single MCMC chains with large iteration numbers and subsample multiple
completed datasets with large lag values.

Usually, a small number $m$ of multiple completed datasets can achieve efficiency. Schafer (1997) shows that, even with a very small $m$, the Monte Carlo error is relatively ignorable compared to the overall inferential uncertainty, and it can be explicitly accounted by the rules that combine the $m$ completed-data analyses.

Given the $m$ completed data sets, analysts can perform completed data analysis using standard techniques or software, and combine the point and variance estimates of interest from each imputed dataset by a simple set of formulas derived by Rubin (1987), called combining rules. Let $Q$ denote the scalar quantity of interest in the population, such as mean, regression coefficients or proportions. For $l = 1, \ldots, m$, denote the estimate of $Q$ as $q_l$ and its variance estimate as $u_l$ in the $l$th completed dataset. We calculate the quantities

\[ \bar{q}_m = \frac{\sum_{l=1}^{m} q_l}{m}, \quad \bar{u}_m = \frac{\sum_{l=1}^{m} u_l}{m}, \quad \text{and} \quad b_m = \sum_{l=1}^{m} (q_l - \bar{q}_m)^2 / (m-1). \]

Then, $\bar{q}_m$ is used as the mean estimate of $Q$, $\bar{u}_m$ is the within sampling variance, and $b_m$ represents the between imputation variance. Rubin’s combining rule (Rubin, 1987) provides the variance estimator

\[ T_m = (1 + 1/m) b_m + \bar{u}_m. \quad (1.3) \]

When $m$ is large, $Q$ follows a normal distribution with mean $\bar{q}_m$ and variance $T_m$. When $m$ is modest or small, the distribution of $Q$ is Student $t_{\nu_{df}}(\bar{q}_m, T_m)$ with degrees of freedom $\nu_{df} = \frac{(m-1)\left(1+\bar{u}_m/[(1+1/m)b_m]\right)^2}{(m-2)}$ (Rubin and Schenker, 1987). Barnard and Rubin (1999) adjust the degree of freedom for small sample size. The combining rules are derived from a Bayesian perspective under diffuse prior distributions and large-sample normality assumptions. Rubin (1987) shows conditions under which MI results in statistically valid frequentist randomization inference. Evaluations via repeated simulation studies have shown that MI leads to inferences that tend to be well calibrated from a frequentist standpoint. Si and Reiter (2011) compared the inference using direct posterior sampling and combining rules for MI. Rubin (1987)
also derived the combining rules for multivariate $Q$, and tests of significance for multicomponent null hypotheses are derived by Li et al. (1991), Meng and Rubin (1992) and Reiter (2007).

Reiter and Raghunathan (2007) describe the multiple adaptations of MI in recent developments. In addition to missing data problems, measurement error can be handled by MI (Brownstone and Valletta, 1996; Cole et al., 2006; Durrant and Skinner, 2006). MI has been extensively implemented to protect data confidentiality (Reiter, 2002, 2005, 2009), for example, full synthesis (Raghunathan et al., 2003), partial synthesis (Reiter, 2003, 2004) and sampling with synthesis (Drechsler and Reiter, 2010; Si and Reiter, 2010). The combining rules are modified to yield statistical validity. If the data records available for analysis are only part of the records used for imputation, Reiter (2008) proposed a two-stage MI approach and derived the corresponding combining rules for unbiased and valid inference.

As noted by Schafer (1997), inference by MI includes the two distinct imputation and analysis phases. This is one of the attractions of MI, for example, the missing data problems are handled by imputers who have more expertise in statistical techniques than the analysts who can just perform the standard analysis on completed datasets and combine the results. The combining rules are derived under the implicit assumptions of consistency between the imputation model and the analysis model. The quantity of interest $Q$ in the analysis model is not necessarily a function of parameters of the imputation model, and discrepancy may arise when the imputation and analysis models differ. With the violation of congeniality, whether the MI procedure is still valid has been investigated by Meng (1994) and Rubin (1996). If the analysis model has more assumptions than the imputation model, or the imputation model has more assumptions than the analysis model but the extra assumptions are true, MI is still able to offer valid inferences. If the extra assumptions during imputation are false, the resulting inference will be invalid. For the reason of congeniality,
the imputation model should be general and flexible enough to preserve possible correlations among all variables.

The condition stated by Rubin (1987) for proper MI is not restricted to any specific parametric models. This suggests potential benefits in the use of nonparametric Bayesian models to implement large and general imputation models. This can enhance the theoretical arguments for congeniality, and more importantly, facilitate practical computation with efficient algorithms.

1.4 Nonparametric Bayesian Methods

Nonparametric Bayesian models are commonly defined as probability models with infinitely many parameters (Bernardo and Smith, 2000). Nonparametric Bayesian models avoid restrictions of parametric assumptions, and as special cases, embed them in larger encompassing nonparametric models (Müller and Quintana, 2004). Müller and Quintana (2004) and Gershmana and Blei (2012) present overviews of nonparametric Bayesian inference. Relevant nonparametric Bayesian models include Dirichlet process (DP), Gaussian process, Polya trees, Chinese restaurant process, Indian buffet process, Beta process, wavelets, neural network, splines, classification and regression trees (CART) and their variations and extensions. They are typically applied to inference problems such as density estimation, clustering, regression, survival analysis and model validation. In this thesis, I use Dirichlet process mixture models (Antoniak, 1974). These have become popular across many application fields including econometrics (Chib and Hamilton, 2002; Hirano, 2002), social science (Kyung et al., 2010) and finance (Rodríguez and Dunson, 2011).

I use lower case $z_i$ to denote the observation for individual $i$. For independent and identically distributed samples $z_1, \ldots, z_N$, Bayesian nonparametric mixture models
assume the realizations are generated from a convolution

\[ z_i \sim \int k(\cdot | \theta) P(d\theta), \quad (1.4) \]

where \( k(\cdot | \theta) \) is a given parametric kernel distribution indexed by the parameter \( \theta \) and \( P \) is a mixing distribution. With a DP prior (Ferguson, 1973, 1974; Sethuraman, 1994) on \( P \), we obtain the well known Dirichlet process mixture (DPM) models (Escobar and West, 1995). Ferguson (1973) introduces the DP as a random probability measure that can perform as a prior distribution with large support and analytically manageable posterior inference. A random probability distribution \( P \) is generated by a DP, \( P \sim DP(\alpha, P_0) \), if for any fixed \( k \) and any measurable partition \( B_1, \ldots, B_k \) of the sample space we have \( (P(B_1), \ldots, P(B_k)) \sim \text{Dirichlet}(\alpha P_0(B_1), \ldots, \alpha P_0(B_k)) \).

Here \( \alpha \) is the precision parameter and \( P_0 \) is the base measure. The expectation is \( E(P(B)) = P_0(B) \) and the variance is \( V(P(B)) = P_0(B)(1 - P_0(B))/(1 + \alpha) \).

The DP is a distribution over distributions and has three main representations: Polya urn scheme (Blackwell and MacQueen, 1973), the Chinese restaurant process (Ishwaran and James, 2003) and the stick-breaking process (Sethuraman, 1994). The stick-breaking construction is useful for computation, and \( P \sim DP(\alpha, P_0) \) can be represented as

\[ P(\cdot) = \sum_{h=1}^{\infty} \pi_h \delta_{\theta_h}(\cdot), \quad \theta_h \overset{iid}{\sim} P_0 \quad (1.5) \]

\[ \pi_h = V_h \prod_{l<h}(1 - V_l), \quad V_h \overset{iid}{\sim} \text{Beta}(1, \alpha). \quad (1.6) \]

Here, the DP is represented as infinite mixtures of points masses and is an almost surely discrete distribution. However, the discreteness is inappropriate for many continuous distribution cases. This restriction can be released through the convolution (1.4) known as DPM models. The kernel functions can be Gaussian distributions (Lo,
1984; Escobar and West, 1995) for continuous data and multinomial distributions for categorical data (Dunson and Xing, 2009).

Posterior computations for the DPM models based on MCMC algorithms have been discussed by Neal (2000), MacEachern and Müller (1998) and West et al. (1994). The stick-breaking representation of DP facilitates efficient Gibbs updating algorithms. Ishwaran and James (2001) introduced the blocked Gibbs sampler under truncation of the stick breaking prior distributions. Walker (2007) proposed slice sampling algorithm for the posterior computation of DPM models by inducing a latent and uniformly distributed variable without truncation. Papaspiliopoulos and Roberts (2008) used retrospective MCMC method to sample from the exact posterior distributions. Papaspiliopoulos (2008) combined the characteristics of the retrospective sampler (Papaspiliopoulos and Roberts, 2008) and the slice sampler (Walker, 2007) and proposed exact blocked Gibbs sampler without predetermined truncation number of classes.

Many recent developments in nonparametric Bayesian methods have focused on inducing more flexibility of the locally different mixing distributions $P_h$ to share some common global feature and become dependent among the members. Typical approaches induce dependence in the weights $\pi_h$ and/or component specific parameters $\theta_h$ (MacEachern, 1999, 2000; Teh et al., 2006; Gelfand et al., 2005; Rodríguez et al., 2008). Hierarchical DP (Teh et al., 2006) and nested DP (Rodríguez et al., 2008) construct dependence structure across groups by assuming respectively that the base measure is also generated from another DP and that the atoms are distributions from independent Dirichlet processes. These extensions of DP for a family of a priori exchangeable distributions allow us to simultaneously cluster groups and cluster individuals within groups. MacEachern (2000) proposed a single-$\pi$ dependent DP by inducing covariates $x$ into the component specific distributions. The component specific parameters $\theta_h(x)$ form a random sample of stochastic processes
and follow some given distribution, while the weights $\pi$ have the same construction as DP. This single-$\pi$ dependent DP assumes the shapes of the conditional base measures $P_0(x)$ do not depend on $x$, and it does not afford sufficient flexibility or accommodate collections of independent distributions (MacEachern, 2000). To induce dependence structure in the weights $\pi_h$, a probit stick-breaking process can be used (Rodríguez and Dunson, 2011; Chung and Dunson, 2009) by replacing the Beta distribution in the stick-breaking process by probit links of normal random variables with covariates. The construction of weights is reminiscent of the continuation ratio probit model (Agresti, 2003) in survival analysis. These sequential ordinal models can be computed using data augmentation algorithm introduced by Albert and Chib (2001).

In this thesis, I use the DPM methods to construct the imputation model. I modify the dependent DPM models treating attrition indicator $W$ as a covariate to handle non-ignorable missing values.

1.5 Outline

The reminder of the thesis implements nonparametric Bayesian models to handle incomplete categorical variables of high dimension under the framework of multiple imputation, including ignorable missingness and potentially non-ignorable missingness mechanisms in panel studies. Chapter 2 starts with item non-response in categorical variables in cross sectional large scale surveys, which typically suffer from missing data problems. Standard methods of missing data analysis cannot handle the high dimensional scenario efficiently. I propose a fully Bayesian, joint modeling approach for multiple imputation of large scale categorical data using Dirichlet process mixture of multinomial distributions. This nonparametric Bayesian method flexibly captures the complex dependency structure while the regularization of the Dirichlet process prior facilitates efficient computation. I illustrate repeated sampling properties of
the approach using simulated data. I apply the methodology to impute missing student background data in the 2007 Trends in International Mathematics and Science Study (Foy and Olson, 2009).

With the nonparametric Bayesian imputation engine, I extend to consider a mix of potentially non-ignorable attrition and item nonresponse in multiple wave panel studies. In Chapter 3, I present semi-parametric selection models that include an additive non-ignorable model for attrition and a nonparametric Bayesian joint model for the collected variables. The posterior computation algorithms for the semi-parametric selections are developed under different prior choices for the regression coefficients in the attrition model. I use simulation studies to investigate the performance of the approach and then apply the semi-parametric selection model to the data from the 2007-2008 AP Yahoo News Presidential election panel.

In Chapter 4, I handle non-ignorable attrition with a pattern mixture model. I develop a dependent Bayesian latent pattern mixture model: the variables are modeled via latent classes and the attrition is treated as a covariate in the class allocation weights. I also develop a joint Bayesian latent pattern mixture model: the attrition and the variables are modeled jointly via latent classes. Using simulation studies, I first investigate the benefits of refreshment samples for these two models under the assumption of conditional independence. I then release the conditional independence assumption. I apply the models to the 2007-2008 AP Yahoo News panel data.

In Chapter 5, I discuss the ongoing work and further research directions. Extensions to multilevel categorical data using hierarchical Dirichlet process prior and the corresponding posterior computation algorithms for Bayesian pattern mixture models, are included in appendices at the end of the thesis.
2

Nonparametric Bayesian Multiple Imputation for Incomplete Categorical Variables in Large-Scale Surveys

2.1 Introduction

Large-scale surveys of educational progress are group-level assessments targeting policy relevant subgroups. Examples include the National Assessment of Educational Progress, the Programme for International Student Assessment and the Trends in International Mathematics and Science Study (TIMSS). In addition to cognitive data based on tasks that are designed to elicit what students know and can do, these national and international educational surveys collect a large number of background variables on students, such as their demographics, interests, activities and study habits. Background data on teachers and schools are also collected. Background data are relevant in multiple ways for data analysis, modeling and policy research. Background variables are used for deriving score files and for reporting in operational analysis. Literature shows that ignoring the effect of background variables may lead to biased inference (Thomas, 2002).
Background variables are often collected via self-reports, which make these measures fallible. Particularly, missing data are common in low stakes assessments of educational outcomes. Students who are asked to fill out tests and questionnaires without direct benefit sometimes choose to skip questions or tasks, or to leave out sections or groups of questions. In the case of student background data, students may choose to leave out questions due to reasons that do not directly relate to test motivation. For example, some questions may be considered sensitive or too personal for them to be willing to respond. Finally, time constraints of school based assessments lead to omission of responses. Students tend to spend more time on cognitive items and leave out the background questions. All these processes will lead to non-negligible amounts of missing responses. For example, for the Grade 4 students' profile in the 2007 TIMSS data file that I analyze, only 4385 out of 90505 students have complete records on a set of 80 background variables.

Missing observations also can be due to design issues, which systematically produce a number of non-observed responses for some or all students. The missing data also could be a mixture of structured missing and respondent-selected missing. The researchers and survey users desire to know more than what observed data can tell and what are hidden under the missing values.

In this chapter, I present a fully Bayesian, joint modeling approach to multiple imputation for high-dimensional categorical data. The approach is motivated by missing values among background variables in TIMSS. In such high dimensions (80 categorical variables), typical multiple imputation methods for categorical data, like log-linear models and sequential regression strategies (Raghunathan et al., 2001), can fail to capture complex dependencies and become difficult to implement effectively. I model the implied contingency table of the background variables as a mixture of independent multinomial distributions, estimating the mixture distributions nonparametrically with Dirichlet process prior distributions as in Dunson and
Xing (2009). This is related to the approach of Vermunt et al. (2008), who also use mixtures of multinomials for multiple imputation of categorical data. Their approach requires an \textit{ad hoc} selection of a fixed number of mixture components and uses repeated maximum likelihood estimation on bootstrapped samples to approximate draws of imputation model parameters. A fully Bayesian approach avoids both of these approximations while remaining computationally efficient.

The reminder of this chapter is organized as follows: Section 2.2 presents motivations for the need of the nonparametric imputation method. Section 2.3 describes latent class analysis and the nonparametric Bayes approach used for imputation. Section 2.4 provides the posterior computation algorithms. Section 2.5 describes simulations on different cases to investigate the performances of the proposed approach. Section 2.6 illustrates the imputations on the TIMSS data. Finally section 2.7 concludes with a discussion of further extensions.

2.2 Review of Missing Categorical Data Methods

For a limited number of categorical variables, one common choice is to treat the variables as continuous and assume that they follow multivariate normal distributions. This procedure is available in SAS PROC MI and MIANALYZE (Yuan, 2006) and in the R package NORM (Schafer, 1997). These use data augmentation to update the missing values and the parameters iteratively - the mean vector and covariance matrix for the multivariate normal distribution - and collect multiple datasets after convergence. Each collected non-integer value is rounded to be the nearest feasible integer, which is used as the imputation. Graham and Schafer (1999) show that MI with the multivariate normal model is rather robust to the violations of normality. However, imputation methods that treat the categorical data as continuous, e.g., as multivariate normal, can work well for some problems but are known to fail in others, even in low dimensions (Graham and Schafer, 1999; Ake, 2005; Finch, 2010; Yucel
et al., 2011). Allison (2000) and Horton et al. (2003) point that this rounding may lead to serious bias. Bernaards et al. (2007) proposed an adaptive rounding procedure to reduce the bias. Even though this multivariate normal imputation sometimes may perform well, it can work only for ordered categorical data, such as dichotomous or ordinal. Unordered or nominal categorical variables can be problematic.

Gaussian Latent Factor Models (Muthén, 1983) are routinely used for modeling of dependence and accommodating arbitrary mixtures of continuous, binary and ordered categorical data through an underlying Gaussian latent factor structure. Hence they are candidates for imputation as well. For unordered categorical variables, we link each observed value $z_{ij}$ to an underlying continuous variable $t_{ij}$, and we generate nominal $z_{ij}$ via thresholding of the underlying $t_{ij}$. This leads to complex posterior updating procedure with sticky mixing. The covariance matrix $\Sigma_{p \times p}$ for $\{t_{ij}\}$ needs extra constraints for identification, while these identification constraints make sampling $\Sigma$ from the posterior distributions difficult. When $p$ is large the computation cost can be expensive.

Log-linear model is a natural choice for imputation of categorical data (Schafer, 1997). However, log-linear models have known limitations in high dimensions (Erosheva et al., 2002). Model selection becomes very challenging, as the number of possible models is enormous. With high dimensions it is impossible to enumerate all possible log-linear models, so that automated model selection procedures—which are complicated to implement with missing data—are necessary. With sparse tables, as is the case in practice with high dimensional categorical data, many cells of the observed contingency table randomly equal zero. Maximum likelihood estimates of the log-linear model coefficients corresponding to zero margins cannot be determined, so that one either has to assume that those cells have expected values equal to zero (Bishop et al., 1975), which results in biased estimates of observed non-zero cell probabilities, or has to ensure that models do not include problematic cells, which
artificially restricts the range of possible models. The latter can be problematic for missing data imputation, in that subsequent estimates of complex interactions could be attenuated due to insufficiently complex imputation models.

Besides these specified joint imputation models, another popular variant under MI is sequential regression modeling, also called multiple imputation by chained equations (MICE) (Van Buuren and Oudshoorn, 1999; Raghunathan et al., 2001; Su et al., 2010). Rather than specifying a joint distribution of all variables for imputation, MICE constructs series of univariate conditional models to impute the variables with missing values one by one, after rearranging the missingness pattern to be monotone. For categorical variables, the models will typically be a set of logistic/probit or multinomial logistic/probit regression models. However, when the number of variables is large, we need to specify a large number of sequential imputation models, which is quite time-consuming. Typically, the default sequential imputation models only contain main effects or linearity without any higher-order association terms. This will fail to capture the complexity dependence relationship and lead to biased inference.

To improve the sequential imputation model in MICE, Burgette and Reiter (2010) proposed using classification and regression trees (CART) as the conditional models for imputation. CART performs flexibly to catch complex data structure without parametric assumptions. The imputer needs to control the size of trees, and CART will fit the relationship models automatically. CART partitions the predictor space by binary splits so that the obtained subsets of units have relatively homogeneous outcomes. The series of splits can be effectively represented by a tree structure, and result in models with many interaction effects. When one variable has many possible levels, CART may keep splitting this variable and ignore other significant variables with fewer levels.

However, for either MICE or the sequential CART, these conditional regression
models lack an strong statistical support: a joint distribution may not exist that corresponds to the series of specified condition distributions. For example, the order in which variables are placed in the chain could impact the imputations (Li et al., 2012a,b).

2.3 Mixture Models for Multiple Imputation

When confronted with high dimensional categorical data with nontrivial item non-response, I desire a multiple imputation approach that (i) avoids the difficulties of model selection and estimation inherent in log-linear models, (ii) has theoretical grounding as a a coherent Bayesian joint model, and (iii) offers efficient computation. I propose MI using a nonparametric Bayes approach for multivariate unordered categorical data with missing values: Dirichlet Process Mixture of Products of Multinomial distributions (DPMPM). This model was initially proposed by Dunson and Xing (2009) to capture positional dependence within transcription factor binding motifs. This nonparametric Bayes approach defines a prior with full support on the whole space of distributions for multivariate unordered categorical variables. Under a proper base measure, the resulting posterior distribution can closely approximate any joint distribution of these variables. DP (Ferguson, 1973, 1974) mixture models have been used widely for density estimation and dimension reduction. They provide a flexible framework and favor a sparse structure to allow efficient computation.

The number of mixture components $K$ is unknown a priori and will be inferred through the posterior computation. This is one of the advantages in the proposed approach over MI using latent class analysis (Vermunt et al., 2008), where the imputation model is essentially a finite mixture of independent product of multinomial distributions. In latent class analysis with MI, the number of mixture components $K$ is fixed as a selected positive integer. Vermunt et al. (2008) use BIC, AIC and a variant AIC referred as AIC3 to choose the value of $K$ for model selection. This
latent class approach is also called latent structure analysis (Lazarsfeld and Henry, 1968). Formann (2007) implemented latent class analysis to categorical data when the missingness may not be ignorable and the mechanism for missingness is modeled jointly with the variables of interest. Latent class analysis is a special case of the general diagnostic model (Von Davier, 2010), implemented in the software mdtm.

I describe the nonparametric Bayes approach for multiple imputation in detail. First I introduce the latent class analysis before generalizing to the infinite mixture models.

2.3.1 Finite Mixture of Products of Multinomial Distributions

Denote $X_{ij}$ as the corresponding category of variable $j$ for individual (student) $i$, for $i = 1, \ldots, N$ and $j = 1, \ldots, p$, where $N$ is the total number of individuals and $p$ is the total number of background variables. The possible values of $X_{ij}$ for variable $j$ fall in $\{1, \ldots, d_j\}$, where $d_j$ is the total number of categories for variable $j$ ($d_j \geq 2$). Let $c_j$ be the chosen value of variable $j$, where $c_j \in \{1, \ldots, d_j\}$. We obtain a $d_1 \times d_2 \times \cdots \times d_p$ contingency table $D$ with cell $(c_1, \ldots, c_p)$ containing the count $\sum_{i=1}^{N} I(X_{i1} = c_1, \ldots, X_{ip} = c_p)$ and cell probability $\theta_{c_1, \ldots, c_p}$, where $\theta_{c_1, \ldots, c_p} = \Pr(X_{i1} = c_1, \ldots, X_{ip} = c_p)$, for $c_j = 1, \ldots, d_j$ and $j = 1, \ldots, p$. I build models on the cell probabilities $\theta = \{\theta_{c_1, \ldots, c_p}\}$.

Suppose we have a finite number of $K$ latent classes, and all the individuals will be allocated into the $K < \infty$ clusters. The individuals inside one particular class will share common parameters and characteristics. Thus the values of the probability vector for the multinomial distributions inside any particular class are the same.

For $i = 1, \ldots, N$, let $s_i \in \{1, \ldots, K\}$ indicate the class for individual $i$, and let $\pi_h = \Pr(s_i = h)$. I assume that $\pi = (\pi_1, \ldots, \pi_K)$ is the same for all individuals. Within any class, I suppose that each of the $p$ variables independently follows a class-specific multinomial distribution. Inside one class, the joint probability for all
variables will be the product of the marginal probabilities for each variable. Let
\( Pr(X_{ij} = c_j | s_i = h) = \psi_{hcj}^{(j)} \) is the probability of \( X_{ij} = c_j \) given the allocation of individual \( i \) to class \( h \).

The finite mixture of products of multinomial distributions can be specified as

\[
X_{ij} | s_i, \psi \sim \text{Discrete}(\psi_{s_i1}, \ldots, \psi_{s_ip}) \quad \text{for all } i, j
\]

\[
s_i | \pi \sim \text{Discrete}(\pi_1, \ldots, \pi_K) \quad \text{for all } i.
\]  

(2.1)

For complete data, Dunson and Xing (2009) show that any joint distribution of \( X_{i1}, \ldots, X_{ip} \) can be characterized in (2.1) for some \( K \). By choosing sufficiently large value of \( K \), the finite mixture model should accurately pick up the first, second and higher-order observed moments of the \( p \) categorical variables (McLachlan and Peel, 2000).

Rich literature work has been focused on several issues of finite mixture models. First, lack of identification of the allocation probabilities \( \pi \) and the component specific parameters \( \Psi = \{ \psi_{hcj}^{(j)}, h = 1, \ldots, K, j = 1, \ldots, p \} \) can cause label switching problems. Different values of \( \pi \) and \( \Psi \) yield the same joint distribution \( \theta \), which is the interest. Even if \( \pi \) and \( \Psi \) are not identified, \( \theta \) is uniquely identified. Also, under-identifiability makes the interpretation more problematic, while latent classes may not be interpretable. Interpretation is not a concern when the goal is imputation. Meanwhile, over-fitting is less of a problem than under-fitting. In practice even relatively simple dependence structures may require allocation of individuals to many different classes leading to a large effective number of parameters. However, this is less problematic than ignoring important interactions among variables in the imputation model.

Model selection is another important concern, especially how to choose a sufficiently large value of \( K \). Vermunt et al. (2008) use criteria including BIC, AIC and AIC3, which combine model fitting and parsimony into one single value. In
their analysis, these three different criteria fail to select the same model consistently, where the chosen numbers of latent classes are quite different from others. It is hard to compromise to select a single $K$. Because data are often very sparse with most of the cells in the huge contingency table being empty, a unique Maximum likelihood estimate (MLE) of the parameters often does not exist even when a modest $K$ is chosen. Hence, inferences on the dependence structure in the observations may be severely biased. Meanwhile, the determination of $K$ ignores the uncertainty induced by $K$. It is not realistic to assume the same value of $K$ regardless the sample size.

The latent class methods can capture the associations among the observed data. However, choosing a sufficiently large enough number of latent classes is crucial. Using a large fixed number will complicate posterior inference and add to the computational cost, especially when the number of categories and/or variables is large. Nonparametric Bayes can solve the problem.

### 2.3.2 Infinite Mixture of Products of Multinomial Distributions

Selection of a single $K$ ignores the uncertainty about $K$. This could result in underestimation of variance in parameter estimates, which goes against Rubin’s (1987) recommendations for generating multiple imputations. In practical terms, underestimation of uncertainty could lead to unjustifiably precise multiple imputation inferences and reduced confidence interval coverage rates. It would be nice to have a flexible and efficient imputation method for a large collection of categorical variables with complicated relationship while favoring a sparse representation. This motivates a nonparametric Bayes extension, which avoids the selection of $K$ and tends to occupy only the first few components. The number of selected classes will be determined by the data and can grow with the sample size.

Dunson and Xing (2009) developed a nonparametric Bayes approach on multivariate unordered categorical data through a Dirichlet process mixture of products of
multinomial distributions. Use a Bayesian nonparametric prior $P$ on $\theta = \{\theta_{c_1, \ldots, c_p}\}$, where $Pr(X_{i1} = c_1, \ldots, X_{ip} = c_p) = \theta_{c_1, \ldots, c_p}$, specified as

$$\theta_{c_1, \ldots, c_p} = \sum_{h=1}^{\infty} \pi_h \prod_{j=1}^{p} \psi^{(j)}_{hc_j} \text{ for all } (c_1, \ldots, c_p) \in D \quad (2.2)$$

$$\psi^{(j)}_h \sim P_{0j} \text{ for } j = 1, \ldots, p, h = 1, \ldots, \infty \quad (2.3)$$

$$\pi \sim Q. \quad (2.4)$$

Here $P_{0j}$ is a probability measure on the $d_j$-dimensional probability simplex. I use the Dirichlet distributions for all $P_{0j}$ to describe properties of the probabilities $\psi^{(j)}_h$. Let $Q$ be a probability measure on the countably infinite probability simplex. I treat $Q$ as a DP. The well-known clustering property of the DP provides a nonparametric prior for the number of mixture components. The prior on $\theta$ is $P$, and $P \sim DP(\{P_{0j}, j = 1, \ldots, p\})$. I select the stick-breaking representation to facilitate straightforward posterior updating. The stick-breaking representation (Sethuraman, 1994) of DP implies that

$$\pi_h = V_h \prod_{l<h} (1 - V_l)$$

$$V_h \stackrel{iid}{\sim} \text{Beta}(1, \alpha) \text{ for } h = 1, \ldots, \infty, \quad (2.5)$$

where $\alpha$ is a precision parameter characterizing $Q$. If $\alpha$ is small, then $\pi_h$ will decrease stochastically rapidly to zero as $h$ increases. We can see that the prior favors a sparse representation in that most of the weight will be on the first few components. Moreover, setting a hyper-prior on $\alpha$ (Escobar and West, 1998) will allow the data more control of the sparsity with a Bayesian penalty for model complexity. I use the conjugate hyper-prior $\alpha \sim \text{Gamma}(a_\alpha, b_\alpha)$. 

24
To clarify, with the expression of latent class index $s$ the model is summarized as

$$X_{ij} \mid s_i, \psi \sim \text{Multinomial}(\psi_{s_{i1}}, \ldots, \psi_{s_{id_j}}) \quad \text{for all } i,j \quad (2.6)$$

$$s_i \mid \pi \sim \text{Multinomial}(\pi_1, \ldots, \pi_\infty) \quad \text{for all } i \quad (2.7)$$

$$\pi_h = V_h \prod_{l<h} (1 - V_l) \quad \text{for } h = 1, \ldots, \infty \quad (2.8)$$

$$V_h \sim \text{Beta}(1, \alpha) \quad (2.9)$$

$$\alpha \sim \text{Gamma}(a_\alpha, b_\alpha) \quad (2.10)$$

$$\psi_{hj} = (\psi_{h1}^{(j)}, \ldots, \psi_{hd_j}^{(j)}) \sim \text{Dirichlet}(a_{j1}, \ldots, a_{jd_j}) \quad (2.11)$$

All the $X_{ij}$ are conditionally independent given the latent class $z_i$ and the stick-breaking random variables $V = \{V_h\}$. Meanwhile, the probability vectors $\Psi = \{\psi_{h}^{(j)}\}$ are exchangeable across different classes.

For complete data, Dunson and Xing (2009) also showed that the prior $P$ has full support on the whole space $\prod_{d_1 \ldots d_p}$, with $\prod_{d_1 \ldots d_p}$ containing all $d_1 \times \cdots \times d_p$ probability matrices with elements $\theta_{c_1 \ldots c_p}$ and $\sum_{c_1=1}^{d_1} \cdots \sum_{c_p=1}^{d_p} \theta_{c_1 \ldots c_p} = 1$. The resulting posterior distribution is consistent.

### 2.4 Posterior Computation and Inference

If we set all $P_{0j}$ as finite Dirichlet distributions and $Q$ as a DP, the posterior distributions can be modified to be conjugate and then can be updated through a straightforward Gibbs sampler. I implement two posterior computation algorithms.

#### 2.4.1 Blocked Gibbs Sampler

The first one is blocked Gibbs sampler (Ishwaran and James, 2001), which truncates the infinite stick-breaking probabilities at a large number $H^*$. The truncated stick-breaking process is expressed as $\theta = \sum_{h=1}^{H^*} \pi_h \Psi_h$ or $s_i \sim \sum_{h=1}^{H^*} V_h \prod_{l<h} (1 - V_l) \delta_h$. Here, one makes $H^*$ as large as possible while still offering fast computation. Using
an initial proposal for $H^*$, say $H^* = 100$, analysts can examine the posterior distributions of the sampled number of unique classes across MCMC iterates to diagnose if $H^*$ is large enough. Significant posterior mass at a number of classes equal to $H^*$ suggests that the truncation limit be increased. A Gibbs sampler can be derived to estimate the posterior distributions of the parameters and the missing values. The steps for posterior computations are specified as following:

**Step 1:** Update $s_i \in \{1, \ldots, H^*\}$ from multinomial distribution with probabilities

$$Pr(s_i = h|\cdot) = \frac{\pi_h \prod_{j=1}^p \psi_h^{(j)} X_{ij}}{\sum_{k=1}^{H^*} \pi_k \prod_{j=1}^p \psi_k^{(j)} X_{ij}},$$

where $X_{ij} \in X_{com}$ and $X_{com} = \{X_{obs}, X_{mis}\}$, for $i = 1, \ldots, N$ and $j = 1, \ldots, p$.

**Step 2:** Update $V_h$ from conjugate Beta distributions, for $h = 1, \ldots, H^* - 1$

$$(V_h|\cdot) \sim \text{Beta}(1 + n_h, \alpha + \sum_{k=h+1}^{H^*} n_k),$$

where $n_h = \sum_{i=1}^N I(s_i = h)$. Here, $I(\cdot) = 1$ when the condition inside the parentheses is true and $I(\cdot) = 0$ otherwise. Set $V_{H^*} = 1$ under truncation, and calculate $\pi_h$ from $\pi_h = V_h \prod_{k<h} (1 - V_k)$.

**Step 3:** Update $\psi_h^{(j)} = (\psi_h^{(j)}_1, \ldots, \psi_h^{(j)}_{d_j})$ from conjugate Dirichlet distributions

$$(\psi_h^{(j)}|\cdot) \sim \text{Dirichlet} \left( a_{j1} + \sum_{i:s_i=h} I(X_{ij} = 1), \ldots, a_{jd_j} + \sum_{i:s_i=h} I(X_{ij} = d_j) \right).$$

**Step 4:** Update $\alpha$ from conjugate Gamma distributions

$$(\alpha|\cdot) \sim \text{Gamma}(a_\alpha + H^* - 1, b_\alpha - \log \pi_{H^*}).$$
Step 5: Update $X_{mis}$ from multinomial distributions. Include a response indicator $r_{ij}$, where $r_{ij} = 1$ if $X_{ij}$ is observed and $r_{ij} = 0$ if $X_{ij}$ is missing. We need to update $X_{ij}$ when $r_{ij} = 0$. Draw new values for those $X_{ij}$’s with $(i, j) \in \{(i, j) : r_{ij} = 0\}$, 

\[
(X_{ij}|-) \sim \text{Multinomial}(\{1, \ldots, d_j\}; \psi_{s_i,1}^{(j)}, \ldots, \psi_{s_i,d_j}^{(j)}).
\]

2.4.2 Slice Sampling Algorithm

The second approach is called slice sampling (Walker, 2007) by introducing a vector of latent variables $u = \{u_1, \ldots, u_N\}'$. Define the joint and complete likelihood of $u$ and $X$ given $X_{mis}$, $V$ and $\Psi$ as

\[
\prod_{i=1}^{N} \left\{ \sum_{h=1}^{\infty} I(u_i < \pi_h) \prod_{j=1}^{p} \prod_{l=1}^{d_j} (\psi_{h,l}^{(j)})^{I(X_{ij} = l)} \right\}.
\]  

(2.12)

Including $s$, the joint likelihood can also be expressed as

\[
\prod_{i=1}^{N} \left\{ I(u_i < \pi_{s_i}) \prod_{j=1}^{p} \prod_{l=1}^{d_j} (\psi_{s_i,l}^{(j)})^{I(X_{ij} = l)} \right\}.
\]  

(2.13)

I update parameters and impute missing values based on data augmentation with latent classes and the uniformly distributed variable. Using the approach of Papaspiliopoulos (2008), the efficiency of the slice sampling algorithm proposed by Dunson and Xing (2009) can be improved through approximating the posterior distribution of $V_h$, which is a truncated Beta distribution, by the Beta distribution without truncation.

The steps for posterior computations are as follows:

Step 1: Update $u_i$, for $i = 1, \ldots, N$, from $\text{Uniform}(0, \pi_{s_i})$.

Step 2: Update $\psi_{h}^{(j)}$, for $h = 1, \ldots, k^*$, where $k^* = \max(s_1, \ldots, s_N)$,

\[
(\psi_{h}^{(j)}|-) \sim \text{Dirichlet}(a_{j1} + \sum_{i:s_i=h} I(X_{ij} = 1), \ldots, a_{jd_j} + \sum_{i:s_i=h} I(X_{ij} = d_j)).
\]
If there are no $s_i$’s equal to $h$, we generate $\psi_h^{(j)}$ from the prior.

**Step 3:** Update $V_h$, for $h = 1, \ldots, k^*$, where $k^* = \max(s_1, \ldots, s_N)$. Rather than a truncated Beta distribution

$$\text{Beta}(1, \alpha_0)I \left( \left[ \max_{i:s_i = h} \left\{ \frac{u_i}{\prod_{l < h} (1 - V_l)} \right\}, 1 - \max_{i:s_i > h} \left\{ \frac{u_i}{\prod_{l < h, l \neq h} (1 - V_l)} \right\} \right] \right),$$

we generate $V_h$ from

$$(V_h|-) \sim \text{Beta}(1 + n_h, \alpha + \sum_{k=h+1}^{k^*} n_k),$$

where $n_h = \sum_{i=1}^{N} I(s_i = h)$, and then directly calculate $\pi_h$.

**Step 4:** Update $s_i$, for $i = 1, \ldots, N$, from the multinomial distribution

$$Pr(s_i = h|-) = \frac{I(h \in A_i) \prod_{j=1}^{p} \psi_{pX_{ij}}^{(j)}}{\sum_{l \in A_i} \prod_{j=1}^{p} \psi_{pX_{ij}}^{(j)}},$$

where $A_i = \{h : \pi_h > u_i\}$. To identify the elements in $A_1, \ldots, A_N$, first update $V_h$, for $h = 1, \ldots, \tilde{k}$, where $\tilde{k}$ is the smallest value satisfying $\sum_{h=1}^{\tilde{k}} \pi_h > 1 - \min\{u_1, \ldots, u_N\}$, i.e. sample as many of $\pi_h$ until we are sure that we have all the $\pi_h > u_i$.

**Step 5:** Update $\alpha$ from conjugate Gamma distributions

$$(\alpha|-) \sim \text{Gamma}(a_\alpha + k^*, b_\alpha - \sum_{h=1}^{k^*} \log(1 - \pi_h)).$$

**Step 6:** Update $X_{mis}$ from multinomial distributions. Draw new values for those $X_{ij}$’s with $(i, j) \in \{(i, j) : r_{ij} = 0\}$ from

$$(X_{ij}\mid-) \sim \text{Multinomial}\{\{1, \ldots, d_j\}, \psi_{s_1, \ldots, \psi_{s,d_j}}^{(j)}\}.$$
The Gibbs sampler proceeds by initializing the chain. I suggest initializing \( \alpha = 1 \), each \( V_h \) with independent draws from Beta(1, 1) and \( \psi \) with the marginal estimates based on the data. Then, I proceed through each step, repeating many times until convergence. To initialize \( X_{mis} \) for the MCMC, I suggest sampling from marginal distributions of each \( X_j \).

To obtain \( m \) completed datasets for use in multiple imputation, analysts select \( m \) of the sampled \( X_{mis} \) after convergence of the Gibbs sampler. These datasets should be spaced sufficiently so as to be approximately independent (given \( X_{obs} \)). This involves thinning the MCMC samples so that the autocorrelations are close to zero.

2.5 Simulation Studies of Frequentist Performance

I now investigate the performance of the DPMPM multiple imputation method via simulation studies, focusing on repeated sampling properties. I consider two scenarios: a small number of variables (\( p = 7 \)) generated via log-linear models and a somewhat large number of variables (\( p = 50 \)) generated via finite mixtures of multinomial distributions. In both I consider simulated bias of point estimates and coverage rates of 95% confidence intervals, all constructed using Rubin’s (1987) methods of multiple imputation inference.

2.5.1 Case 1: Small \( p \) and Log-linear Model Data Generation

With modest \( p \), it is straightforward to write down and simulate from distributions that encode complex dependence relationships. I generate data comprising \( n = 5000 \) individuals and \( p = 7 \) binary variables as follows. The first five variables are sampled independently from a multinomial distribution with probabilities governed by the
log-linear model with

\[
\log \Pr(X_1, X_2, X_3, X_4, X_5) \propto \sum_{j=1}^{5} -2X_j + \sum_{j=1}^{4} \sum_{j'=j+1}^{5} X_j X_{j'} + X_1 X_2 X_3
- X_2 X_3 X_4 - 2X_3 X_4 X_5 + X_2 X_3 X_5 + X_1 X_4 X_5. \tag{2.14}
\]

I generate \(X_6\) from Bernoulli distributions with probabilities governed by the logistic regression with

\[
\logit \Pr(X_6) = -1 + X_1 + 2.2X_2 - 2.5X_3 + .9X_4 + 1.1X_5 - 2.8X_2 X_3 + 2.3X_3 X_4
- .5X_2 X_4 - 2.4X_3 X_5 + 1.55X_1 X_4 - 2.1X_4 X_5 + 1.2X_3 X_4 X_5. \tag{2.15}
\]

I generate \(X_7\) from Bernoulli distributions with probabilities governed by the logistic regression with

\[
\logit \Pr(X_7) = -.3 + 1.5X_1 - 2.15X_2 - 2.25X_3 + 1.6X_4 - .88X_5 + 1.11X_6 - .96X_2 X_3
+ 2.3X_1 X_3 - .5X_2 X_6 - 2X_5 X_6 + 1.21X_1 X_5 - 2.7X_1 X_2 + 1.5X_1 X_2 X_3. \tag{2.16}
\]

The specific values of the coefficients are not particularly important, except for noting that the higher order interactions with non-zero coefficients generate complex dependencies.

I suppose that \((X_1, X_2, X_7)\) have values missing at random. Let \(X_1\) be missing with probabilities \((0.1, 0.4, 0.4, 0.7)\), respectively, for each of the four combinations of \((X_3, X_4)\). Let \(X_2\) be missing with probabilities \((0.7, 0.4, 0.4, 0.1)\), respectively, for each of the four combinations of \((X_5, X_6)\). Let \(X_7\) be missing with probabilities \((0.5, 0.2, 0.3, 0.7)\), respectively, for each of the four combinations of \((X_5, X_6)\). About 99% of the units have at least one missing value.

After introducing missing data, I implement the DPMPM to create \(m = 5\) multiply-imputed datasets. I set \(H^* = 20\) and run the chains for 50000 iterations, which via experimentation appears sufficient to ensure convergence and offer repeated
simulation results in reasonable time. I also implement a default version of chained equations, using the MICE software package in R (Van Buuren and Oudshoorn, 1999). I repeat the process of generating observed data, introducing missing values and performing multiple imputations for 500 times.

I evaluate the two approaches on regression coefficients for three models. The first model is the log-linear model in (2.14). The second and third models are the two logistic regressions in (2.15) and (2.16), excluding the three-way interactions. These are excluded so as to avoid problems caused by random zeros in the repeated simulations. Random zeros cause logistic regression coefficient estimates and standard errors to blow up, whether in the evaluation or in MICE. I find that random zeros do not cause problems for the DPMPM imputation procedure, which is another advantage.

Figure 2.1 displays average point estimates and 95% confidence interval coverage rates across the 500 simulations. The average point estimates based on DPMPM are closer to the corresponding true values than those based on default MICE. Across all estimands and simulations, the average mean squared error of $\hat{q}_m$ equals .08 when using DPMPM, whereas it equals .13 (50% higher) when using default MICE. The simulated coverage rates of the 95% confidence intervals based on DPMPM generally are closer to 95% than those based on default MICE. Indeed, default MICE results in several rates below 20%. These belong to the three-way interaction terms from (2.14). The DPMPM, in contrast, has reasonable coverage rates for the three-way interactions. The simulated coverage rates below 80% for the DPMPM belong to coefficients in the logistic regression for $X_7$. I find that these rates (77.4%, 65.2%, 70.8%, 49.6%) for the most part are better than the corresponding ones from default MICE (46.8%, 30.4%, 96.4%, 16.6%). The outputs show that the improved coverage rates for DPMPM do not result from unrealistic inflation of variances, as evidenced by the reasonable standard error bars for $\hat{q}_m$ in the left panel of Figure 2.1.
Figure 2.1: Small $p$ simulation results. Simulated average point estimates and 95% confidence interval coverage rates for 45 regression coefficients. True values of coefficients obtained by fitting each model on a very large, completed dataset generated from the data models. Standard error bars for each $\bar{q}_m$ for DPMPM equal $1.96\sqrt{\text{avg. } T_m}$, where “avg. $T_m$” is the average multiple imputation variance estimate across the 500 replications.

Clearly, the DPMPM represents a substantial improvement over default MICE for these simulations. Of course, one can do better than main effects only conditional models when using MICE. Including interaction effects should improve coverage rates. However, we suspect that the complex dependencies in these data would be challenging to identify in practice when specifying the conditional regression models, and that many analysts would use the default application of MICE.
2.5.2 Case 2: Large $p$ and Mixture Model Data Generation

In general, it is computationally cumbersome to simulate large contingency tables with complex dependence structure from log-linear models and logistic regressions. I therefore generate tables from a finite mixture model. I set $p = 50$ and allow the number of levels for each variable to be randomly chosen from 2 to 6. The final table has $d \approx 10^{30}$ cells. I sample $n = 1000$ individuals so that most of the $d$ cells in any sampled dataset are in fact empty. I use $K = 4$ classes such that $(\pi_1 = .3, \pi_2 = .2, \pi_3 = .4, \pi_4 = .1)$. Within any class $h$, I set $\psi$ to differ across $(h, j, x)$ so as to induce complex dependence. Specifically, for each $(h, j, x)$ I set

$$
\psi_{hx}^{(j)} = \max \left( \frac{h(d_j - 1)}{(h + 1)d_j^2}, 0.05h \right), \quad \psi_{hdj}^{(j)} = 1 - \sum_{x=1}^{d_j-1} \psi_{hx}^{(j)}. \tag{2.17}
$$

Although it is difficult to summarize succinctly the degree of dependence that results, I find that in one completed dataset randomly generated from this model only 103 of the $\binom{50}{2} = 1225$ bivariate $\chi^2$ tests of independence had $p$-values exceeding .05.

After generating completed data, I make each of the first 20 variables have 40% values missing completely at random. This results in no complete cases. I implement multiple imputation using the DPMPM with $m = 5$, $H^* = 20$, and 100000 MCMC iterations. I did not implement MICE, as it was computationally too expensive to run in repeated simulations with $p = 50$. With the large numbers of random zeros that result from this simulation, I would be essentially forced to run MICE with main effects only to avoid (randomly) inestimable multinomial regression coefficients. I repeat the process of generating observed data, introducing missing values and performing DPMPM multiple imputation for 100 times. The smaller number of simulations than in Section 4.1 reflects the increased time for evaluation with $p = 50$.

For evaluation purposes, I examine twenty arbitrary conditional probabilities, including (i) four involving 2 variables both with missing data, (ii) four involving 2
variables with only one having missing data, (iii) four involving 3 variables all with missing data, (iv) four involving 3 variables with two having missing data, and (v) four involving 3 variables with one having missing data.

Figure 2.2 displays average point estimates and 95% confidence interval coverage rates across the 100 simulations. The average point estimates based on DPMPM are close to the corresponding true values, and the simulated coverage rates are at least 93%, suggesting reasonable performance. The conservative nature of the simulated coverage rates could be an artifact of the limited number of simulation runs.

2.6 Imputation of TIMSS Background Variables

The TIMSS is conducted by the International Association for the Evaluation of Educational Achievement. Data are collected on a four year cycle and made available for downloads via a dedicated TIMSS website (www.timssandpirls.bc.edu). The goal of TIMSS is to facilitate comparisons of student achievement in mathematics and science across countries. In addition to domain-specific test questions, TIMSS data include background information on students including demographics, amount of educational resources at the home, time spent on homework and attitudes towards mathematics and science. These background variables are all categorical.

I use data from the 2007 TIMSS that comprise 80 background variables on 90505 students (88129 in Grade 4 and 2376 in Grade 5) from 22 countries. Among these 80 variables, most (68) have less than 10% missing values; six variables have between 10% and 30% missing values; only one variable has more than 75%. Missingness rates differ by country but not dramatically so. The TIMSS data file lists reasons for missingness, including omitted (student should have answered but did not), not administered (missing because of the rotation design or unintentional misprint) and not reached (incorrect responses). For purposes of multiple imputation, I do not distinguish response reasons and treat all item nonresponse as missing at random.
Figure 2.2: Large $p$ simulation results. Simulated average point estimates and 95% confidence interval coverage rates for 20 conditional probabilities. True values of probabilities obtained on all completed datasets generated from the data models. Standard error bars for each $\bar{\theta}_m$ for DPMPM equal $1.96\sqrt{\text{avg. } T_m}$, where “avg. $T_m$” is the average multiple imputation variance estimate across the 100 replications.

To create multiple imputations, I run DPMPMs separately in each country. Separate imputation avoids smoothing estimates towards common values, which seems prudent since TIMSS is intended for comparisons across countries. It is possible to extend the DPMPM to allow borrowing information across countries using the hierarchical Dirichlet process described in Appendix A. When variables are not recorded in a country, I remove them from imputation for that country.

For the MCMC, I set $H^* = 20$. The posterior distribution of the number of classes
among individuals (within any country) had nearly all of its mass below twenty, so that we do not expect truncation to impact the imputations materially. Imputations with $H^* = 50$ on a smaller set of countries resulted in similar performance. I also examined different vague prior specifications for $a_\alpha$ and $b_\alpha$—including, for example, $(a_\alpha = 1, b_\alpha = .25)$, $(a_\alpha = 1, b_\alpha = 1)$, and $(a_\alpha = 1, b_\alpha = 2)$—and did not observe noticeable differences in the posterior distributions of $\theta$. In each country, I run the Gibbs sampler for 100000 iterations. MCMC diagnostics of marginal and randomly selected joint components of $\theta$ suggest convergence. Generating 100000 iterations for all 22 countries in series took approximately 3 days using a standard desk top computer. Of course, this time could be significantly reduced by running the countries’ Gibbs samplers in parallel.

To assess the quality of the multiple imputations, I focus on one arbitrarily selected country in the data file. In this country, fifty-nine variables have 10% or less missing values, eleven variables have between 20% and 30% missing values, five variables have between 50% and 70% missing values, and one variable has more than 80% missing values. Only 32 out of 4223 individuals have complete records. For purposes of evaluating the imputations, I increased the MCMC iterations to 500000. This took roughly 10 hours to run.

Comparisons of the marginal distributions of the observed and imputed values (Gelman et al., 2005) show similar distributions; these are not shown here to save space. While comforting, such diagnostics offer only partial insights into the quality of the imputations for multivariate relationships. I therefore consider posterior predictive checks that directly assess the ability of the imputation models to preserve associations, following the approach in He et al. (2010) and Burgette and Reiter (2010). The basic idea is to use the imputation model to generate not only $X_{mis}$ but an entirely new full dataset, i.e., create a completed dataset $D^{(t)} = (X_{obs}, X_{mis}^{(t)})$. 
and a replicated dataset $R^{(l)}$ in which both $X_{\text{obs}}$ and $X_{\text{mis}}$ are simulated from the imputation model. After repeating the process of generating pairs $(D^{(l)}, R^{(l)})$ many times (I use $T = 500$), I compare each $R^{(l)}$ with its corresponding $D^{(l)}$ on statistics of interest. When the statistics are dissimilar, the diagnostic suggests that the imputation model does not generate replicated data that look like the completed data, so that it may not be generating plausible values for the missing data. When the statistics are not dissimilar, the diagnostic does not offer evidence of imputation model inadequacy (with respect to that statistic).

More formally, let $S$ be the statistic of interest, such as a regression coefficient or joint probability. Let $S_{D^{(i)}}$ and $S_{R^{(i)}}$ be the values of $S$ computed with $D^{(l)}$ and $R^{(l)}$, respectively. For each $S$ we compute the two-sided posterior predictive probability,

$$
\text{ppp} = (2/T) \cdot \min \left( \sum_{i=1}^{T} I(S_{D^{(i)}} - S_{R^{(i)}} > 0), \sum_{i=1}^{T} I(S_{R^{(i)}} - S_{D^{(i)}} > 0) \right).
$$

(2.18)

We can see that $\text{ppp}$ is small when $S_{D^{(i)}}$ and $S_{R^{(i)}}$ consistently deviate from each other in one direction, which would indicate that the imputation model is systematically distorting the relationship captured by $S$. For $S$ with small $\text{ppp}$, it is prudent to examine the distribution of $S_{R^{(i)}} - S_{D^{(i)}}$ to evaluate if the difference is practically important.

To obtain the pairs $(D^{(l)}, R^{(l)})$, I add a step to the MCMC that replaces all values of $X_{\text{mis}}$ and $X_{\text{obs}}$ using the parameter values at that iteration. This step is used only for computation of the $\text{ppp}$; the estimation of parameters continues to be based on $X_{\text{obs}}$. When autocorrelations among parameters are high, I recommend thinning the chain so that $\theta$ draws are approximately independent before creating the set of $R^{(l)}$. Further, I advise saving the $T$ pairs of $(D^{(l)}, R^{(l)})$, so that they can be used repeatedly with different $S$.

I present posterior predictive checks for 36 coefficients in a multinomial logistic
regression and 1000 joint probabilities from a contingency table. The multinomial logistic regression predicts how much students agree that they like being in school (like a lot, like a little, dislike a little, and dislike a lot). It has 4.5% missing values. The predictor variables include how much students agree that they have tried their best (4 categories); how much students agree that teachers want students to do their best (4 categories); whether or not students have had something stolen from them at school (2 categories); whether students were hit or hurt by others at school (2 categories); whether students were made to do things by others at school (2 categories); whether or not students were made fun of or called names at school (2 categories); and, whether or not students were left out of activities at school (2 categories). Among all these predictors, the missing data rates range from 4.5% to 5.8%. Investigations of the completed datasets indicate strong associations among the variables.

The contingency table includes whether or not students ever use a computer at home (2 categories); whether or not students ever use a computer at school (2 categories); whether or not students ever use a computer elsewhere (2 categories); how often students use a computer for mathematics homework (5 categories); how often students use a computer for science homework (5 categories); and, how often students spend time playing computer games (5 categories). Among all these variables, the missing data rates range from 10% to 63%.

I consider each coefficient and joint probability as separate $S$. The 1036 values of $ppp$ are displayed in Figure 2.3. For the contingency table, only 27 out of 1000 values are below .05, suggesting that overall the DPMMPM model generates replicated tables that look similar to the completed ones. For the multinomial regression, three out of 36 values of $ppp$ are below .05 but above .01, and two are below .01. These suggest potential model mis-specification involving the associated variables. These five low values correspond to coefficients for the two four-category variables.
Figure 2.3: Frequency distributions of *ppp* for the TIMSS imputation. The left panel is for the 1000 cell probabilities, i.e., the completed-data counts over *N*, and the right panel is for the 36 multinomial regression coefficients.

These variables each have two levels with moderately low marginal probabilities (between 5.2% and 6.8%). Apparently, the DPMPM is somewhat inaccurate at replicating the associations with the outcome at these levels. I note, however, that the standard errors for these coefficients are large compared to the point estimates, so that the impact of modest imputation model mis-specification on multiple imputation inferences for these coefficients is likely to be swamped by sampling variability.
2.7 Conclusions

The Dirichlet process mixture of products of multinomial distributions offers a fully Bayesian, joint model for multiple imputation of large-scale, incomplete categorical data. The approach is flexible enough to capture complex dependencies automatically and computationally efficient enough to be applied in large datasets. Although based on mixture models, the approach avoids \textit{ad hoc} selection of a fixed number of classes, thereby reducing risks of using too few classes while fully estimating uncertainty in posterior distributions.

This approach can serve as the basis for additional methodological and applied developments. Many surveys include both categorical and continuous data. The nonparametric Bayesian approach could be extended to handle mixed data, for example by letting the continuous data be modeled as mixtures of independent normal distributions within latent classes. Many education surveys have data with hierarchical structures, such as students within teachers or schools within counties. The nonparametric Bayesian approach could be extended to include random effects that account for such hierarchical structure. Finally, many educational surveys employ multiple imputation to create “plausible values” of students’ abilities (e.g., Mislevy et al., 1992). To the best of my knowledge, current practice typically is to impute plausible values and missing background characteristics separately. Doing so simultaneously, or at least imputing missing background variables conditional on plausible values, may offer improved accuracy when estimating associations between background variables and student proficiency.
3 Semi-parametric Selection Models for Potentially Non-ignorable Attrition in Panel Study with Refreshment Sample

3.1 Introduction

Panel studies are commonly used in surveys organized by government, academic disciplines, and business sectors. The key feature of panel or longitudinal surveys is that the same individuals are interviewed repeatedly at different points in time. Panel surveys offer rich data that facilitate statistical models of complex relationships for attitudes and behaviors in economic (Baltagi and Song, 2006), political (Hillygus, 2005), educational (Buckley and Schneider, 2006), and social science (Western, 2002) fields. The American National Election Study (Bartels, 1999), the General Social Survey, the Panel Survey on Income Dynamics (PSID), and the Current Population Survey, are major surveys in U.S. containing panel components. Millions of dollars are spent annually to collect high quality and large amounts of data. Typical longitudinal studies contain multiple waves across a long-term time period, where each wave collects a large number of item responses. For example, the 2007-2008 Associ-
ated Press/Yahoo News election panel (APYN), conducted over the internet (data available online http://www.knowledgenetworks.com/ganp/election2008), includes 11 waves in one year. Each wave has around 110 questions.

Missing data problems are common and especially severe in panel surveys. Panel attrition and item nonresponse are two main inevitable issues that reduce effective sample sizes and introduce bias or uncertainty. Panel attrition occurs when individuals who participate in the initial wave of a panel drop out in the follow-up waves because they cannot be located or refuse continuous participation. Take the multiple-decade PSID (Zabel, 1998; Fitzgerald et al., 1998; Lillard and Panis, 1998) for an example, which was first fielded in 1968. By 1989 nearly half of the initial sample members were lost due to cumulative attrition and mortality. Even with a much shorter study, the APYN study was conducted over the course of 2008 election cycle. During the final wave, only 61 percent of the baseline survey participants stayed in the panel. In addition to panel attrition, item nonresponse is also common because participants refuse to provide answers to some particular question or do not have much to say about the question. Both panel attrition and item nonresponse reduce effective sample sizes and decrease analysts’ abilities to discover longitudinal traits. Attrition may result in an available sample that is not representative of the target population, and therefore yield biased inferences if the tendency to drop out is systematically related to the substantive outcome of interest (Olsen, 2005; Behr et al., 2005; Bhattacharya, 2008).

Fundamentally, panel attrition is a problem of non-response. The missing values are a mix of non-ignorable and random missingness. When missing values are treated as a different category or imputed by an ad hoc plug-in value, relationships can be distorted and inference can be biased. Complete case analysis, perhaps with the use of post-stratification weights (Vandecasteele and Debels, 2007), ignores attrition. List wise deletion creates a balanced sub panel under the assumption that
the nonresponse is MCAR. Wide recognition among experts about the panel attri-
tion and item nonresponse makes this complete case analysis unrepresentative and
risky. Most analysts tend to assume that the values are MAR, where missingness
depends only on observed data. Re-weighting and single imputation methods (hot
deck, nearest neighbor, last-observation-carried-forward, etc ) have been used to ad-
just the nonresponse due to MAR. However, it has been shown that these approaches
can increase the variability and do not correct panel nonresponse adequately (Gel-
man, 2007; Vandecasteele and Debels, 2007; Hogan and Daniels, 2008). Common
MI approaches are appropriate for missing values due to MAR or MCAR. MI can
also handle NMAR. When the panel attrition is expected as NMAR, the missingness
mechanism should be modeled.

Selections models (Hausman and Wise, 1979; Kenward, 1998; Scharfstein et al.,
1999) and pattern mixture models (Little, 1993; Kenward et al., 2003; Lin et al.,
2004; Roy and Daniels, 2008) are the two main methods to model the attrition
in longitudinal data analysis. However, the panel data alone cannot estimate or
identify the model for missingness mechanism; untestable assumptions about the
attrition process are necessary. Strongly informative priors (Hogan and Daniels,
2008) and additional deterministic constraints (Little and Wang, 1996) are assumed
for identification. Schluchte (1982), Brown (1990), and Diggle and Kenward (1994)
proposed methods for non-random drop-out process. These methods make various
assumptions about the missing data mechanisms, and their utility in practical studies
depends critically on whether these assumptions apply in the specific application.

To reduce reliance on untestable assumptions, I consider the use of refreshment
samples. Refreshment samples include new, randomly sampled respondents who are
given the same questionnaire at the same time as a second or subsequent wave of the
panel. Refreshment samples are available for many large scale panel studies, such
as the National Educational Longitudinal Study, the Health and Retirement Survey,
and the APYN study analyzed in this paper. Rotating panels can be interpreted as equivalent as refreshment samples, for example, in the Survey of Income and Program Participation and the Current Population Survey. Hirano et al. (2001) show that refreshment samples provide information that can be used to assess the effects of panel attrition and to correct for biases via statistical modeling. However, they have not been fully utilized for statistical modeling. Analysts typically use the refreshment samples only for exploratory checks or comparisons.

To my knowledge, no literature has addressed the utilization of refreshment samples in large-scale longitudinal studies. One challenge lies in efficient methods for handling missing values in incomplete large-scale categorical datasets. I meet this challenge by introducing a joint Bayesian selection model that offers interpretation of the missingness mechanism and multiple imputations for valid inference. I call this approach a semi-parametric selection model: use a parametric model for missingness mechanism and nonparametric model for the completed data.

I describe the semi-parametric selection model in detail in Section 3.2 and the computation algorithms for posterior updating in Section 3.3. I use simulation studies to evaluate the procedure in Section 3.4. I apply the procedure to the APYN data set in Section 3.5. I summarize the chapter and discuss future research directions in Section 3.6.

3.2 Methods

I first introduce the selection model for non-ignorable panel attrition with refreshment samples. I then describe the imputation models for large-scale categorical data. All the variables considered here are either binary or nominal. The method can be extended to handle mixed data types of continuous and ordered categorical variables.

Consider a two wave panel of $N_p$ individuals with a refreshment sample of $N_r$ new subjects in the second wave. Suppose that in the baseline panel, $p_1$ response variables
are of interest: \( Y_1 = (Y_{11}, \ldots, Y_{1p_1}) \). The corresponding \( p_2 \) outcome variables in wave 2 are \( Y_2 = (Y_{21}, \ldots, Y_{2p_2}) \). Here I assume \( p_2 = p_1 \). In the panel, \( Y_1 \) have item nonresponse among the \( N_p \) individuals. For a subset of \( N_{cp} \) individuals (\( N_{cp} < N_p \)), \( Y_2 \) are collected and have item nonresponse; the remaining subset of \( N_{ip} = N_p - N_{cp} \) individuals drop out of the panel. The refreshment sample includes \( Y_2 \) as well and is subject to item nonresponse among the \( N_r \) individuals. By design, \( Y_1 \) are all missing for the individuals in the refreshment sample.

For each individual \( i = 1, \ldots, N = N_p + N_r \), let \( W_i = 1 \) if individual \( i \) would remain in the wave 2 if included in wave 1; let \( W_i = 0 \) if individual \( i \) would drop out of wave 2 if included in wave 1. We note that \( W_i \) is fully observed for all individuals in the panel but is missing for the individuals in the refreshment sample. The individuals in the refreshment sample are not provided the chance to respond in wave 1, so \( W \) is missing for them. I collect \( q \) demographic variables \( X = (X_1, \ldots, X_q) \) for all \( N \) individuals, which are also subject to item nonresponse. These are invariant across time. Let \( p = q + p_1 + p_2 \) be the total number of variables and \( Z = (Z_1, \ldots, Z_p) = (X_1, \ldots, X_q, Y_{11}, \ldots, Y_{1p_1}, Y_{21}, \ldots, Y_{2p_2}) \) denote all the variables.

In sum, I have mixed types of missingness due to attrition, item nonresponse and design as follows:

- \( W \) in the refreshment sample by design,
- \( Y_1 \) in the refreshment sample by design,
- \( Y_2 \) for the \( N_{ip} \) individuals in the panel due to attrition,
- \( Y_1 \) in the panel due to item nonresponse,
- \( Y_2 \) for the \( N_{cp} \) individuals in the panel due to item nonresponse,
- \( Y_2 \) in the refreshment sample due to item nonresponse,
Table 3.1: Graphical representation for panel and refreshment (ref) sample.

<table>
<thead>
<tr>
<th></th>
<th>Baseline</th>
<th>Wave 1</th>
<th>Wave 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Panel</td>
<td>X</td>
<td>$Y_1$</td>
<td>$Y_2$, $W=1$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$Y_2=?$, $W=0$</td>
</tr>
<tr>
<td>Ref</td>
<td>X</td>
<td>$Y_1=?$</td>
<td>$Y_2$, $W=?$</td>
</tr>
</tbody>
</table>

- $X$ in the panel due to item nonresponse,
- $X$ in the refreshment sample due to item nonresponse.

Table 3.1 displays the graphical representation of the data in the panel and refreshment sample. I impute all the above missing parts to obtain a completed and balanced panel plus refreshment sample. In this chapter, I assume the item nonresponse are MAR and include the potentially non-ignorable attrition into the modeling procedure.

3.2.1 Attrition Model

Historically, two main models are used for attrition in panel studies from a selection model perspective. The first model, which is also the most commonly used, allows the probability of attrition to depend only on observed values under the MAR assumption. The attrition tendency depends on the lagged $Y_1$ in wave 1 or the demographic information $X$. The second model, developed by Hausman and Wise (1979) (HW), allows the probability of attrition to depend on unobserved values. The HW model assumes the attrition depends only on the contemporaneous variables $Y_2$ rather than the lagged and observed information. Without refreshment samples, one can only include either $Y_1$ or $Y_2$ in the model for $W$, but cannot allow both together due to lack of identifiability. When the refreshment samples are available, MAR and HW models have testable implications and can be embedded in the more general and identified additive non-ignorable model (AN), proposed by Hirano et al. (2001). The AN model assumes the probability of attrition depends on both the observed and
unobserved variables through an additive format without any interaction between $Y_1$ and $Y_2$. Main effects, potential interaction terms among the variables in wave 1, and potential interaction terms among the variables in wave 2 can be included and identified in the AN model. Hirano et al. (2001) demonstrate the AN model for binary responses, but it can be generalized to any data type under any joint distributions.

I use the AN model that includes the main effects for the variables shown as below

$$
\text{logit } \Pr(W_i = 1|Y_1, Y_2, X) = \beta_0 + \beta_{X_1}X_1 + \cdots + \beta_{X_q}X_q
+ \beta_{Y_{11}}Y_{11} + \cdots + \beta_{Y_{1p_1}}Y_{1p_1} + \beta_{Y_{21}}Y_{21} + \cdots + \beta_{Y_{2p_2}}Y_{2p_2}.
$$

The logistic regression model can be changed to other proper link function for binary response, such as probit. If $\beta_{Y_{21}} = \cdots = \beta_{Y_{2p_2}} = 0$ and at least one coefficient among $\{\beta_{X_1}, \ldots, \beta_{X_q}, \beta_{Y_{11}}, \ldots, \beta_{Y_{1p_1}}\}$ is away from 0, the missing mechanism will be MAR. If $\beta_{Y_{11}} = \cdots, \beta_{Y_{1p_1}} = 0$ and at least one value among $\{\beta_{Y_{21}}, \ldots, \beta_{Y_{2p_2}}\}$ is away from 0, the attrition model will be HW. This additive non-ignorable attrition model can be identified by the panel data and refreshment sample. Note that when we use categorical variables as covariates in the regression model, dummy coding will yield more binary indicator variables depending on the category levels for each variables.

The number of coefficients will be increased. I use one coefficient for each variable here to simply the notation.

3.2.2 Prior on $\beta$

For the prior distribution on $\beta$, I consider three main kinds: diffuse prior, weakly informative prior, and shrinkage prior distributions. As diffuse prior distributions, I use independent and identical flat univariate normal distributions with a relatively large variance. The posterior computation of $\beta$ can be based on conjugate full conditional updates. As weakly informative prior distribution, I use Cauchy distributions
(Gelman et al., 2008) with scale 10 for the intercept and scale 2.5 for the slopes. This Cauchy prior (i.e., Student $t$ distribution with 1 degree of freedom) requires modification of the posterior computation, such as Metropolis updating, the adaptive rejection sampling (Gilks and Wild, 1992), or involvement of the normal-inverse Gamma conjugacy, to achieve better mixing behavior. I describe the computation algorithms in detail in Section 3.3.

Shrinkage prior distributions are useful when we believe non-zero coefficients are sparse. I develop shrinkage prior distributions under the Bayesian Lasso in the generalized linear regression models with logit or probit link functions. To motivate this development under data augmentation, I first review the Bayesian Lasso prior for normal models.

To estimate the regression coefficients $\beta = (\beta_1, \ldots, \beta_p)'$ in the normal model $y_{n1} = \mu_{1n} + X_{n,p} \beta_{p1} + \epsilon, \epsilon \sim N(0, \sigma^2 I_n)$, Lasso (Tibshirani, 1996) estimates are viewed as $L_1$-penalized least squares estimates to achieve the minimum in terms of $\beta$: $\min_{\beta}(\tilde{y} - X\beta)'(\tilde{y} - X\beta) + \lambda \sum_{j=1}^p |\beta_j|$ for some $\lambda \geq 0$, where $\tilde{y} = y - \bar{y}1_n$. Tibshirani (1996) suggests that Lasso estimates can be interpreted as posterior mode estimates when the regression parameters have independent and identical Laplace (i.e., double-exponential) prior distributions. Under a conditional Laplace prior, a fully Bayesian analysis process was proposed by Park and Casella (2008) using $\pi(\beta|\sigma^2) = \prod_{j=1}^p \frac{\lambda}{2\sigma} e^{-\lambda |\beta_j|/\sigma}$. Conditioning on $\sigma^2$ is necessary to guarantee a unimodal posterior distribution. The Laplace distribution can be written as a scale mixture of normal distributions with an exponential mixing density for the scale (Andrews and Mallows, 1974; West, 1987): $\frac{a}{2} e^{-a|x|} = \int_0^{\infty} \frac{1}{\sqrt{2\pi a^2 s}} e^{-x^2/(2as)} \frac{1}{2} e^{-a^2/2s} ds, \ a > 0$.

This representation is used to construct the Gibbs sampler for Bayesian Lasso
through hierarchical forms (Park and Casella, 2008) as follows

\[
(y|\mu, X, \beta, \sigma^2) \sim N_n(\mu 1_n + X \beta, \sigma^2 I_n) \tag{3.2}
\]

\[
(\beta|\sigma^2, \tau_{1}^2, \ldots, \tau_{p}^2) \sim N_p(0, \sigma^2 D_\tau) \tag{3.3}
\]

\[
D_\tau = \text{diag}(\tau_{1}^2, \ldots, \tau_{p}^2) \tag{3.4}
\]

\[
(\sigma^2, \tau_{1}^2, \ldots, \tau_{p}^2) \sim \pi(\sigma^2) \prod_{j=1}^{p} \frac{\lambda^2/2 e^{-\lambda^2 \tau_{j}^2/2}} {\Gamma(\delta_{1}, \delta_{2})}. \tag{3.5}
\]

Full conditional distributions for \(\beta, \sigma^2, \lambda^2\) and \(\tau_{1}^2, \ldots, \tau_{p}^2\) are easy to sample (the intercept \(\mu\) may be given an independent, flat prior, and the full conditional distribution will be normal with mean \(\bar{y}\) and variance \(\sigma^2/n\)). Hans (2009) provides a direct characterization of the posterior computation for the regression coefficients that does not include the intermediary variables \(\tau_{1}^2, \ldots, \tau_{p}^2\).

Both Park and Casella (2008) and Hans (2009) considered normal linear models. I extend to generalized linear models under Bayesian Lasso, where the attrition is the binary response variable. Gramacy and Polson (2012) also developed a new MCMC scheme for regularized logistic regression.

Posterior computation on Bayesian generalized linear models is complicated by the fact that no conjugate prior exists for the parameters other than for normal regression. Albert and Chib (1993) propose data augmentation involving latent Gaussian variables for binary probit regression models, where conjugate priors are available and the MCMC simulation by iteratively sampling from the full conditional distributions for the latent normal variables and model parameters is straightforward and easy to implement. The binary variable \(y_i = 1\) if the latent auxiliary variable \(t_i > 0\), and \(y_i = 0\) otherwise. Assume \(t_i = X_i\beta + \epsilon_i\) and \(\epsilon_i \sim N(0, 1)\).

However, the strong correlation between the latent Gaussian variables and the regression coefficients makes the mixing very slow and sticky. Holmes and Held
(2006) propose joint updating of the latent variables and \( \beta \) to improve the mixing performance and sampling efficiency for the conventional data augmentation algorithm of probit and logistic regression models. I introduce the auxiliary variables during posterior computation.

### 3.2.3 Model for Survey Variables

When the value of \( p_1 \) or \( p_2 \) is large, to impute the missing values in the categorical variables \( Y_1 \) and \( Y_2 \) is a challenging issue, as discussed in Chapter 2. I propose to adopt the Dirichlet Process Mixture of Products of Multinomial distributions (DPMPM) for panel study with refreshment samples. In particular, I use one joint model for the demographic variables \( X \), the response variables in the first wave \( Y_1 \) and the response variables in the second wave \( Y_2 \). They are assumed as exchangeable given the latent class.

Denote \( Z_{ij} \) as the corresponding category of variable \( j \) for individual \( i \), for \( i = 1, \ldots, N \) and \( j = 1, \ldots, p \). The possible values of \( Z_{ij} \) for variable \( j \) fall in \( \{1, \ldots, d_j\} \), where \( d_j \) is the total number of categories for variable \( j \) (\( d_j \geq 2 \)). Let \( c_j \) be the chosen value of variable \( j \), where \( c_j \in \{1, \ldots, d_j\} \). I obtain a \( d_1 \times d_2 \times \cdots \times d_p \) contingency table with cell \((c_1, \ldots, c_p)\) containing the count \( \sum_{i=1}^{N} I(Z_{i1} = c_1, \ldots, Z_{ip} = c_p) \) and cell probability \( \theta_{c_1,\ldots,c_p} \), where \( \theta_{c_1,\ldots,c_p} = \Pr(Z_{i1} = c_1, \ldots, Z_{ip} = c_p) \), for \( c_j = 1, \ldots, d_j \) and \( j = 1, \ldots, p \). I model the joint cell probabilities \( \Theta = \{\theta_{c_1,\ldots,c_p}\} \). The nonparametric model for mixture of products of multinomial distributions can be specified as

\[
\Pr(Z_{i1} = c_1, \ldots, Z_{ip} = c_p) = \theta_{c_1,\ldots,c_p},
\]

\[
\theta_{c_1,\ldots,c_p} = \sum_{h=1}^{\infty} \pi_h \prod_{j=1}^{p} \psi_{hc_j}^{(j)},
\]

(3.7)

where \( \pi = (\pi_1, \ldots, \pi_\infty)' \) is a vector of class allocation probabilities, and \( s_i \in \{1, \ldots, \infty\} \) is the latent class index. Here, \( \pi_h = \Pr(s_i = h) \), and \( \Pr(Z_{ij} = c_j | s_i = h) = \psi_{hc_j}^{(j)} \) is the probability of \( Z_{ij} = c_j \) given the allocation of individual \( i \) to class \( h \).
To clarify, with the expression of latent class index $s$, the selection model for attrition and the imputation model (AN+DPMPM) under prior specification are summarized as

$$\text{logit/probit } \Pr(W_i = 1) = Z_i\beta$$

(3.8)

$$Z_{ij} \sim \text{Multinomial}\{\{1, \ldots, d_j\}, \psi_{s_{ij}, \ldots, s_{i,d_j}}^{(j)}\}$$

(3.9)

$$s_i \sim \text{Multinomial}(\pi_1, \ldots, \pi_{\infty}) \text{ for all } i$$

(3.10)

$$\pi_h = V_h \prod_{i < h} (1 - V_i) \text{ for } h = 1, \ldots, \infty$$

(3.11)

$$V_h \sim \text{Beta}(1, \alpha)$$

(3.12)

$$\psi_h^{(j)} = (\psi_{h1}^{(j)}, \ldots, \psi_{hd_j}^{(j)}) \sim \text{Dirichlet}(a_{j1}, \ldots, a_{jd_j})$$

(3.13)

$$\beta \sim f(\beta).$$

(3.14)

The prior distribution $f(\beta)$ for $\beta$ can be multivariate normal distributions with mean $\mu_\beta$ and covariance matrix $\Sigma_\beta$, or a weakly informative prior of Cauchy distribution, or a Bayesian Lasso prior. All the $Z_{ij}$ are conditionally independent given the latent class $s_i$ and the stick-breaking random variables $V = \{V_h\}$. The probability vectors $\Psi = \{\psi_h^{(j)}\}$ are exchangeable across different classes.

### 3.3 Posterior Computations

I truncate the infinite stick-breaking variables at a large number $H^*$ and run blocked Gibbs sampler (Ishwaran and James, 2001) for the allocation probabilities and component parameters. The truncated stick-breaking process is expressed as $\pi = \sum_{h=1}^{H^*} \nu_h \psi_h$ or $s_i \sim \sum_{h=1}^{H^*} V_h \prod_{l < h} (1 - V_i) \delta_h$.

When the total number of binary variables after dummy coding (denoted as $p^*$) is modest, Metropolis updates involving dependent accept-reject steps for the logistic regression coefficients $\beta$ can be implemented. I use the MLE of $\beta$ and the
corresponding covariance matrix for initial values and parameters in the proposal multivariate normal distribution for in the Metropolis step conditional on imputed dataset. The MCMC algorithms are easy to implement and the acceptance rate is in a reasonable range.

If the number of binary variables $p^a$ is large, under logistic regression and normal priors, we can use the adaptive rejection sampling (Gilks and Wild, 1992) to draw samples of $\beta$ one by one. Adaptive rejection sampling is proposed to sample from any univariate log-concave probability density function. As the adaptive sampling proceeds, the rejection envelope and the squeezing function converge to the density function. This technique is intended particularly for applications of Gibbs sampling to Bayesian models with non-conjugacy. However, this univariate updating may ignore potentially strong posterior dependence between the coefficients and can be computationally expensive.

I also implement data augmentation that uses auxiliary variables to jointly update the regression coefficients (Holmes and Held, 2006). The posterior computation updates the latent variables $t$ from its marginal distribution after integrating over $\beta$: 

$$
\pi(\beta, t | y) = \pi(t | y) \pi(\beta | t).
$$

For probit regression under the prior $N(0, \nu)$ for $\beta$, we have $\pi(t | w) \propto N(0, I_n + Z \nu Z') \text{Ind}(w, t)$, where $\text{Ind}(w, t)$ is an indicator function which truncates the multivariate normal distribution of $t$ to the appropriate region. Updating the multivariate variables $t$ is hard, but it can be done easily via univariate updates. The univariate updating by Gibbs will be $(t_i | t_{-i}, w_i) \sim N(m_i, v_i)I(t_i > 0)$ if $w_i = 1$, and $N(m_i, v_i)I(t_i \leq 0)$ otherwise, where $t_{-i}$ denotes the auxiliary variables $t$ with the $i$th
variable removed. The related parameters \((m_i, v_i)\) are obtained by
\[
\begin{align*}
\pi(\beta|t) &= N(B, V), \quad B = VZ't, \quad V = (\nu^{-1} + Z'Z)^{-1} \\
m_i &= Z_iB - g_i(t_i - Z_iB), \quad v_i = 1 + g_i \\
g_i &= h_i/(1 - h_i), \quad h_i = (H)_{ii}, \quad H = ZZ'.
\end{align*}
\]
(3.15) (3.16) (3.17)

The posterior mean \(B\) must be recalculated after updating to each \(t_i\) using \(B = B^{old} + S_i(t_i - t_i^{old})\), where \(B^{old}\) and \(t_i^{old}\) denote the values of \(B\) and \(t_i\) before the update of \(t_i\), and \(S_i\) denotes the \(i\)th column of \(S = VZ'\).

For the probit regression, the weakly informative Cauchy prior can be also used with the auxiliary variable models for joint updating. I use the conditional normal-inverse Gamma updating for the mean and variance to replace the marginal updating for the Cauchy distribution in the data augmentation for the binary response variable. The prior distribution \(N(0, \nu)\) for \(\beta\) is replaced by \(N(0, \nu)\) and \(1/\nu \sim \text{Gamma}(1/2, \sigma_0^2/2)\). If \(\sigma_0 = 2.5\), this introduces a Cauchy prior with scale 2.5 on \(\beta\). We need to add one more step by drawing the scales \(\nu\) from conjugate inverse Gamma distribution in the posterior updating.

If the link function is logistic, Holmes and Held (2006) introduce another set of intermediate variables \(\lambda\) to facilitate the Bayesian auxiliary variable models: update \({t, \beta}\) jointly given \(\lambda\): \(\pi(t, \beta|Z, \lambda) = \pi(t|Z, \lambda) \pi(\beta|Z, \lambda)\). We have
\[
\begin{align*}
w_i &= \begin{cases} 1 & \text{if } t_i > 0 \\ 0 & \text{otherwise} \end{cases} \\
t_i &= Z_i\beta + \epsilon_i, \quad \epsilon_i \sim N(0, \lambda_i), \quad \lambda_i = (2\delta_i)^2 \\
\delta_i &\sim \text{Kolmogorov-Smirnov distribution} \\
\beta &\sim N(0, \Sigma_0).
\end{align*}
\]
(3.18) (3.19) (3.20) (3.21)

The full conditional distribution of \(\beta\) given \(t\) and \(\lambda\) is still normal: \((\beta|t, \lambda) \sim N(B, V)\), where \(B = V(Z'Z)^{-1}t, V = (\Sigma_0^{-1} + Z'Z)^{-1}\), and \(\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_N)\). Auxiliary
variables $t$ are updated based on the marginalized distribution after integrating $\beta$ out. We have

$$ (t_i|t_{-i}, w_i) \sim \begin{cases} N(m_i, v_i)I(t_i > 0) & \text{if } w_i = 1 \\ N(m_i, v_i)I(t_i \leq 0) & \text{otherwise} \end{cases} $$ (3.22)

$$ m_i = Z_iB - g_i(t_i - Z_iB), \quad v_i = \lambda_i(1 + g_i) $$ (3.23)

$$ g_i = h_i/(\lambda_i - h_i), \quad h_i = (H)_{ii}, \quad H = ZVZ' $$ (3.24)

Recalculate $B = B^{\text{old}} + S_i(t_i - t_i^{\text{old}})/\lambda_i$ where $S_i$ denotes the $i$th column of $S = VZ'$. $\pi(\lambda|t, \beta)$ does not have a standard form. $\lambda$ is generated using rejection sampling.

I modify and extend Bayesian Lasso and the auxiliary variable models. I combine them to handle Bayesian binary probit regression under the shrinkage prior. The $\sigma^2$ in Bayesian Lasso will be 1 in the probit transformation. I use a flat prior $N(0, 1000)$ on the intercept inside $\beta$. The summary of the modeling process and prior distributions are shown following the notation for the attrition model with the link of probit as below

$$ w_i = \begin{cases} 1 & \text{if } t_i > 0 \\ 0 & \text{otherwise} \end{cases} $$ (3.25)

$$ t_i = Z_i\beta + \epsilon_i, \quad \epsilon_i \sim N(0, 1) $$ (3.26)

$$ (\beta|\tau_1^2, \ldots, \tau_p^2) \sim N_p(0_{p+1}, D_{\tau}), \quad D_{\tau} = \text{diag}(1000, \tau_1^2, \ldots, \tau_p^2) $$ (3.27)

$$ (\tau_1^2, \ldots, \tau_p^2) \sim \prod_{j=1}^{p} \lambda^2/2e^{-\lambda^2\tau_j^2/2}, \quad \lambda^2 \sim \text{Gamma}(\delta_1, \delta_2). $$ (3.28)

The posterior updating under Bayesian Lasso draws samples from full conditional distributions for the latent auxiliary variables and the parameters. It proceeds as follows

1. Update $\beta$ from

$$ (\beta|\cdot) \sim N(B, V), $$
where $B = (D_{\tau}^{-1} + Z'Z)^{-1}Z't$ and $V = (D_{\tau}^{-1} + Z'Z)^{-1}$.

(2). Update auxiliary variables $t$ from $N(0, I_n + Z D_t Z') \text{Ind}(w, t)$

$$
(t_i | t_{-i}, w_i) \sim \begin{cases} 
N(m_i, v_i) I(t_i > 0) & \text{if } w_i = 1 \\
N(m_i, v_i) I(t_i \leq 0) & \text{otherwise}
\end{cases}
$$

$$
m_i = Z_i B - g_i(t_i - Z_i B), \quad v_i = 1 + g_i
$$

$$
g_i = h_i/(1 - h_i), \quad h_i = (H)_{ii}, \quad H = ZVZ'.
$$

Recalculate $B = B^{\text{old}} + S_i(t_i - t_i^{\text{old}})$ where $S_i$ denotes the $i$th column of $S = VZ'$.

(3). Update $\tau_j^2$ with $1/\tau_j^2$, independently for $j = 1, \ldots, p$, conditionally drawn from Inverse-Gaussian distribution with parameters $\mu' = \sqrt{\lambda^2/\beta_j^2}$ and $\lambda' = \lambda^2$ in the parameterization of the Inverse-Gaussian density given by

$$
f(x) = \sqrt{\frac{\nu}{2\pi}} x^{-\nu/2} \exp\left\{ -\frac{(x - \mu')^2}{2\mu' x} \right\}
$$

for $x > 0$. To generate random variables from an Inverse-Gaussian distribution

- simulate $\nu \sim N(0, 1)$,
- $y = \nu^2$,
- $x = \mu' + \mu'^2 y/(2\lambda') - \mu'/(2\lambda') \sqrt{4\mu' \lambda' y + \mu'^2 y^2}$,
- simulate $z \sim U(0, 1)$,
- if $z \leq \mu'/(\mu' + x)$ then return $x$; else return $\mu'^2/x$.

The inverse of the generated value will be the value for $\tau_j^2$ for this iteration.

(4). Update $\lambda^2$ from

$$
(\lambda^2 | -) \sim \text{Gamma}(\delta_1 + p, \text{rate} = \delta_2 + \sum_{j=1}^{p} \tau_j^2/2).
$$
The steps for posterior computations for the related parameters in the whole semi-parametric \( AN + DPMPM \) models and imputation for missing values are specified as following

**Step 1:** Update \( s_i \in \{1, \ldots, H^*\} \) from multinomial distribution with probabilities

\[
\Pr(s_i = h|\cdot) = \frac{\pi_h \prod_{j=1}^{p} \psi_{hZ_{ij}}^{(j)}}{\sum_{k=1}^{H^*} \pi_k \prod_{j=1}^{p} \psi_{kZ_{ij}}^{(j)}},
\]

where \( Z_{ij} \in Z_{com} \) and \( Z_{com} = \{X_{mis}, X_{obs}, Y_{1mis}, Y_{1obs}, Y_{2obs}, Y_{2mis}\} \), for \( i = 1, \ldots, N \) and \( j = 1, \ldots, p \).

**Step 2:** Update \( V_h \) from conjugate Beta distributions, for \( h = 1, \ldots, H^* \)

\[
(V_h|\cdot) \sim \text{Beta}(1 + n_h, \alpha + \sum_{k=h+1}^{H^*} n_k),
\]

where \( n_h = \sum_{i=1}^{N} I(s_i = h) \), and calculate \( \pi_h \) from \( \pi_h = V_h \prod_{k<h}(1 - V_k) \).

**Step 3:** Update \( \psi_h^{(j)} = (\psi_{h1}^{(j)}, \ldots, \psi_{hd_j}^{(j)}) \) from conjugate Dirichlet distributions

\[
(\psi_h^{(j)}|\cdot) \sim \text{Dirichlet}(a_{j1} + \sum_{i:s_i = h} I(Z_{ij} = 1), \ldots, a_{jd_j} + \sum_{i:s_i = h} I(Z_{ij} = d_j)).
\]

**Step 4:** Update \( \alpha \) from conjugate Gamma distributions

\[
(\alpha|\cdot) \sim \text{Gamma}(a_\alpha + H^* - 1, b_\alpha - \log \pi_{H^*}).
\]

**Step 5:** Update \( \beta^{(t)} = \{\beta_0, \beta_{\gamma'}, j' = 1, \ldots, p^*\} \) at iteration \( t, t \geq 2 \)

1. **Metropolis algorithm under Cauchy prior:** draw \( \beta^* \) from multivariate normal proposal distribution with mean \( \beta_{Z^*} \) and covariance matrix \( \Sigma_{Z^*} \), where
$Z^*$ is the data frame for $Z$ after dummy coding, $\beta_{Z^*}$ is MLE; accept $\beta^{(t)} = \beta^*$ with probability

$$r = \frac{\exp(Z^*_i \beta^* W_i)}{[1 + \exp(Z^*_i \beta^*)] \prod_{j=1}^{p^*} C(\beta^*_j; 0, 2.5)} \frac{\exp(Z^*_i \beta^{(t-1)} W_i)}{[1 + \exp(Z^*_i \beta^{(t-1)})] \prod_{j=1}^{p^*} C(\beta^{(t-1)}_j; 0, 2.5)},$$

where $C(\beta; 0, v)$ represents the pdf of a Cauchy distribution with mean 0 and scale $v$ at the point $\beta$. Otherwise, set $\beta^{(t)} = \beta^{(t-1)}$.

2. **Adaptive rejection sampling**: draw one sample based on adaptive rejection sampling procedure for each univariate $\beta$.

3. **Joint updating with auxiliary variable**:

   • Under probit link and normal prior $N(0, \Sigma_0)$

   The full conditional distribution of $\beta$ given $t$ is still normal.

   $$ (\beta|t, \lambda) \sim N(B, V), $$

   where $B = V(Z't)$ and $V = (\Sigma_0^{-1} + Z'Z)^{-1}$. Auxiliary variables $t$ are updated based on the marginalized distribution after integrating $\beta$ out.

   $$ (t_i|t_{-i}, w_i) \sim \begin{cases} N(m_i, v_i) I(t_i > 0) & \text{if } w_i = 1 \\ N(m_i, v_i) I(t_i \leq 0) & \text{otherwise}, \end{cases} $$

   where $m_i = Z_iB - g_i(t_i - Z_iB)$, $v_i = 1 + g_i$, $g_i = h_i/(1 - h_i)$, $h_i = (H)_{ii}$, and $H = ZVZ'$. Then recalculate $B = B^{\text{old}} + S_i(t_i - t_i^{\text{old}})$.

   • Under logit link and normal prior $N(0, \Sigma_0)$

   The full conditional distribution of $\beta$ given $t$ and $\lambda$ is still normal

   $$ (\beta|t, \lambda) \sim N(B, V). $$
where $B = V(Z^\prime \Lambda^{-1} t)$, $V = (\Sigma_0^{-1} + Z^\prime \Lambda^{-1} Z)^{-1}$, and $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$.

Auxiliary variables $t$ are updated based on the marginalized distribution after integrating $\beta$ out

$$
(t_i | t_{-i}, w_i) \sim \begin{cases} 
N(m_i, v_i)I(t_i > 0) & \text{if } w_i = 1 \\
N(m_i, v_i)I(t_i \leq 0) & \text{otherwise}
\end{cases}.
$$

where $m_i = Z_i B - g_i(t_i - Z_i B)$, $v_i = \lambda_i (1 + g_i)$, $g_i = h_i / (\lambda_i - h_i)$, $h_i = (H)_{ii}$, and $H = ZVZ^\prime$. Recalculate $B = B^{\text{old}} + S_i(t_i - t_i^{\text{old}})/\lambda_i$, where $S_i$ denotes the $i$th column of $S = VZ'$. Then sample $\lambda$ using rejection sampling.

- Under probit link and Cauchy prior $\sigma_0$
  
  Add the step for updating the variance from Inverse-Gamma distributions.

- Under probit link and Bayesian Lasso
  
  See the Steps (1)-(4) in this section.

**Step 6:** Update $Y_{1\text{mis}}$ in the panel and refreshment samples

$$(Y_{i,1j}^\ast | \_\_ \_ ) \sim \text{Multinomial}(\{1, \ldots, d_j\}, \psi_{s_i1}^{(j)}, \ldots, \psi_{s_id_j}^{(j)}).$$

For the missing $Y_1$ values in the panel, accept the new draw with the probability

$$\Pr(W_i = 1 | Y_{i,1j}^\ast, \_\_\_)^{(W_i = 1)} \Pr(W_i = 0 | Y_{i,1j}^\ast, \_\_\_)^{(W_i = 0)}.$$

**Step 7:** Update $Y_{2\text{nmis}-pl}$ in the panel when $W=0$: $Y_{2\text{nmis}}$ is NMAR in the panel and is conditional on the imputation model related terms with $W = 0$. For $\{(i, j) : Y_{i,2j} \in Y_{2\text{nmis}-pl}\}$, I draw $Y_{2\text{nmis}-pl}$ from the conditional distribution on imputation, which is a multinomial distribution

$$(Y_{i,2j}^\ast | \_\_ \_ \_ ) \sim \text{Multinomial}(\{1, \ldots, d_j\}, \psi_{s_i1}^{(j)}, \ldots, \psi_{s_id_j}^{(j)}),$$

and accept draws with probability $\Pr(W_i = 0 | Y_{i,2j}^\ast, \_\_\_ \_)$.
Step 8: Update $Y_{2mis-ref}$ in the refreshment sample due to MAR, which is independent of the attrition model. For $\{(i, j) : Y_{i,2j} \in Y_{2mis-ref}\}$

$$(Y_{i,2j}|\cdot) \sim \text{Multinomial}(\{1, \ldots, d_j\}, \psi_{s_1}, \ldots, \psi_{s_d}).$$

Step 9: Update $Y_{2mis-pl}$ in the panel when $W = 1$. For $\{(i, j) : Y_{i,2j} \in Y_{2mis-pl}\}$, I draw $Y_{2mis-pl}$ from the conditional distribution on imputation

$$(Y_{i,2j}^*|\cdot) \sim \text{Multinomial}(\{1, \ldots, d_j\}, \psi_{s_1}, \ldots, \psi_{s_d}),$$

and accept draws with probability $Pr(W_i = 1|Y_{i,2j}^*, \cdot)$.

Step 10: Update $W_{mis}$ in the refreshment sample from a Bernoulli distribution with success probability

$$Pr(W_i = 1|\cdot) = \begin{cases} \frac{\exp(Z_i^* \beta)}{1 + \exp(Z_i^* \beta)} & \text{logit link} \\ \Phi(Z_i^* \beta) & \text{probit link} \end{cases}$$

Step 11: Update $X_{mis}$ in the panel and in the refreshment sample due to MAR from multinomial distributions. For $\{(i, j) : X_{i,j} \in X_{mis}\}$

$$(X_{i,j}|\cdot) \sim \text{Multinomial}(\{1, \ldots, d_j\}, \psi_{s_1}, \ldots, \psi_{s_d}).$$

For missing $X$ in the panel, the new draw $X_{ij}^*$ will be accepted with probability

$$Pr(W_i = 1|X_{i,j}^*, \cdot)^{(W_i=1)}Pr(W_i = 0|X_{i,j}^*, \cdot)^{(W_i=0)}.$$

3.4 Simulation Study

I use simulation studies to implement and evaluate the whole AN+DPMPM procedure. Here, I focus on the regression coefficients $\beta$ of the AN model for attrition, since the performance of DPMPM has been discussed in Chapter 2 and it works well for plausible imputation. In this simulation study, I only assume the existence of
non-ignorable missingness due to attrition and suppose there is no item nonresponse due to MAR or MCAR. First, I consider a logit link for the attrition model and use Metropolis updating and adaptive rejection sampling algorithms for the posterior computation. Then, under a probit link, I implement the posterior updating with the Bayesian auxiliary variables.

3.4.1 Case 1: Logistic Regression

I use repeated sampling studies to test the validity of the imputation procedure from a frequentist randomization perspective. Consider a two wave panel with modest number of variables. Suppose $Y_1$ has $p_1 = 20$ variables; $Y_2$ has $p_2 = 20$ variables and $X$ has $q = 40$ variables. All are binary. I assume there are four subpopulations and simulate three stick-breaking variables from Beta$(1, \alpha_0 = 2)$. The allocation weights $\pi_0$ for the four subpopulations are $(.6239, .0558, .0131, .3073)$. During each repeated study, I draw the latent class indicator based on $\pi_0$ and generate $N_p = 2000$ samples from the panel study and a refreshment sample with size $N_r = 1000$. The attrition model is a logistic regression model \( \logit \Pr(W=1) = Z\beta \). The true values for the regression coefficients $\beta_0$ are the sequence from -4 to 4 with a lag .1. This is an arbitrary assumption. I also try other arbitrary settings, and the computation procedure is always able to recover the truth. Around 44% of the individuals drop out of the panel during wave 2 after participation in wave 1.

In the MCMC runs, I generate the initial values for $V_h$ independently from Beta$(1, 2)$ and the initial value for each latent class indicator $s_i$ from a multinomial distribution from 1 to 4 with equal probability. The starting values for $\Psi$ are the corresponding marginal probabilities obtained from the initial dataset, where the starting values for the missing data are set as the truth. I also tried arbitrary starting values for the missing values, and the MCMC chain will converge to the truth as long as enough iterations are run. To guarantee all the chains have converged
during the repetition, I set the truth for the missing data as the starting points. I use the weakly informative Cauchy prior distributions as prior on $\beta$: with scale 10 for the intercept and scale 2.5 for the slope. In the Metropolis updating step for the regression coefficients $\beta$, I use the MLE as the mean for the proposed multivariate normal distribution for $\beta$, and the corresponding estimate for the covariance matrix as the covariance matrix, based on the iteratively imputed dataset.

The truncation level in the blocked Gibbs sampler is $H^* = 10$. The MCMC chains, with 10000 iterations and a thinning value of 10, have good mixing behavior and converge to the true values fast even under randomly selected initial values. I collect several marginal probabilities for convergence diagnostics. The posterior mode for the number of distinctly occupied classes is 6.

I run 50 repeated sampling studies. Figure 3.1 displays the key results from the simulations. As shown in the left-top plot of Figure 3.1, the AN + DPMPM imputation can recover the truth of $\beta$. For the 50 runs, the average values of acceptance rates in the Metropolis step are between .08 to .28, which is a reasonable range. The average values of the Hellinger distance are between .380 and .395, which means the imputed data distributions are reasonably stationary across the repetitions. The 95% Bayesian credible intervals are reasonably well-calibrated with coverage rates only within simulation error of .95.

I also replace the Metropolis updating step for the coefficients with adaptive rejection sampling (ARS), using $N(0, 25)$ as prior distribution for each coefficient. I only run $T = 5000$ iterations here because the computation speed for the ARS is slow, compared to the Metropolis-Gibbs updating algorithm. But even for the small number of iterations, the performance for the ARS-Gibbs algorithm works well and the posterior mean estimates are close to the truth.
Figure 3.1: Simulation outputs under logit link. Top-left: posterior mean and (2.5%, 97.5%) quantiles of $\beta$ for one simulation; top-right: nominal coverage rates for 95% credible intervals of $\beta$, bottom-left: acceptance rates for the M-H step, and bottom-right: the Hellinger distance between imputed and original datasets.
3.4.2 Case 2: Bayesian Probit Auxiliary Variable Models

In this simulation, I use a flat prior \( N(0, 1000I_{p+1}) \) for \( \beta \). The other settings are the same in Case 1. I introduce an auxiliary variable for updating \( \beta \) in the data augmentation step for the probit model. I change the initial values for \( \beta \) and run multiple chains for the convergence check. The outputs are illustrated in Figure 3.2. The convergence behavior is good and the truth can be recovered no matter what the initial values are. Fifty repeated sampling studies show that the nominal coverage rates for 95% Bayesian posterior credible intervals are close to 95%.

Table 3.2 presents the marginal probabilities for \( Pr(Y_2 = 1) \) and the conditional probabilities \( Pr(Y_2 = 1 \mid W) \), where \( W=0/1 \), estimated on only the completed panel data, only the refreshment sample and both the completed panel and refreshment sample (all). Table 3.3 displays the joint probabilities \( Pr(Y_2 = 1, W) \), where \( W = \)
Table 3.2: The marginal probability $Pr(Y_2 = 1)$ and the conditional probability $Pr(Y_2 = 1 \mid W)$ for all measures in wave 2 after imputation for completed panel data, refreshment sample and all data for semi-parametric model simulation study

<table>
<thead>
<tr>
<th>variable</th>
<th>$Y$</th>
<th>$(Y \mid W = 1)$</th>
<th>$(Y \mid W = 0)$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>p</td>
<td>r</td>
<td>p</td>
</tr>
<tr>
<td>$Y_{2,1}$</td>
<td>.716</td>
<td>.710</td>
<td>.676</td>
</tr>
<tr>
<td>$Y_{2,2}$</td>
<td>.691</td>
<td>.688</td>
<td>.647</td>
</tr>
<tr>
<td>$Y_{2,3}$</td>
<td>.698</td>
<td>.694</td>
<td>.649</td>
</tr>
<tr>
<td>$Y_{2,4}$</td>
<td>.673</td>
<td>.714</td>
<td>.630</td>
</tr>
<tr>
<td>$Y_{2,5}$</td>
<td>.719</td>
<td>.723</td>
<td>.660</td>
</tr>
<tr>
<td>$Y_{2,6}$</td>
<td>.695</td>
<td>.702</td>
<td>.652</td>
</tr>
<tr>
<td>$Y_{2,7}$</td>
<td>.718</td>
<td>.703</td>
<td>.666</td>
</tr>
<tr>
<td>$Y_{2,8}$</td>
<td>.684</td>
<td>.700</td>
<td>.631</td>
</tr>
<tr>
<td>$Y_{2,9}$</td>
<td>.702</td>
<td>.707</td>
<td>.642</td>
</tr>
<tr>
<td>$Y_{2,10}$</td>
<td>.693</td>
<td>.705</td>
<td>.641</td>
</tr>
<tr>
<td>$Y_{2,11}$</td>
<td>.711</td>
<td>.710</td>
<td>.657</td>
</tr>
<tr>
<td>$Y_{2,12}$</td>
<td>.702</td>
<td>.717</td>
<td>.662</td>
</tr>
<tr>
<td>$Y_{2,13}$</td>
<td>.680</td>
<td>.699</td>
<td>.646</td>
</tr>
<tr>
<td>$Y_{2,14}$</td>
<td>.697</td>
<td>.723</td>
<td>.657</td>
</tr>
<tr>
<td>$Y_{2,15}$</td>
<td>.701</td>
<td>.696</td>
<td>.643</td>
</tr>
<tr>
<td>$Y_{2,16}$</td>
<td>.710</td>
<td>.713</td>
<td>.629</td>
</tr>
<tr>
<td>$Y_{2,17}$</td>
<td>.706</td>
<td>.717</td>
<td>.654</td>
</tr>
<tr>
<td>$Y_{2,18}$</td>
<td>.699</td>
<td>.720</td>
<td>.633</td>
</tr>
<tr>
<td>$Y_{2,19}$</td>
<td>.715</td>
<td>.704</td>
<td>.627</td>
</tr>
<tr>
<td>$Y_{2,20}$</td>
<td>.692</td>
<td>.714</td>
<td>.634</td>
</tr>
</tbody>
</table>

0/1, estimated on only the completed panel data, only the refreshment sample and both the completed panel and refreshment sample (all). We can see that the panel data and the refreshment sample have the same marginal, conditional and joint distributions for $Y_2$. I also check this for $Y_1$ and find the same conclusion. The conditional probability $Pr(Y_2 = 1 \mid W = 1)$ is different from $Pr(Y_2 = 1 \mid W = 0)$, which means that $Y_2$ depends on $W$ and the attrition is non-ignorable.

3.4.3 Case 3: Regularized Probit Auxiliary Variable Models

These posterior computation algorithms work very well in my simulation studies. However, my APYN data analysis shows that under flat prior distributions, Bayesian
Table 3.3: The joint probability $Pr(Y_2 = 1, W)$ for all measures in wave 2 after imputation, for completed panel data, refreshment sample and all data for semi-parametric model simulation study.

<table>
<thead>
<tr>
<th>variable</th>
<th>$(Y, W = 1)$</th>
<th>$(Y, W = 0)$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>p r all</td>
<td>p r all</td>
</tr>
<tr>
<td>$Y_{2,1}$</td>
<td>.386 .374 .382</td>
<td>.330 .336 .332</td>
</tr>
<tr>
<td>$Y_{2,2}$</td>
<td>.370 .348 .362</td>
<td>.322 .340 .328</td>
</tr>
<tr>
<td>$Y_{2,3}$</td>
<td>.371 .367 .369</td>
<td>.327 .327 .327</td>
</tr>
<tr>
<td>$Y_{2,4}$</td>
<td>.360 .367 .362</td>
<td>.313 .347 .324</td>
</tr>
<tr>
<td>$Y_{2,5}$</td>
<td>.377 .362 .372</td>
<td>.342 .361 .348</td>
</tr>
<tr>
<td>$Y_{2,6}$</td>
<td>.372 .363 .369</td>
<td>.323 .339 .328</td>
</tr>
<tr>
<td>$Y_{2,7}$</td>
<td>.380 .365 .375</td>
<td>.338 .338 .338</td>
</tr>
<tr>
<td>$Y_{2,8}$</td>
<td>.361 .353 .358</td>
<td>.324 .347 .331</td>
</tr>
<tr>
<td>$Y_{2,9}$</td>
<td>.367 .368 .367</td>
<td>.335 .339 .336</td>
</tr>
<tr>
<td>$Y_{2,10}$</td>
<td>.366 .353 .362</td>
<td>.327 .352 .335</td>
</tr>
<tr>
<td>$Y_{2,11}$</td>
<td>.375 .376 .375</td>
<td>.336 .334 .335</td>
</tr>
<tr>
<td>$Y_{2,12}$</td>
<td>.378 .369 .375</td>
<td>.324 .348 .332</td>
</tr>
<tr>
<td>$Y_{2,13}$</td>
<td>.369 .368 .369</td>
<td>.311 .331 .318</td>
</tr>
<tr>
<td>$Y_{2,14}$</td>
<td>.375 .352 .367</td>
<td>.322 .371 .338</td>
</tr>
<tr>
<td>$Y_{2,15}$</td>
<td>.367 .354 .363</td>
<td>.334 .342 .336</td>
</tr>
<tr>
<td>$Y_{2,16}$</td>
<td>.359 .358 .359</td>
<td>.351 .355 .352</td>
</tr>
<tr>
<td>$Y_{2,17}$</td>
<td>.374 .352 .366</td>
<td>.332 .365 .343</td>
</tr>
<tr>
<td>$Y_{2,18}$</td>
<td>.362 .353 .359</td>
<td>.338 .367 .347</td>
</tr>
<tr>
<td>$Y_{2,19}$</td>
<td>.358 .357 .358</td>
<td>.357 .347 .353</td>
</tr>
<tr>
<td>$Y_{2,20}$</td>
<td>.362 .348 .357</td>
<td>.330 .366 .342</td>
</tr>
</tbody>
</table>

Probit auxiliary variable models tend to suffer from sticky mixing problems. I have to either put an informative prior distribution on $\beta$, such as, $N(0,1)$, or use the shrinkage prior distribution to improve the mixing performance.

To illustrate this, I simulate $p_1 = p_2 = 10$ and $q = 30$ binary variables. The attrition model is a probit regression model $Pr(W=1) = \Phi(Z\beta)$. I use the Bayesian auxiliary variable model (Holmes and Held, 2006) to facilitate the computation for probit models and implement Bayesian Lasso (Park and Casella, 2008) to shrink the estimate of the regression coefficients $\beta$ to be sparse. Assume the true values $\beta_t$ are $\{\log(.36*1), \ldots, \log(.36*10), .01, \ldots, .01, .18*1, \ldots, .18*9\}$. Around 25% samples drop out of the panel. I use the true values $\beta_t$ as the initial values $\beta^{(0)}$. The hyper
prior distribution for the tuning parameter $\lambda$ in the Bayesian Lasso is Gamma$(1, 1)$ and the initial value $\lambda^{(0)}$ is set as .1. The initial value of $\tau_j^2$ is the inverse of the generated variable from Inverse-Gaussian distribution with parameters $\sqrt{\lambda^{(0)}^2 / \beta_j^{(0)}^2}$ and $\lambda^{(0)}^2$, for $j = 1, \ldots, p$. I choose a flat $N(0, 1000)$ as the prior distribution of the intercept $\beta_{1}^{(0)}$. I assume the same initial settings for other parameters as the previous simulation study (the cluster structure, sample size and component probabilities). I run the MCMC chain for $T = 50000$ iterations with a burn-in value of 2000 and a thinning value of 100.

After convergence check, the outputs in Figure 3.3 show that the posterior mean values of the regression coefficients are shrunk towards 0. Larger shrinking effects correspond to larger values of the scale $\tau$. The posterior mean for the tuning parameter $\lambda$ in the double exponential prior distribution is 3, which illustrates the shrinkage effect. Figure 3.3 illustrates that the posterior computation algorithm under shrinkage prior distributions works well but has odd behaviors for the last four coefficients related to $Y_2$. This shows that the shrinkage may affect the contribution of $Y_2$ to $W$ and therefore the imputation performance for $Y_2$, even though the marginal probabilities of $Y_2$ can be recovered well. Shrinkage leads the missingness mechanisms to be close to MAR or even MCAR.

I also implement the Bayesian auxiliary variable models for logistic regression. Assume a prior $N(0, 100I_{p+1})$ on $\beta$ and update $\lambda$ by rejection sampling process. However, the mixing is very sticky here, because of very correlated behaviors used by the latent intermediary variable $\lambda$ for each individual.

3.5 AP Yahoo News Panel Data Analysis

The 2007-2008 Associated Press/Yahoo News election panel (APYN) was conducted among the probability-based, web-enabled KnowledgePanel over the internet. It
was a one-year survey with 11 interview waves from November 2007 to December 2008. The study uses known published sampling frames that cover 99% of the US population. Most of the demographic information for the participants is matched from this public population profile. Sampled non-internet households are provided a laptop computer or MSN TV unit and free internet service.

One fundamental motivation of the APYN project was to study the electoral behaviors and attitudes toward the upcoming 2008 Presidential election. The baseline survey (wave 1) fielded 3548 KnowledgePanel members in November 2007, and 2735 cases responded. After the initial wave, the baseline respondents who stayed in the panel were fielded follow-up waves. Wave-to-wave attrition rates or completion rates
vary across time towards the end of the panel. Three external refresh cross sections were also collected: a sample of 697 new respondents in wave 3 (January, 2008), 576 new respondents in wave 6 (September, 2008), and 464 new respondents in wave 9 (October, 2008). I focus on the baseline survey in November 2007 as wave 1 and the September 2008 survey as wave 2 to conduct a two wave panel study with one refreshment sample in the second wave. In wave 2, 1724 members remained in the panel while 1011 (36.97%) respondents dropped out after participating in wave 1.

The APYN study involves various measures on Presidential candidate preferences. Typically these measures can be divided into two main classes. One is about voting participation and the other is on voting choices. The variables I collect for the selection models are summarized in Table 3.4. For example, the variables beginning with $FAV$ directly measure attitudes towards political figures (Barack Obama, John McCain, etc) and the parties (The Democratic Party and The Republican Party); the variables beginning with $LV$ evaluate the participants’ voting participation (registered or not, how much campaign interest, likelihood to vote, etc). These measures on campaign interests are strong predictors of political attitudes and behaviors (Prior, 2010) and key measures for defining likely voters in pre-election polls (Traugott and Tucker, 1984). Bartels (1999) points out that campaign interest is related to panel attrition.

These questions are included in both wave 1 and wave 2 panel. The corresponding answers are either nominal or ordinal with several fixed number of scales. All questions allow for “Not asked”, “Refused/Not Answered” and “Don’t know enough to say” response. I treat these three category levels as item nonresponse, which exists in the panel and refreshment samples. The counts and proportions of item nonresponse for the measures of participants in wave 1, the remaining participants in wave 2 and the new participants in the refreshment sample, are shown in Table 3.4. All variables are subject to item nonresponse, with missing rates varying from
Table 3.4: Collected outcome measures and summary on item nonresponse for semi-parametric selection models.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Interpretation</th>
<th>Levels</th>
<th>W1 (2735)</th>
<th>W2 (1724)</th>
<th>Ref (464)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FAV1-4</td>
<td>Barack Obama</td>
<td>4</td>
<td>550 (2.11)</td>
<td>95 (5.51)</td>
<td>20 (4.31)</td>
</tr>
<tr>
<td>FAV1-9</td>
<td>John McCain</td>
<td>4</td>
<td>709 (25.92)</td>
<td>95 (5.51)</td>
<td>23 (4.96)</td>
</tr>
<tr>
<td>LV1</td>
<td>Are you registered to vote?</td>
<td>2</td>
<td>9 (.33)</td>
<td>8 (.46)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>LV3</td>
<td>interest about campaign</td>
<td>4</td>
<td>6 (.22)</td>
<td>5 (.29)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>CND1</td>
<td>thought to candidates</td>
<td>2</td>
<td>5 (.18)</td>
<td>8 (.46)</td>
<td>3 (.65)</td>
</tr>
<tr>
<td>LV31</td>
<td>likelihood to vote</td>
<td>4</td>
<td>18 (.66)</td>
<td>220 (12.76)</td>
<td>57 (12.28)</td>
</tr>
</tbody>
</table>

.18% up to 25.92% in wave 1, from .29% up to 12.76% for the stayed individuals in wave 2 and from .65% up to 12.28% in the refreshment sample.

The demographic information about the respondents were mainly collected before the study from the public population profiles, as shown in Table 3.5. Age, income, education, race, marriage status, and household information are complete for the individuals in the panel and refreshment samples. PPAGECT4 includes 4 categories: age 18-29, 30-44, 45-59, and 60+. PPEDUCAT includes 4 categories: Less than high school, High school, Some college, and Bachelor’s degree or higher. PPGENDER includes 2 categories: male and female. PPETHM includes 2 categories: White & Non-Hispanic and others. PPINCIMP includes 4 categories: Less than $29,999, $30,000 to $49,999, $50,000 to $74,999, and $75,000 or more. PPMARIT includes 2 categories: married and others including single, divorced, etc. I also use several variables collected in both the panel and refreshment sample of this study, and match them to be treated as demographic information. PARTYID is a recoded variable: do you think yourself a democrat, republican, or neither of the two? ID1 asks the participants whether they consider themselves as liberal/ moderate/conservative. REL3 describes how often attending religious services: more than once a week, once a week, a few times a month, a few times a year, and never. All these demographic variables are subject to item nonresponse. I assume the missing values due to MAR.
Table 3.5: Collected demographic variables and summary on item nonresponse for semi-parametric selection models.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Interpretation</th>
<th>Levels</th>
<th>Panel (2735)</th>
<th>Ref (464)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PPAGECT4</td>
<td>Age</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PPEDUCAT</td>
<td>Education</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PPETHM</td>
<td>Race/Ethnicity</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PPGENDER</td>
<td>Gender</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PPINCIMP</td>
<td>Household Income</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PPMARIT</td>
<td>Marital Status</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PARTYID</td>
<td>party</td>
<td>3</td>
<td>10 (.37)</td>
<td>7 (1.51)</td>
</tr>
<tr>
<td>ID1</td>
<td>liberal or conservative</td>
<td>5</td>
<td>57 (2.08)</td>
<td>10 (2.16)</td>
</tr>
<tr>
<td>REL3</td>
<td>attend religious services</td>
<td>5</td>
<td>20 (.73)</td>
<td>8 (1.72)</td>
</tr>
</tbody>
</table>

and impute them in the procedure.

I aggregate categories for LV3, CND1 and LV31 from the original data. These adjustments are due to the extremely small marginal probabilities for some level of these variables. During dummy coding I always use the last category as the reference level.

I first calculate the marginal probabilities for the variables with missing values from their observed parts, and use these marginal probability distributions to simulate the initial values for the missing data. For the missing values of $Y_2$ due to attrition in the panel, I calculate the conditional probabilities $\Pr(Y_2|W=0)$ using the marginal distributions in the refreshment, the marginal distributions estimated on the subjects stayed in wave 2, and the probability $\Pr(W=1|0)$ based on the panel. Bayes rule is used to obtain the marginal probability distributions for $Y_2$ of dropout subjects. The corresponding initial values are generated based on the calculated probabilities. I use the maximum likelihood estimation of $\beta$, based on the panel data after initializing the missing values and dummy coding, as the starting points for $\beta$. The initial values for $W$ in the refreshment samples are independently drawn from a Bernoulli distribution with success probability $N_{cp}/N_p = .6303$. I also tried the
initial values of $W$ obtained on the starting points of $\beta$ and missing values of $Z$. The outputs are consistent.

The starting value for $\alpha$ is 1, and I generate the stick-breaking variables $V_h$ independently from Beta(1, 1) for $h = 1, \ldots, H^*$, where $H^* = 20$ as the truncation level in the blocked Gibbs sampler. This yields the starting value for $\pi$. The latent class indicator $s$ is drawn from a multinomial distribution with probability vector $\pi$ and sample size $N$. The starting values for $\Psi$ are the corresponding marginal probabilities calculated from the initial completed datasets.

The prior distributions on the parameters are set as flat as possible: Beta(1/4, 1/4) on $\alpha$, Dirichlet $(1, \ldots, 1)$ on $\Psi$, and $N(0, 1)$ independently on the regression coefficients $\beta$ in the probit model for attrition. I run the MCMC chain for $T = 500000$ iterations with a burn-in value of 300000 and a thinning value of 200. For convergence diagnostics, I select the marginal probabilities for all the variables to fall into the first category, which are weighted average across all possible clusters. I also look at the posterior samples of $\alpha$ and the posterior samples for $\Pr(W=1)$. The trace plots for $\beta$ and the marginal probabilities demonstrate convergence and good mixing behaviors. The posterior mean for $\Pr(W=1)$ in the refreshment sample is .7399, which is different from the value in the panel data (.6303). The posterior median value for $\alpha$ is 4.1173 and the mode for the posterior samples of the number of distinctly occupied clusters is 19.

The posterior mean values for the 51 coefficients $\beta$ are shown in Figure 3.4. The order follows dummy variables corresponding to $X$, $Y_1$ and $Y_2$. We can see that the coefficients related to the outcome measures $Y_2$ have relatively larger effect towards the attrition than the demographic variables and the measures $Y_1$. The variables LV3 (interest on campaign news) and CND1 (thoughts given to candidates) are strong predictors for attrition.

Table 3.6 presents the posterior median and 95% Bayesian posterior credible
Figure 3.4: The posterior mean estimates for regression coefficients $\beta$ in the probit attrition model with prior $N(0,1)$ of APYN study.

interval for the regression coefficients on the demographic variables. From the estimates and the 95% confidence intervals, we can infer that: Republican interviewers tend to stay in the panel compared to those from Independent or some other party; participants who attend religious a few times one year have more tendency to drop out than those who never attend; older (60+) tend to stay in the panel comparing to the young (18-29); participants with education level of high school or college, have more tendency for attrition than those with a Bachelor’s degree or higher; white participants have less tendency for dropout than other races; female have more tendency for attrition than male; and married participants have more tendency for attrition than those with other marriage status.

Table 3.7 displays the posterior median and 95% Bayesian posterior credible intervals for regression coefficients on outcome variables of wave 1 and wave 2 for the
Table 3.6: Posterior median and 95% Bayesian posterior credible interval for regression coefficients on demographic variables for the attrition model.

<table>
<thead>
<tr>
<th>$\beta$</th>
<th>posterior median</th>
<th>95% credible interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1.2178</td>
<td>(-.1338 2.4546)</td>
</tr>
<tr>
<td>PARTYID (Democrat)</td>
<td>.0735</td>
<td>(-.1422 .2939)</td>
</tr>
<tr>
<td>PARTYID (Republican)</td>
<td>.1483</td>
<td>(-.0766 .3908)</td>
</tr>
<tr>
<td>ID1 (very liberal)</td>
<td>-.0992</td>
<td>(-.5185 .2956)</td>
</tr>
<tr>
<td>ID1 (somewhat liberal)</td>
<td>-.0125</td>
<td>(-.3490 .2884)</td>
</tr>
<tr>
<td>ID1 (moderate)</td>
<td>.0388</td>
<td>(-.2423 .2975)</td>
</tr>
<tr>
<td>ID1 (very conservative)</td>
<td>-.0822</td>
<td>(-.3277 .1576)</td>
</tr>
<tr>
<td>REL3 (more than once a week)</td>
<td>-.0659</td>
<td>(-.3000 .1671)</td>
</tr>
<tr>
<td>REL3 (once a week)</td>
<td>-.0303</td>
<td>(-.2082 .1407)</td>
</tr>
<tr>
<td>REL3 (a few times a month)</td>
<td>-.1913</td>
<td>(-.4408 .0701)</td>
</tr>
<tr>
<td>REL3 (a few times a year)</td>
<td>.0015</td>
<td>(-.1590 .1610)</td>
</tr>
<tr>
<td>AGE (18-29)</td>
<td>-.2857</td>
<td>(-.5056 -.0780)</td>
</tr>
<tr>
<td>AGE (30-44)</td>
<td>-.0801</td>
<td>(-.2701 .1118)</td>
</tr>
<tr>
<td>AGE (45-59)</td>
<td>-.0529</td>
<td>(-.2234 .1288)</td>
</tr>
<tr>
<td>EDU (less than high school)</td>
<td>-.2233</td>
<td>(-.4853 .0240)</td>
</tr>
<tr>
<td>EDU (high school)</td>
<td>-.2386</td>
<td>(-.4331 -.0619)</td>
</tr>
<tr>
<td>EDU (some college)</td>
<td>-.2767</td>
<td>(-.4392 -.1104)</td>
</tr>
<tr>
<td>RACE (white &amp; non-Hispanic)</td>
<td>.0910</td>
<td>(-.0796 .2534)</td>
</tr>
<tr>
<td>GENDER (male)</td>
<td>.0967</td>
<td>(-.0208 .2270)</td>
</tr>
<tr>
<td>INC (less than $29,999)</td>
<td>-.0004</td>
<td>(-.1970 .1773)</td>
</tr>
<tr>
<td>INC ($30,000 to $49,999)</td>
<td>.0271</td>
<td>(-.1556 .2265)</td>
</tr>
<tr>
<td>INC ($50,000 to $74,999)</td>
<td>.0852</td>
<td>(-.1086 .2684)</td>
</tr>
<tr>
<td>MARIT (married)</td>
<td>-.1924</td>
<td>(-.3362 -.0554)</td>
</tr>
</tbody>
</table>

attrition model. In wave 1, $\beta$ for LV3 shows that participants with a great deal or quite a bit of interest following news about the campaign have more tendency to drop out than those with very little interest. In wave 2 $\beta$ for LV3 tells that participants with a great deal or quite a bit of interest following news about the campaign tend to stay in the panel than those with very little interest. In wave 2, the $\beta$ for FAV1-4 shows that participants who have very favorable impression on Barack Obama have less tendency for dropout that those who are very unfavorable towards him. In wave 2, the $\beta$ for FAV1-4 shows that participants who have very favorable impression on Barack Obama have more tendency for dropout that those who are very unfavorable
Table 3.7: Posterior median estimates and 95% Bayesian posterior credible intervals for regression coefficients on outcome variables for the attrition model.

<table>
<thead>
<tr>
<th></th>
<th>posterior median</th>
<th>95% credible interval</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>wave 1</strong> Y₁</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FAV1-4 (very favorable)</td>
<td>.4445</td>
<td>(.0576 .8175)</td>
</tr>
<tr>
<td>FAV1-4 (somewhat favorable)</td>
<td>.1146</td>
<td>(-.1603 .3699)</td>
</tr>
<tr>
<td>FAV1-4 (somewhat unfavorable)</td>
<td>-.0424</td>
<td>(-.2974 .1843)</td>
</tr>
<tr>
<td>FAV1-9 (very favorable)</td>
<td>-.1148</td>
<td>(-.5573 .2544)</td>
</tr>
<tr>
<td>FAV1-9 (somewhat favorable)</td>
<td>-.0660</td>
<td>(-.3837 .2353)</td>
</tr>
<tr>
<td>FAV1-9 (somewhat unfavorable)</td>
<td>-.0165</td>
<td>(-.3061 .2472)</td>
</tr>
<tr>
<td>LV1 (registered )</td>
<td>-.3078</td>
<td>(-.8572 .1736)</td>
</tr>
<tr>
<td>LV3 (a great deal)</td>
<td>-.3196</td>
<td>(-.7095 .0040)</td>
</tr>
<tr>
<td>LV3 (quite a bit)</td>
<td>-.2278</td>
<td>(-.5818 .0537)</td>
</tr>
<tr>
<td>LV3 (only some)</td>
<td>-.1309</td>
<td>(-.3402 .0826)</td>
</tr>
<tr>
<td>CND1 (a lot)</td>
<td>-.0412</td>
<td>(-.2396 .1485)</td>
</tr>
<tr>
<td>LV31 (certain will not vote 1)</td>
<td>.2870</td>
<td>(-.2255 .7181)</td>
</tr>
<tr>
<td>LV31 (certain will vote 2)</td>
<td>.0046</td>
<td>(-.3374 .3095)</td>
</tr>
<tr>
<td>LV31 (certain will vote 3)</td>
<td>.1358</td>
<td>(-.1768 .4534)</td>
</tr>
<tr>
<td><strong>wave 2</strong> Y₂</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FAV1-4 (very favorable)</td>
<td>-.8405</td>
<td>(-1.7521 .0793)</td>
</tr>
<tr>
<td>FAV1-4 (somewhat favorable)</td>
<td>-.2416</td>
<td>(-.7960 .2603)</td>
</tr>
<tr>
<td>FAV1-4 (somewhat unfavorable)</td>
<td>-.2875</td>
<td>(-.7389 .1661)</td>
</tr>
<tr>
<td>FAV1-9 (very favorable)</td>
<td>-.3508</td>
<td>(-1.2759 .5039)</td>
</tr>
<tr>
<td>FAV1-9 (somewhat favorable)</td>
<td>-.3919</td>
<td>(-1.2150 .3398)</td>
</tr>
<tr>
<td>FAV1-9 (somewhat unfavorable)</td>
<td>.0251</td>
<td>(-.5283 .5444)</td>
</tr>
<tr>
<td>LV1 (registered )</td>
<td>.7701</td>
<td>(-.2869 2.0617)</td>
</tr>
<tr>
<td>LV3 (a great deal)</td>
<td>2.2434</td>
<td>(1.5866 2.9423)</td>
</tr>
<tr>
<td>LV3 (quite a bit)</td>
<td>1.9705</td>
<td>(1.3846 3.1238)</td>
</tr>
<tr>
<td>LV3 (only some)</td>
<td>.2358</td>
<td>(-1.627 .6315)</td>
</tr>
<tr>
<td>CND1 (a lot)</td>
<td>-1.8289</td>
<td>(-2.4085 -1.3648)</td>
</tr>
<tr>
<td>LV31 (certain will not vote 1)</td>
<td>-.2849</td>
<td>(-1.4451 1.0705)</td>
</tr>
<tr>
<td>LV31 (certain will vote 2)</td>
<td>-.4305</td>
<td>(-1.1143 .4114)</td>
</tr>
<tr>
<td>LV31 (certain will vote 3)</td>
<td>-.2425</td>
<td>(-.7637 .3252)</td>
</tr>
</tbody>
</table>

towards him. The tendency for attrition of these two variables is not stationary or consistent across the waves. The $\beta$ for CND1 tells that participants with more thoughts on Presidential candidates tend to drop out in wave 2 comparing to those with fewer thoughts.
I collect $m = 50$ multiple completed datasets with attrition indicators $W$ with a lag value 1000. I obtain the mean estimates for $Pr(Y_2 = 1)$ and for the conditional probabilities $Pr(Y_2 = 1|W)$ based on the $m = 50$ multiple completed datasets, where $W = 0/1$, estimated on only the completed panel data, only the refreshment sample and all data combining panel and refreshment sample, as shown in Table 3.8. Except the variable LV3 (.244 vs .356) with a bit difference (.112), the marginal probabilities $Pr(Y_2 = 1)$ are similar across panel and refreshment samples. The conditional probabilities $Pr(Y_2 = 1|W = 1)$, have close values across panel and refreshment samples. $Pr(Y_2 = 1|W = 0)$ are also similar between panel and refreshment samples except for LV3 (.118 vs .235). The joint probability $Pr(Y_2 = 1, W = 0)$ is similar for as shown in Table 3.9, which includes the joint probabilities $Pr(Y_2 = 1, W)$ for the panel, refreshment sample and both data sources.

I also collect the 95% confidence intervals for these marginal, conditional and joint probabilities using Rubin’s rule (Rubin, 1987) and compare them on panel data and refreshment sample, as illustrated in Figure 3.5. Ideally, if the refreshment sample is a random sample from the same population as the panel data, all the marginal, joint, and conditional probabilities should be the same as each other, as in the simulation study. This does not hold in the APYN data application. The probability $Pr(W = 1)$ only on the refreshment sample is .7984, and different from that in the panel (.6303). Hence, the marginal, joint, and conditional probabilities of $Y_2$ will be a bit different between the panel and refreshment samples.

If the conditional probability $Pr(Y_2 = 1 \mid W = 1)$ is different from $Pr(Y_2 = 1 \mid W = 0)$, $Y_2$ depends on $W$ and the attrition is non-ignorable. The outputs in Table 3.8 display the existence of non-ignorable attrition. I also collect the marginal probabilities and the 95% confidence intervals under MI for all the possible categories of the six $Y_2$ variables, for the stayed and dropout individuals in the panel, as in Figure 3.6. These marginal probabilities for the two groups are essentially conditional
Table 3.8: The marginal probability $Pr(Y_2 = 1)$ and the conditional probability $Pr(Y_2 = 1 \mid W)$ for all measures in wave 2 after imputation, estimated on only the completed panel data $p$, only the refreshment sample $r$ and all data $all$ for the APYN study in semi-parametric selection models.

<table>
<thead>
<tr>
<th></th>
<th>$Y_2$</th>
<th>$(Y_2 \mid W = 1)$</th>
<th>$(Y_2 \mid W = 0)$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$p$</td>
<td>$r$</td>
</tr>
<tr>
<td>FAV1-4</td>
<td>.311</td>
<td>.346</td>
<td>.289</td>
</tr>
<tr>
<td>FAV1-9</td>
<td>.184</td>
<td>.158</td>
<td>.181</td>
</tr>
<tr>
<td>LV1</td>
<td>.875</td>
<td>.892</td>
<td>.902</td>
</tr>
<tr>
<td>LV3</td>
<td>.244</td>
<td>.356</td>
<td>.318</td>
</tr>
<tr>
<td>CND1</td>
<td>.715</td>
<td>.726</td>
<td>.652</td>
</tr>
<tr>
<td>LV31</td>
<td>.119</td>
<td>.103</td>
<td>.102</td>
</tr>
</tbody>
</table>

Table 3.9: The joint probability $Pr(Y_2 = 1, W)$ for all measures in wave 2 after imputation estimated on only the completed panel data $p$, only the refreshment sample $r$ and all data $all$ for the APYN study in selection models.

<table>
<thead>
<tr>
<th></th>
<th>$(Y_2, W = 1)$</th>
<th>$(Y_2, W = 0)$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$p$</td>
<td>$r$</td>
</tr>
<tr>
<td>FAV1-4</td>
<td>.183</td>
<td>.243</td>
</tr>
<tr>
<td>FAV1-9</td>
<td>.115</td>
<td>.118</td>
</tr>
<tr>
<td>LV1</td>
<td>.569</td>
<td>.668</td>
</tr>
<tr>
<td>LV3</td>
<td>.200</td>
<td>.293</td>
</tr>
<tr>
<td>CND1</td>
<td>.411</td>
<td>.509</td>
</tr>
<tr>
<td>LV31</td>
<td>.064</td>
<td>.062</td>
</tr>
</tbody>
</table>

probabilities depending on the attrition group. We can see that the mean estimates of two probabilities are different from each other and some variables have relatively large value differences, such as LV31, CND1, LV3 and LV1. This shows that the distribution of $Y_2$ depends on the attrition $W$ and the attrition is non-ignorable.

I build eight regression models based on the outcome measures in wave 2 which are of interest to political scientists on the multiply collected datasets after imputation. First, I build three logistic regression models for the attrition indicator $W$: one includes LV31 (Indicate how you strongly feel about your likelihood to vote: certain will NOT vote 1 - Certain to vote 4) as the covariate, one includes LV3 (How much interest do you have in following news about the campaign for President, a great deal,
Figure 3.5: The marginal probability $Pr(Y_2 = 1)$, the conditional probability $Pr(Y_2 = 1|W)$ and the joint probability $Pr(Y_2 = 1, W)$ estimated from the completed panel and refreshment based on multiple $m = 50$ imputed datasets of selection models; left 6 cases are for $W = 1$ and right 6 cases are for $W = 0$. 
Figure 3.6: Comparison on marginal probability and 95% confidence intervals under MI for $Y_2$ between stayed and dropout samples in the completed panel data of selection models.

quite a bit, only some, and other: very little, or no interest at all?) as the covariate, and the last one includes CND1 (How much thoughts you give to the Presidential candidates: a lot vs others) as the covariate.

Second, I build a logistic regression with LV1 (are you registered to vote?) as the outcome and several demographic variables PPAGECT4 (age 18-29, 30-44, 45-59, and 60+), PPEDUCAT (less than high school, high school, some college, and bachelor’s degree or higher), PPGENDER (male and female), PPETHM (white and others), PPINCIMP (income from low to high) and PPMARIT (married and others) as the covariates. The categorical variables, denoted as $X_1=(age, edu, gender, race, income, marriage status)$, are included in regression models as covariates after dummy coding.
I also estimate two multinomial logistic regression models respectively treating LV3 (in the MI analysis, recoded as 3-level variables: a great deal, quite a bit, and others, for the consideration of random zeros), and LV31 (in the MI analysis, recoded as 3-level variables: certain will NOT vote 1 to certain will vote 3, for the consideration of random zeros) as outcomes with $X_1$ as predictors. I build another logistic regression with CND1 as the outcome variable and $X_1$ as predictor variables. In the MI analysis, I merge some categories of these variables to avoid the random zero fitting problems on refreshment samples only for inference comparison in the multinomial logistic regression models. This is done based on the frequency distribution check of the cross tabulated contingency tables.

Finally, I build a multinomial logistic regression with the response variable FAV1-4 to measure interviewers’ preference towards the candidate "Barack Obama", which includes 4 different categories: "very favorable", "somewhat favorable", "somewhat unfavorable" and "very unfavorable". I accumulate the last two categories "somewhat unfavorable" and "very unfavorable" as one level to avoid the random zero fitting problems in the MI analysis. The predictors I select include all the demographic variables in Table 3.5, denoted as $X_2=$(age, edu, gender, race, income, marriage status, partyID, ID1, REL3).

These eight regression models are used to evaluate and analyze the multiply completed datasets. First I compare the analysis between the completed panel and refreshment sample. Figure 3.7 displays the mean estimates and the 95% confidence intervals of the regression coefficients from the 8 models under MI on the completed panel data and refreshment sample. We can see that the mean estimates are close and the confidence intervals overlap with each other for the regression coefficients, except for several intercepts. One possible reason for the differences on intercepts is that the reference level has extremely small marginal probabilities. The analysis of these regression models is almost the same between the panel and refreshment
samples. We can conclude that the refreshment samples are representative of the panel data. The wider confidence intervals are due to the uncertainty induced by imputation.

Next I present estimates of the eight regression models combining the completed panel and refreshment samples. For the coefficients $\beta$, the mean estimates, variance estimates obtained by Rubin’s rule (Rubin, 1987), and 95% confidence intervals across the $m = 50$ completed datasets after MI are summarized in Table 3.10-3.18. Based on the estimates, we can infer the relationship between the predictors and the outcome variables with 95% confidence level.

Table 3.10 displays that participants with more likelihood to vote tend to stay in the panel comparing to those with less likelihood. Table 3.11 shows that participants with more interest following news on campaign tend to stay in the panel longer than those with less interest, which is also consistent with the output related to LV3 of wave 2 in Table 3.7. Table 3.12 shows that participants with a lot of thoughts towards Presidential candidates tend to stay in the panel longer than those with few thoughts. The analysis makes sense in practice. However, the conclusion is just the opposite of that from the output on CND1 from Table 3.7. We can see that strong co-linearity exists among the outcome measures as predictors in the attrition model. The estimation of the coefficient of CND1 in the large attrition model is affected by other predictors, such as LV3 and LV31, which are strongly correlated to CND1.

Table 3.13 displays that older and female participants with higher education level have more tendency to be registered for voting than the younger and male participants with lower education level. White participants tend to register comparing to those from other races. Participants with annual income with $75,000 or more tend to be registered comparing to those with lower income of less than $29,999.

Table 3.14 displays that older participants with higher education level tend to have more interest following news on campaign comparing to the younger with lower
Figure 3.7: Comparison of regression coefficients and the 95% confidence intervals under MI in the eight analysis models between completed panel data and refreshment samples for selection models in APYN study. From top left to bottom right, the models in sequence are $W \sim LV31$, $W \sim LV3$, $LV1 \sim X_1$, $LV3 \sim X_1$, $LV31 \sim X_1$, $CND1 \sim X_1$, $W \sim CND1$ and $FAV1-4 \sim X_2$. 
Table 3.10: Estimates and 95% confidence intervals for regression coefficients with W as outcome and LV31 as covariate on all completed data: certain will NOT vote 1 - Certain to vote 4 (reference) for LV31, in selection models.

<table>
<thead>
<tr>
<th>$\beta$</th>
<th>mean</th>
<th>variance</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>.795</td>
<td>.008</td>
<td>(.622 .969)</td>
</tr>
<tr>
<td>LV31-1</td>
<td>-.580</td>
<td>.078</td>
<td>(-1.136 -.024)</td>
</tr>
<tr>
<td>LV31-2</td>
<td>-.690</td>
<td>.110</td>
<td>(-1.351 -.029)</td>
</tr>
<tr>
<td>LV31-3</td>
<td>-.509</td>
<td>.081</td>
<td>(-1.074 .057)</td>
</tr>
</tbody>
</table>

education. Participants with high income level ($75,000 or more) have more interest following news on campaign than those with income less than $50,000.

Table 3.15 shows that younger participants with lower education level have less likelihood to vote than the older with higher education level. Male have less likelihood to vote than female. Participants with income less than $29,999 have less likelihood to vote than those with a high income level with $75,000 or more.

Table 3.16 shows that older and single participants with higher education level and higher income tend to have given more thoughts to Presidential candidates than the younger with other marriage status, lower education level and lower income.

Table 3.17 shows that participants with a Bachelor’s degree or higher education level have more favorable impression on Barack Obama than those with some college degree, high school, or less than high school; White participants tend to be unfavorable to Barack Obama than those from other races. Table 3.18 displays that Demographic party has more supports for Obama, while Republicans have more unfavorable impressions on him. Liberal participants have more favorable impression on Obama than the conservative.

3.5.1 Discussions on Semi-parametric Models for the APYN Data

For the 21 collected outcome measures in Table 4.3 and 24 demographic variables in Table 4.4 from APYN study in Chapter 4, I tried different prior choices for the regression coefficients $\beta$ in the attrition model. I first use very flat prior distributions
Table 3.11: Estimates and 95% confidence intervals for regression coefficients with W as outcome and LV3 as covariate on all completed data: level 1-a great deal, level 2- quite a bit, level 3-only some, and level 4- very little or none (reference) for LV3 in selection models.

<table>
<thead>
<tr>
<th>β</th>
<th>mean</th>
<th>variance</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-.311</td>
<td>.028</td>
<td>(-.643 .021)</td>
</tr>
<tr>
<td>LV3-1</td>
<td>1.888</td>
<td>.156</td>
<td>( 1.098 2.677)</td>
</tr>
<tr>
<td>LV3-2</td>
<td>1.736</td>
<td>.177</td>
<td>(.894 2.578)</td>
</tr>
<tr>
<td>LV3-3</td>
<td>.093</td>
<td>.055</td>
<td>(-.375 .560)</td>
</tr>
</tbody>
</table>

Table 3.12: Estimates and 95% confidence intervals for regression coefficients with W as outcome and CND1 as covariate on all completed data: level 1-a lot, level 2 (reference)-fewer for CND1 in selection models.

<table>
<thead>
<tr>
<th>β</th>
<th>mean</th>
<th>variance</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>.379</td>
<td>.005</td>
<td>(.244 .514)</td>
</tr>
<tr>
<td>CND1</td>
<td>.898</td>
<td>.106</td>
<td>(.247 1.549)</td>
</tr>
</tbody>
</table>

identically and independently on the β, such as \(N(0,1000)\). The MCMC chain fails to converge and the estimates of the coefficients related to variables LV3, LV31 and CND1 are stuck in implausible regions (absolute values are large and close to 20). I use \(N(0,25)\) as prior distribution and implement the adaptive rejection sampling with boundary \((-5,5)\) for the posterior computation of β. The estimates of the coefficients that were blown up get stuck around the boundary during the updating. For data augmentation, the maximum likelihood estimates of β do not exist based on some imputed datasets. I cannot use the maximum likelihood estimates for the proposal distribution in the Metropolis step. Other choices for the proposal are hard to make because of extreme small/large acceptance rates. Weakly informative prior distributions cannot solve the mixing problems and also yield too large estimate values for those several coefficients. This illustrates that there are serious complete or quasi-complete separation problems among the variables.

I implement the Bayesian Lasso prior to shrink the estimates of the coefficients. The variables are believed to be strongly correlated with each other. The mixing
Table 3.13: Estimates and 95% confidence intervals for regression coefficients with LV1 as outcome and $X_1$ as covariate on all completed data: level 1-yes, level 2-no/other (reference) for LV1 in selection models.

<table>
<thead>
<tr>
<th>$\beta$</th>
<th>mean</th>
<th>variance</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>3.996</td>
<td>.162</td>
<td>(3.203 4.789)</td>
</tr>
<tr>
<td>AGE (18-29)</td>
<td>-1.263</td>
<td>.061</td>
<td>(-1.747 -.780)</td>
</tr>
<tr>
<td>AGE (30-44)</td>
<td>-1.355</td>
<td>.050</td>
<td>(-1.793 -.917)</td>
</tr>
<tr>
<td>AGE (45-59)</td>
<td>-1.050</td>
<td>.046</td>
<td>(-1.470 -.631)</td>
</tr>
<tr>
<td>EDU (less than high school)</td>
<td>-1.984</td>
<td>.058</td>
<td>(-2.455 -1.513)</td>
</tr>
<tr>
<td>EDU (high school)</td>
<td>-1.616</td>
<td>.041</td>
<td>(-2.012 -1.220)</td>
</tr>
<tr>
<td>EDU (some college)</td>
<td>-0.808</td>
<td>.040</td>
<td>(-1.201 -.416)</td>
</tr>
<tr>
<td>RACE (white &amp; non-Hispanic)</td>
<td>0.367</td>
<td>.026</td>
<td>(.048 .687)</td>
</tr>
<tr>
<td>GENDER (male)</td>
<td>-0.224</td>
<td>.018</td>
<td>(-.485 .038)</td>
</tr>
<tr>
<td>INC (less than $29,999)</td>
<td>-0.579</td>
<td>.038</td>
<td>(-.959 -.200)</td>
</tr>
<tr>
<td>INC ($30,000 to $49,999)</td>
<td>-0.237</td>
<td>.036</td>
<td>(-.608 .133)</td>
</tr>
<tr>
<td>INC ($50,000 to $74,999)</td>
<td>0.060</td>
<td>.042</td>
<td>(-.341 .460)</td>
</tr>
<tr>
<td>MARIT (married)</td>
<td>0.119</td>
<td>.020</td>
<td>(-.158 .400)</td>
</tr>
</tbody>
</table>

behavior is improved and achieves convergence very quickly. All the regression coefficients are shrunk towards to 0, shown in Figure 3.8, while the values corresponding the variables with extremely large coefficient estimates in the model with diffuse priors are still relatively larger than the remaining values. The shrinkage on $\beta$ reduces the affects from the variables on the attrition. I need to determine a balance between the shrinkage effect and the imputation performance. Too strong shrinkage control will violates Rubin’s rule for a proper multiple imputation procedure. But we need informative prior distributions to facilitate the MCMC mixing performance. To balance these two concerns, I choose $N(0, 1)$ identically and independently for $\beta$ in the analysis presented in Section 3.5.

Because of the quasi-complete separation problems, the selection models cannot hold if these massive variables are included as predictors at the same time. This could be another reason for the posterior computation that becomes sensitive to the prior choices. I reduce the set of the variables as in Table 3.4 and Table 3.5 in this Chapter, and include a relatively small number of binary covariates (51) after
Table 3.14: Estimates and 95% confidence intervals for regression coefficients with LV3 as outcome and $X_1$ as covariate on all completed data: C1-a great deal (level 1), C2-quite a bit (level 2), C3-only some (level 3), and level 4- little or none (reference) for LV3 in selection models.

<table>
<thead>
<tr>
<th></th>
<th>C1</th>
<th></th>
<th>C2</th>
<th></th>
<th>C3</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean</td>
<td>var</td>
<td>mean</td>
<td>var</td>
<td>mean</td>
<td>var</td>
</tr>
<tr>
<td>Intercept</td>
<td>2.924</td>
<td>.137</td>
<td>2.407</td>
<td>.147</td>
<td>1.512</td>
<td>.161</td>
</tr>
<tr>
<td></td>
<td>(2.196, 3.652)</td>
<td>(1.651, 3.164)</td>
<td>(.718, 2.305)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AGE (18-29)</td>
<td>-1.617</td>
<td>.061</td>
<td>-.855</td>
<td>.063</td>
<td>-.223</td>
<td>.068</td>
</tr>
<tr>
<td></td>
<td>(-2.101, -1.134)</td>
<td>(-1.351, -1.360)</td>
<td>(-.735, .288)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AGE (30-44)</td>
<td>-1.525</td>
<td>.046</td>
<td>-.962</td>
<td>.049</td>
<td>-.410</td>
<td>.052</td>
</tr>
<tr>
<td></td>
<td>(-1.945, -1.104)</td>
<td>(-1.400, -.525)</td>
<td>(-.857, .037)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AGE (45-59)</td>
<td>-1.123</td>
<td>.040</td>
<td>-.480</td>
<td>.042</td>
<td>-.299</td>
<td>.041</td>
</tr>
<tr>
<td></td>
<td>(-1.514, -.732)</td>
<td>(-.880, -.079)</td>
<td>(-.696, .097)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EDU (&lt;high school)</td>
<td>-1.900</td>
<td>.077</td>
<td>-1.745</td>
<td>.074</td>
<td>-1.091</td>
<td>.079</td>
</tr>
<tr>
<td></td>
<td>(-2.445, -1.355)</td>
<td>(-2.281, -1.209)</td>
<td>(-1.646, -.536)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EDU (high school)</td>
<td>-1.607</td>
<td>.048</td>
<td>-1.319</td>
<td>.044</td>
<td>-.792</td>
<td>.047</td>
</tr>
<tr>
<td></td>
<td>(-2.037, -1.177)</td>
<td>(-1.730, -.908)</td>
<td>(-1.217, -.368)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EDU (some college)</td>
<td>-.842</td>
<td>.042</td>
<td>-.693</td>
<td>.041</td>
<td>-.410</td>
<td>.044</td>
</tr>
<tr>
<td></td>
<td>(-1.244, -.440)</td>
<td>(-1.088, -.298)</td>
<td>(-.823, .003)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RACE (white&amp;non-His)</td>
<td>-.032</td>
<td>.036</td>
<td>.017</td>
<td>.032</td>
<td>.105</td>
<td>.037</td>
</tr>
<tr>
<td></td>
<td>(-.403, .339)</td>
<td>(-.336, .370)</td>
<td>(.276, .487)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GENDER (male)</td>
<td>.046</td>
<td>.021</td>
<td>.048</td>
<td>.020</td>
<td>.005</td>
<td>.018</td>
</tr>
<tr>
<td></td>
<td>(-.241, .333)</td>
<td>(-.227, .323)</td>
<td>(.261, .271)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INC ($\leq$ 29,999)</td>
<td>-.770</td>
<td>.044</td>
<td>-.712</td>
<td>.047</td>
<td>-.377</td>
<td>.049</td>
</tr>
<tr>
<td></td>
<td>(-1.181, -.358)</td>
<td>(-1.139, -.286)</td>
<td>(-.810, .057)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INC ($30,000-$49,999)</td>
<td>-.602</td>
<td>.042</td>
<td>-.316</td>
<td>.039</td>
<td>-.134</td>
<td>.046</td>
</tr>
<tr>
<td></td>
<td>(-1.005, -.198)</td>
<td>(-.703, .071)</td>
<td>(-.558, .290)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INC ($50,000-$74,999)</td>
<td>-.097</td>
<td>.050</td>
<td>-.166</td>
<td>.047</td>
<td>-.006</td>
<td>.053</td>
</tr>
<tr>
<td></td>
<td>(-.536, .342)</td>
<td>(-.591, .258)</td>
<td>(-.457, .445)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MARIT (married)</td>
<td>.024</td>
<td>.024</td>
<td>.016</td>
<td>.023</td>
<td>.085</td>
<td>.026</td>
</tr>
<tr>
<td></td>
<td>(-.281, .328)</td>
<td>(-.283, .314)</td>
<td>(-.233, .403)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

dummy coding in to the selection model. The selection models perform well now.

Another important concern is that I only include main effects of these variables into the selection model. It is possible that the high-way interaction terms between these variables (no interaction between $Y_1$ and $Y_2$ allowed for identification) are also significant predictors for the attrition. The main-effect selection model may ignore the complex dependency structure. However, the computation and the involvement
Table 3.15: Estimates and 95% confidence intervals for regression coefficients with LV31 as outcome and $X_1$ as covariate on all completed data: C1-Level "certain will NOT vote" 1, C2-Level "Certain to vote" 2, C3-Level "Certain to vote" 3, Reference is Level "Certain to vote" 4 for LV31 in selection models.

<table>
<thead>
<tr>
<th></th>
<th>C1</th>
<th></th>
<th>C2</th>
<th></th>
<th>C3</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean</td>
<td>var</td>
<td>mean</td>
<td>var</td>
<td>mean</td>
<td>var</td>
</tr>
<tr>
<td>Intercept</td>
<td>-4.315</td>
<td>.134</td>
<td>-3.790</td>
<td>.154</td>
<td>-3.142</td>
<td>.155</td>
</tr>
<tr>
<td></td>
<td>(-5.033, -3.596)</td>
<td>(-4.563, -3.018)</td>
<td>(-3.919, -2.366)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AGE (18-29)</td>
<td>1.823</td>
<td>.075</td>
<td>1.440</td>
<td>.081</td>
<td>1.049</td>
<td>.090</td>
</tr>
<tr>
<td></td>
<td>(1.287, 2.359)</td>
<td>(.880, 1.999)</td>
<td>(.459, 1.640)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AGE (30-44)</td>
<td>1.870</td>
<td>.053</td>
<td>1.449</td>
<td>.059</td>
<td>.707</td>
<td>.067</td>
</tr>
<tr>
<td></td>
<td>(1.418, 2.321)</td>
<td>(.972, 1.927)</td>
<td>(.196, 1.218)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AGE (45-59)</td>
<td>1.361</td>
<td>.056</td>
<td>.826</td>
<td>.058</td>
<td>.507</td>
<td>.073</td>
</tr>
<tr>
<td></td>
<td>(.8956, 1.826)</td>
<td>(.354, 1.298)</td>
<td>(.024, 1.039)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EDU (&lt;high school)</td>
<td>2.068</td>
<td>.072</td>
<td>1.729</td>
<td>.078</td>
<td>.423</td>
<td>.146</td>
</tr>
<tr>
<td></td>
<td>(1.540, 2.596)</td>
<td>(1.179, 2.278)</td>
<td>(-.331, 1.176)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EDU (high school)</td>
<td>1.724</td>
<td>.048</td>
<td>1.063</td>
<td>.058</td>
<td>.116</td>
<td>.051</td>
</tr>
<tr>
<td></td>
<td>(1.296, 2.152)</td>
<td>(.590, 1.535)</td>
<td>(-.327, .560)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EDU (some college)</td>
<td>.846</td>
<td>.044</td>
<td>.603</td>
<td>.052</td>
<td>.149</td>
<td>.041</td>
</tr>
<tr>
<td></td>
<td>(.433, 1.258)</td>
<td>(.155, 1.050)</td>
<td>(-.248, .546)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RACE (white&amp;non-His)</td>
<td>-.187</td>
<td>.024</td>
<td>.033</td>
<td>.034</td>
<td>.136</td>
<td>.043</td>
</tr>
<tr>
<td></td>
<td>(-.492, .118)</td>
<td>(-.331, .396)</td>
<td>(-.272, .544)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GENDER (male)</td>
<td>.178</td>
<td>.018</td>
<td>.056</td>
<td>.024</td>
<td>.321</td>
<td>.025</td>
</tr>
<tr>
<td></td>
<td>(.087, .443)</td>
<td>(-.250, .362)</td>
<td>(.011, .631)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INC ($\leq 29,999)</td>
<td>.547</td>
<td>.040</td>
<td>.433</td>
<td>.055</td>
<td>.235</td>
<td>.064</td>
</tr>
<tr>
<td></td>
<td>(.155, .939)</td>
<td>(-.028, .894)</td>
<td>(-.263, .734)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INC ($30,000-$49,999)</td>
<td>.236</td>
<td>.040</td>
<td>.291</td>
<td>.052</td>
<td>.354</td>
<td>.052</td>
</tr>
<tr>
<td></td>
<td>(-.157, .628)</td>
<td>(-.157, .739)</td>
<td>(-.093, .801)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INC ($50,000-$74,999)</td>
<td>.001</td>
<td>.043</td>
<td>-.032</td>
<td>.061</td>
<td>.050</td>
<td>.069</td>
</tr>
<tr>
<td></td>
<td>(-.405, .406)</td>
<td>(-.519, .454)</td>
<td>(-.467, .566)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MARIT (married)</td>
<td>-.089</td>
<td>.020</td>
<td>-.072</td>
<td>.027</td>
<td>-.137</td>
<td>.027</td>
</tr>
<tr>
<td></td>
<td>(.366, .188)</td>
<td>(-.397, .253)</td>
<td>(-.459, .185)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

of high-way interaction terms requires a trade off, especially under high dimension when we apply dummy coding for the categorical variables into the regression model.

3.6 Conclusions

I impute missing values due to potentially non-ignorable attrition and random item nonresponse simultaneously from a selection model perspective. I focus on a two
Table 3.16: Estimates and 95% confidence intervals for regression coefficients with CND1 as outcome and $X_1$ as covariate on all completed data in selection models.

<table>
<thead>
<tr>
<th>$\beta$</th>
<th>mean</th>
<th>variance</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>2.599</td>
<td>.042</td>
<td>(2.198 3.000)</td>
</tr>
<tr>
<td>AGE (18-29)</td>
<td>-1.416</td>
<td>.026</td>
<td>(-1.734 -1.099)</td>
</tr>
<tr>
<td>AGE (30-44)</td>
<td>-1.291</td>
<td>.021</td>
<td>(-1.572 -1.010)</td>
</tr>
<tr>
<td>AGE (45-59)</td>
<td>-0.897</td>
<td>.017</td>
<td>(-1.156 -0.639)</td>
</tr>
<tr>
<td>EDU (less than high school)</td>
<td>-1.091</td>
<td>.033</td>
<td>(-1.450 -0.732)</td>
</tr>
<tr>
<td>EDU (high school)</td>
<td>-0.897</td>
<td>.016</td>
<td>(-1.147 -0.648)</td>
</tr>
<tr>
<td>EDU (some college)</td>
<td>-0.293</td>
<td>.015</td>
<td>(-0.533 -0.053)</td>
</tr>
<tr>
<td>RACE (white &amp; non-Hispanic)</td>
<td>0.049</td>
<td>.012</td>
<td>(-0.169 0.268)</td>
</tr>
<tr>
<td>GENDER (male)</td>
<td>-0.044</td>
<td>.008</td>
<td>(-0.224 0.137)</td>
</tr>
<tr>
<td>INC (less than $29,999)</td>
<td>-0.389</td>
<td>.018</td>
<td>(-0.650 -0.129)</td>
</tr>
<tr>
<td>INC ($30,000 to $49,999)</td>
<td>-0.306</td>
<td>.016</td>
<td>(-0.557 -0.056)</td>
</tr>
<tr>
<td>INC ($50,000 to $74,999)</td>
<td>-0.217</td>
<td>.018</td>
<td>(-0.483 0.048)</td>
</tr>
<tr>
<td>MARIT (married)</td>
<td>-0.139</td>
<td>.010</td>
<td>(-0.332 0.054)</td>
</tr>
</tbody>
</table>

wave panel study with one refreshment sample available in the second wave, and each wave includes high dimensional variables. I propose a procedure featuring a semi-parametric selection model: a parametric model for missingness mechanism and a nonparametric model for completed data. This procedure can generate plausible imputation values for the collected variables in the APYN study. The potential separation problems of the data make the selection models fragile. The computation algorithm to update the regression coefficients in the selection model is discussed in detail. To avoid sticky updating for massive number of regression coefficients if more variables and high-way interaction terms are included, I propose Bayesian latent pattern mixture models that provide solutions in Chapter 4.
Table 3.17: Estimates and 95% confidence intervals for regression coefficients with FAV1-4 as outcome and $X_2$ as covariate on all completed data: C1-very favorable (level 1), C2-somewhat favorable (level 2), C3-somewhat unfavorable (level 3), and reference is very unfavorable for FAV1-4 in selection models.

<table>
<thead>
<tr>
<th>$\beta$</th>
<th>C1</th>
<th></th>
<th>C2</th>
<th></th>
<th>C3</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-.028</td>
<td>.269</td>
<td>-.038</td>
<td>.166</td>
<td>-.696</td>
<td>.203</td>
</tr>
<tr>
<td></td>
<td>(-1.048, .993)</td>
<td>(.837, .761)</td>
<td>(-1.580, .189)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AGE (18-29)</td>
<td>-.218</td>
<td>.063</td>
<td>.279</td>
<td>.057</td>
<td>-.202</td>
<td>.057</td>
</tr>
<tr>
<td></td>
<td>(-.711, .276)</td>
<td>(.188, .747)</td>
<td>(.671, .267)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AGE (30-44)</td>
<td>-.182</td>
<td>.043</td>
<td>.120</td>
<td>.038</td>
<td>.009</td>
<td>.034</td>
</tr>
<tr>
<td></td>
<td>(-.591, .227)</td>
<td>(.265, .505)</td>
<td>(.353, .370)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AGE (45-59)</td>
<td>-.136</td>
<td>.036</td>
<td>.144</td>
<td>.030</td>
<td>-.107</td>
<td>.034</td>
</tr>
<tr>
<td></td>
<td>(-.510, .237)</td>
<td>(.195, .483)</td>
<td>(.471, .256)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EDU (&lt;high school)</td>
<td>-.531</td>
<td>.083</td>
<td>-.738</td>
<td>.086</td>
<td>-.310</td>
<td>.075</td>
</tr>
<tr>
<td></td>
<td>(-1.097, .035)</td>
<td>(-1.314, -.162)</td>
<td>(-.848, .228)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EDU (high school)</td>
<td>-.616</td>
<td>.040</td>
<td>-.524</td>
<td>.035</td>
<td>-.281</td>
<td>.037</td>
</tr>
<tr>
<td></td>
<td>(-1.010, -.222)</td>
<td>(-.889, -.159)</td>
<td>(-.657, .094)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EDU (some college)</td>
<td>-.463</td>
<td>.029</td>
<td>-.437</td>
<td>.026</td>
<td>-.460</td>
<td>.028</td>
</tr>
<tr>
<td></td>
<td>(-.797, -.128)</td>
<td>(-.755, -.119)</td>
<td>(-.785, -.135)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RACE (white&amp;non-His)</td>
<td>-.1104</td>
<td>.037</td>
<td>-.426</td>
<td>.033</td>
<td>-.177</td>
<td>.043</td>
</tr>
<tr>
<td></td>
<td>(-1.481, -.727)</td>
<td>(-.780, -.072)</td>
<td>(-.584, .231)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GENDER (male)</td>
<td>-.049</td>
<td>.020</td>
<td>-.029</td>
<td>.017</td>
<td>.047</td>
<td>.019</td>
</tr>
<tr>
<td></td>
<td>(-.324, .227)</td>
<td>(-.285, .227)</td>
<td>(-.221, .314)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INC (≤ $29,999)</td>
<td>-.301</td>
<td>.051</td>
<td>-.292</td>
<td>.044</td>
<td>-.115</td>
<td>.042</td>
</tr>
<tr>
<td></td>
<td>(-.743, .142)</td>
<td>(-.705, .121)</td>
<td>(-.515, .286)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INC ($30,000-$49,999)</td>
<td>-.266</td>
<td>.043</td>
<td>-.231</td>
<td>.035</td>
<td>-.082</td>
<td>.034</td>
</tr>
<tr>
<td></td>
<td>(-.674, .143)</td>
<td>(-.600, .138)</td>
<td>(-.444, .279)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INC ($50,000-$74,999)</td>
<td>.031</td>
<td>.042</td>
<td>-.030</td>
<td>.037</td>
<td>-.180</td>
<td>.040</td>
</tr>
<tr>
<td></td>
<td>(-.370, .432)</td>
<td>(-.408, .348)</td>
<td>(-.573, .213)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MARIT (married)</td>
<td>-.277</td>
<td>.025</td>
<td>.082</td>
<td>.022</td>
<td>.133</td>
<td>.022</td>
</tr>
<tr>
<td></td>
<td>(-.585, .030)</td>
<td>(-.209, .373)</td>
<td>(-.161, .427)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 3.18: Estimates and 95% confidence intervals for regression coefficients with FAV1-4 as outcome and $X_2$ as covariate on all completed data: C1-very favorable (level 1), C2-somewhat favorable (level 2), C3-somewhat unfavorable (level 3), and reference is very unfavorable for FAV1-4 in selection models CONT. Note: w-week, m-month, y-year.

<table>
<thead>
<tr>
<th>$\beta$</th>
<th>C1</th>
<th></th>
<th>C2</th>
<th></th>
<th>C3</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean</td>
<td>var</td>
<td>mean</td>
<td>var</td>
<td>mean</td>
<td>var</td>
</tr>
<tr>
<td>PARTYID(Democrat)</td>
<td>1.482</td>
<td>.053</td>
<td>.626</td>
<td>.042</td>
<td>.211</td>
<td>.062</td>
</tr>
<tr>
<td></td>
<td>(1.032, 1.933)</td>
<td>(.222, 1.029)</td>
<td>(-.280, .701)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PARTYID(Republican)</td>
<td>-1.303</td>
<td>.055</td>
<td>-1.277</td>
<td>.041</td>
<td>-1.325</td>
<td>.048</td>
</tr>
<tr>
<td></td>
<td>(-1.763, -.843)</td>
<td>(-1.273, -.481)</td>
<td>(-.755, .106)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ID1(very liberal)</td>
<td>3.417</td>
<td>.243</td>
<td>1.946</td>
<td>.205</td>
<td>.683</td>
<td>.430</td>
</tr>
<tr>
<td></td>
<td>(2.449, 4.386)</td>
<td>(1.056, 2.835)</td>
<td>(-.606, 1.973)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ID1(somewhat liberal)</td>
<td>2.897</td>
<td>.157</td>
<td>1.962</td>
<td>.096</td>
<td>1.082</td>
<td>.116</td>
</tr>
<tr>
<td></td>
<td>(2.119, 3.676)</td>
<td>(1.355, 2.570)</td>
<td>(.413, 1.751)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ID1(moderate)</td>
<td>2.149</td>
<td>.105</td>
<td>1.697</td>
<td>.050</td>
<td>1.411</td>
<td>.049</td>
</tr>
<tr>
<td></td>
<td>(1.511, 2.788)</td>
<td>(1.258, 2.137)</td>
<td>(.978, 1.844)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ID1(very conservative)</td>
<td>1.271</td>
<td>.109</td>
<td>.851</td>
<td>.049</td>
<td>1.071</td>
<td>.040</td>
</tr>
<tr>
<td></td>
<td>(.621, 1.921)</td>
<td>(.417, 1.285)</td>
<td>(.678, 1.465)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>REL3(more than once/w)</td>
<td>0.35</td>
<td>.067</td>
<td>-.499</td>
<td>.061</td>
<td>-.060</td>
<td>.061</td>
</tr>
<tr>
<td></td>
<td>(-.473, .543)</td>
<td>(-.984, -.015)</td>
<td>(-.545, .425)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>REL3 (once/w)</td>
<td>-.152</td>
<td>.043</td>
<td>-2.283</td>
<td>.038</td>
<td>.039</td>
<td>.042</td>
</tr>
<tr>
<td></td>
<td>(-.560, .255)</td>
<td>(-.665, .100)</td>
<td>(-.365, .443)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>REL3 (a few times/m)</td>
<td>-.081</td>
<td>.078</td>
<td>-.177</td>
<td>.066</td>
<td>.115</td>
<td>.074</td>
</tr>
<tr>
<td></td>
<td>(-.629, .468)</td>
<td>(-.681, .327)</td>
<td>(.419, .648)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>REL3 (a few times/y)</td>
<td>-.149</td>
<td>.038</td>
<td>-.031</td>
<td>.032</td>
<td>-.019</td>
<td>.039</td>
</tr>
<tr>
<td></td>
<td>(-.531, .232)</td>
<td>(-.381, .320)</td>
<td>(-.408, .370)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

89
Figure 3.8: Posterior mean estimates for $\beta$ under Bayesian Lasso in selection models for APYN study.
4.1 Introduction

Existing model-based methods with refreshment samples are based on selection models for attrition process, for example, Hirano et al. (2001), Deng et al. (2012) and Chapter 3 in this thesis. Kenward (1998) showed that in cross-sectional data estimates could be sensitive to the misspecification of the marginal distribution in the model for missingness. The additive non-ignorable selection model for attrition in Chapter 3 only includes the main effects of the variables. If there exists complex dependency structure, the main-effect only additive non-ignorable model can result in inaccurate imputations. Such interactions are difficult to select in high dimensions. Additionally, even for the main-effect only additive non-ignorable model, posterior computation for the selection model has sticky mixing behavior and is subject to colinearity problems.

The difficulty of modeling interactions in the attrition model motivates my next topic: using pattern mixture models when refreshment samples are available. To my
knowledge, no one has developed pattern mixture models in the context of refreshment samples.

In this chapter, I modify the DPMPM distributions developed in Chapter 2 by including the attrition indicator into the construction for latent classes. This is realized through two main directions. First, I include the attrition indicator as a covariate in the weights for the DP and develop a dependent DP (DDP) process. Second, I estimate cluster membership jointly on the collected variables and attrition indicator.

The structure of this chapter is organized as follows. Section 4.2 introduces common strategies for pattern mixture models without refreshment samples and presents a simple model for use with refreshment samples in low dimensions. I also describe the data from the APYN study analyzed in this chapter. Section 4.3 reviews latent class models for longitudinal studies with dropout but without refreshment samples in the literature. Section 4.4 proposes Bayesian latent pattern mixture (BLPM) models from two perspectives: dependent BLPM and joint BLPM. Section 4.5 investigates the benefits of refreshment sample for the dependent BLPM and joint BLPM models under conditional independence (CI) assumption. Section 4.6 releases the CI assumption and studies the performance of the BLPM models. Section 4.7 summarizes the chapter.

4.2 Pattern Mixture Models and Data Description

4.2.1 Pattern Mixture Models without Refreshment Samples

Because there are no data to estimate \( f(Z|W=0) \) directly in pattern mixture models, analysts must adopt unverifiable assumptions to identify parameters. One approach is to make untestable assumptions on the missing data mechanism by setting the model deterministically. Hogan and Daniels (2008, Chapter 8) discussed common identification strategies including MAR constraints, interior family constraints, non-
future dependence restrictions, and extrapolation. Another solution is to relate the unidentified parameters to identified parameters using strongly informative prior distributions, as described in Hogan and Daniels (2008, Chapter 9).

Another kind of restriction is to choose pre-specified constant values and fix the relationship for the parameters of observed data and missing data patterns. For example, Little (1994) proposed a normal pattern mixture model for two repeated continuous and univariate measures with NMAR dropouts, such that

\[(Y_{i1}, Y_{i2}|W_i=k) \sim N(\mu^{(k)}, \Sigma^{(k)})\] (4.1)

\[W_i \sim \text{Bernoulli}(\delta),\] (4.2)

where \(k = \{0, 1\}\). This model is analyzed under the assumptions

\[\Pr(W_i=1|Y_{i1}, Y_{i2}) = g(Y_{i1}^*) = g(Y_{i1} + \lambda Y_{i2}).\] (4.3)

Here, \(Y_{i2}\) given \(Y_{i1}^* = Y_{i1} + \lambda Y_{i2}\) is independent of \(W_i\) and follows a normal distribution that is the same across the two patterns. The resulting constraints are just sufficient for identifications. Multiple values of \(\lambda\) are specified by the analyst to enable sensitivity analysis. This assumption is similar to the additive non-ignorable model (Hirano et al., 2001), but the latter estimates \(\lambda\) from the data, which becomes possible due to extra information in the refreshment sample.

Little and Wang (1996) extend to multivariate normal pattern mixture models for \(p\) repeated measures with NMAR dropouts and covariates, assuming that

\[(Y_i|X_i, W_i=k) \sim N(X_i\theta^{(k)}, \Sigma^{(k)})\] (4.4)

\[(W_i|X_i) \sim \text{Bernoulli}(\delta(X_i)),\] (4.5)

where \(k = \{0, 1\}\). Under the assumption for identifiability \(Pr(W=1|Y_1, Y_2, X) = f(CY_1 + \Lambda Y_2, X)\), the constants \(C\) and \(\Lambda\) are pre-determined to enable sensitivity analysis.

In the additive non-ignorable model with refreshment samples, \(C\) and \(\Lambda\) can be estimated from the model.
Table 4.1: A simple two wave panel data with two variables

<table>
<thead>
<tr>
<th>Subsample</th>
<th>$Y_1$</th>
<th>$Y_2$</th>
<th>$W$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complete panel</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Incomplete panel</td>
<td>0</td>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td>Refreshment</td>
<td>-</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>1</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 4.2: Directly estimated terms (√) for the simple example

<table>
<thead>
<tr>
<th>$Y_2=1&amp;Y_1=1$</th>
<th>$Y_2=0&amp;Y_1=1$</th>
<th>$Y_2=1&amp;Y_1=0$</th>
<th>$Y_2=0&amp;Y_1=0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$W=1$</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>$W=0$</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
</tbody>
</table>

4.2.2 Pattern Mixture Models in Two Wave Panel with One Refreshment Sample

I start from the simple example with two binary variables (ignore X here) from Hirano et al. (2001), shown in Table 4.1, to illustrate how the identification issues of pattern mixture models are solved when the refreshment sample is available. Denote $Pr(Y_2=1|Y_1=y_1, W=w)$ by $g_{y_1w}$ and $Pr(Y_1=y_1, W=w)$ by $r_{y_1w}$.

Three constraints for unidentified terms are obtained from the Table 4.2 as follows

$$1 = \sum_{ij} Pr(Y_2=i, Y_1=j|W=0) \quad (4.6)$$

$$Pr(Y_1=1|W=0) = Pr(Y_2=1, Y_1=1|W=0) + Pr(Y_2=0, Y_1=1|W=0) \quad (4.7)$$

$$Pr(Y_2=1|W=0) = Pr(Y_2=1, Y_1=1|W=0) + Pr(Y_2=1, Y_1=0|W=0). \quad (4.8)$$

Here $Pr(Y_2=1|W=0)$ can be calculated from the refreshment sample

$$Pr(Y_2=1) = Pr(Y_2=1|W=0)Pr(W=0) + Pr(Y_2=1|W=1)Pr(W=1). \quad (4.9)$$

Equivalently, we have the following facts: since $Y_1$ and $W$ are fully observed, we can directly estimate $(r_{00}, r_{01}, r_{10}, r_{11})$; when $W=1$, we observe $Y_2$ no matter
\( Y_1 = 0 \text{ or } 1 \), and then we can directly estimate \( q_{01} \) and \( q_{11} \): \( Pr(Y_2 = 1) \) is estimable from the refreshment samples, so we have \( Pr(Y_2 = 1) = \sum_{ij} q_{ij} r_{ij} \) as fixed. In sum, we need to estimate two unknown terms \( q_{00} \) and \( q_{10} \) from one constraint \( q_{00} r_{00} + q_{10} r_{10} + q_{01} r_{01} + q_{11} r_{11} = Pr(Y_2 = 1) \). We need another extra constraint. The MAR assumption \( W \perp Y_2 | Y_1 \) provides two extra constraints: \( q_{00} = q_{01} \) and \( q_{10} = q_{11} \). The NMAR assumption \( W \perp Y_1 | Y_2 \) specified in Hausman and Wise (1979) provides two constraints: \( q_{10} r_{10} q_{01} r_{01} = q_{11} r_{11} q_{00} r_{00} \) and \( (1 - q_{10}) r_{10} (1 - q_{01}) r_{01} = (1 - q_{11}) r_{11} (1 - q_{00}) r_{00} \).

Inducing just one extra constraint, I propose a pattern mixture model that assumes sequential models for the binary variables.

\[
W \sim \text{Bern}(\pi)
Pr(Y_1 = 1 | W) = f(\beta_0^* + \beta_1^* W)
Pr(Y_2 = 1 | Y_1, W) = f(\beta_0^{**} + \beta_1^{**} W + \beta_2^{**} Y_1). \tag{4.10}
\]

Suppose \( f(\cdot) \) is any continuous and increasing function with \( \lim_{a \to -\infty} f(a) = 0 \) and \( \lim_{a \to \infty} f(a) = 1 \), for example, logit and probit functions. The model (4.10) embeds the special cases: if MAR, \( \beta_1^* \neq 0 \) and \( \beta_1^{**} = 0 \); if NMAR, \( \beta_1^{**} \neq 0 \). I prove that model (4.10) introduces one constraint and can be just identified under THEOREM 1.

THEOREM 1: For any quadruple \( q_{01}, q_{11} \in (0, 1) \), any quadruple \( r_{y_1 y_2} \in (0, 1) \) with \( \sum_{y_1 y_2} r_{y_1 y_2} = 1 \), and any continuous and increasing function \( f(\cdot) \) with \( \lim_{a \to -\infty} f(a) = 0 \) and \( \lim_{a \to \infty} f(a) = 1 \), there is a unique quintuple \( (\beta_0^{**}, \beta_1^{**}, \beta_2^{**}, \hat{q}_{00}, \hat{q}_{10}) \) with \( \hat{q}_{00}, \hat{q}_{10} \in (0, 1) \) such that the following five conditions are satisfied as

\[
Pr(Y_2 = 1 | Y_1 = 1, W = 1) = f(\beta_0^{**} + \beta_1^{**} + \beta_2^{**}) = q_{11}
Pr(Y_2 = 1 | Y_1 = 1, W = 0) = f(\beta_0^{**} + \beta_2^{**}) = \hat{q}_{10}
Pr(Y_2 = 1 | Y_1 = 0, W = 1) = f(\beta_0^{**} + \beta_1^{**}) = \hat{q}_{01}
Pr(Y_2 = 1 | Y_1 = 0, W = 0) = f(\beta_0^{**}) = \hat{q}_{00}.
\]
\[
\hat{q}_{00}r_{00} + \hat{q}_{10}r_{10} + q_{01}r_{01} + q_{11}r_{11} = q_{00}r_{00} + q_{10}r_{10} + q_{01}r_{01} + q_{11}r_{11}.
\]

(4.11)

The proof of THEOREM 1 is included in Appendix B.

4.2.3 AP Yahoo News Data

I focus on the baseline and wave 9 from the APYN study and pursue a two wave panel study with one refreshment sample available in the second wave. In addition to the measures described in Section 3.5, I collect more outcome measures of interest in wave 1 and wave 2 as shown in Table 4.3 and more demographic variables in the baseline shown in Table 4.4. The variables beginning with FAV directly measure attitudes towards each of the candidates (Barack Obama, John McCain, etc.) and the parties (The Democratic Party and The Republican Party). The variables beginning with INT4 ask for whether any of the following words (excited, interested, angry, hopeful, overwhelmed, etc.) describes how the participants feel about the upcoming presidential election. The variable VOT3 asks if the 2008 general election for the Present were being held today the participant would vote for the Democratic candidate or the Republican candidate. This is of strong interest for political analysis. However, VOT3 has 32.54\% missing values in wave 1, 17.11\% missing values for the respondents staying in wave 2, and 18.32\% missingness in the refreshment sample.

For demographic information, I use the categorical variable DURATION to represent the time in minutes for the participants to finish the whole interview. LV2 described whether the participant voted or not in the 2004 presidential election. VH1 asked which candidate the interviewer voted for the 2004 election for president. REL1 asked whether the interviewer was born-again or evangelical Christian or not. REL2 described the religious preference. These collected demographic measures in the APYN study are also subject to item nonresponse.
Table 4.3: Collected outcome measures for BLPM models in the baseline and wave 2 for APYN study

<table>
<thead>
<tr>
<th>Variable</th>
<th>Interpretation</th>
<th>Levels</th>
<th>W1 (2735)</th>
<th>W2 (1724)</th>
<th>Ref (464)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MD1</td>
<td>heading in R direction</td>
<td>2</td>
<td>14 (.51)</td>
<td>7 (.41)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>MD2</td>
<td>happy</td>
<td>4</td>
<td>4 (.15)</td>
<td>6 (.35)</td>
<td>1 (.22)</td>
</tr>
<tr>
<td>FAV1-1</td>
<td>George W. Bush</td>
<td>4</td>
<td>108 (3.95)</td>
<td>79 (4.58)</td>
<td>8 (1.72)</td>
</tr>
<tr>
<td>FAV1-4</td>
<td>Barack Obama</td>
<td>4</td>
<td>550 (20.11)</td>
<td>95 (5.51)</td>
<td>20 (4.31)</td>
</tr>
<tr>
<td>FAV1-9</td>
<td>John McCain</td>
<td>4</td>
<td>709 (25.92)</td>
<td>95 (5.51)</td>
<td>23 (4.96)</td>
</tr>
<tr>
<td>FAV2-1</td>
<td>The Democratic Party</td>
<td>4</td>
<td>70 (2.56)</td>
<td>53 (3.07)</td>
<td>12 (2.59)</td>
</tr>
<tr>
<td>FAV2-2</td>
<td>The Republican Party</td>
<td>4</td>
<td>106 (3.88)</td>
<td>60 (3.48)</td>
<td>11 (2.37)</td>
</tr>
<tr>
<td>LV1</td>
<td>registered to vote</td>
<td>2</td>
<td>9 (.33)</td>
<td>8 (.46)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>LV3</td>
<td>interest about campaign</td>
<td>4</td>
<td>6 (.22)</td>
<td>5 (.29)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>CND1</td>
<td>thought to candidates</td>
<td>2</td>
<td>5 (.18)</td>
<td>8 (.46)</td>
<td>3 (.65)</td>
</tr>
<tr>
<td>LV31</td>
<td>likelihood to vote</td>
<td>4</td>
<td>18 (.66)</td>
<td>220 (12.76)</td>
<td>57 (12.28)</td>
</tr>
<tr>
<td>VOT3</td>
<td>vote for 2004</td>
<td>2</td>
<td>890 (32.54)</td>
<td>295 (17.11)</td>
<td>85 (18.32)</td>
</tr>
<tr>
<td>INT4-1</td>
<td>Excited</td>
<td>2</td>
<td>5 (.18)</td>
<td>8 (.46)</td>
<td>3 (.65)</td>
</tr>
<tr>
<td>INT4-2</td>
<td>Interested</td>
<td>2</td>
<td>5 (.18)</td>
<td>8 (.46)</td>
<td>3 (.65)</td>
</tr>
<tr>
<td>INT4-3</td>
<td>Frustrated</td>
<td>2</td>
<td>5 (.18)</td>
<td>8 (.46)</td>
<td>3 (.65)</td>
</tr>
<tr>
<td>INT4-4</td>
<td>Bord</td>
<td>2</td>
<td>5 (.18)</td>
<td>8 (.46)</td>
<td>3 (.65)</td>
</tr>
<tr>
<td>INT4-5</td>
<td>Angry</td>
<td>2</td>
<td>5 (.18)</td>
<td>8 (.46)</td>
<td>3 (.65)</td>
</tr>
<tr>
<td>INT4-6</td>
<td>Proud</td>
<td>2</td>
<td>5 (.18)</td>
<td>8 (.46)</td>
<td>3 (.65)</td>
</tr>
<tr>
<td>INT4-7</td>
<td>Hopeful</td>
<td>2</td>
<td>5 (.18)</td>
<td>8 (.46)</td>
<td>3 (.65)</td>
</tr>
<tr>
<td>INT4-8</td>
<td>Overwhelmed</td>
<td>2</td>
<td>5 (.18)</td>
<td>8 (.46)</td>
<td>3 (.65)</td>
</tr>
<tr>
<td>INT4-9</td>
<td>Helpless</td>
<td>2</td>
<td>5 (.18)</td>
<td>8 (.46)</td>
<td>3 (.65)</td>
</tr>
</tbody>
</table>

This significant missing proportion of item nonresponse will lead to reduced sample sizes and even inferential bias if we just consider complete case analysis. I impute this missingness due to item nonresponse and the potentially non-ignorable missing values due to attrition simultaneously. Because of the massive number of outcome measures and demographic variables, the conditional models cannot efficiently take account into all the correlated variables sequentially. I implement the flexible imputation models featuring the DPMPM distributions for the extensive set of categorical variables as described in Chapter 2 and develop Bayesian latent pattern mixture models. This is essentially a generalization of latent class analysis. Latent class analysis, where the number of clusters is pre-determined, becomes a useful tool combining
Table 4.4: Collected demographic variables for BLPM models from the APYN study

<table>
<thead>
<tr>
<th>Variable</th>
<th>Interpretation</th>
<th>Levels</th>
<th>Panel (2735)</th>
<th>Ref (464)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PPAGECT4</td>
<td>Age</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PPDUALIN</td>
<td>Dual Income</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PPEDUCAT</td>
<td>Education</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PPETHM</td>
<td>Race</td>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PPGENDER</td>
<td>Gender</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PPHHSIZE</td>
<td>Household Size</td>
<td>7</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PPHOUSE</td>
<td>Housing Type</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PPINCIMP</td>
<td>Household Income</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PPMARIT</td>
<td>Marital Status</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PPMSACAT</td>
<td>MSA Status</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PPREG4</td>
<td>Region</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PPRENT</td>
<td>Ownership Status</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PPWORK</td>
<td>Employment Status</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>WORKOLD</td>
<td>Employment Status(Alt)</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PPNET</td>
<td>HHs with Internet</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>WEBTV</td>
<td>webtv provided</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PARTYID</td>
<td>party</td>
<td>3</td>
<td>10 (.37)</td>
<td>7 (1.51)</td>
</tr>
<tr>
<td>DURATION</td>
<td>interview time</td>
<td>5</td>
<td>0(0)</td>
<td>0(0)</td>
</tr>
<tr>
<td>LV2</td>
<td>vote or not in 2004</td>
<td>4</td>
<td>13 (.48)</td>
<td>1 (.22)</td>
</tr>
<tr>
<td>VH1</td>
<td>voted candidate in 2004</td>
<td>3</td>
<td>670 (24.50)</td>
<td>122 (26.29)</td>
</tr>
<tr>
<td>ID1</td>
<td>liberal or conservative</td>
<td>5</td>
<td>57 (2.08)</td>
<td>10 (2.16)</td>
</tr>
<tr>
<td>REL1</td>
<td>evangelical Christian</td>
<td>2</td>
<td>51 (1.86)</td>
<td>8 (1.72)</td>
</tr>
<tr>
<td>REL2</td>
<td>religious preference</td>
<td>6</td>
<td>37 (1.35)</td>
<td>11 (2.37)</td>
</tr>
<tr>
<td>REL3</td>
<td>attend religious services</td>
<td>5</td>
<td>20 (0.73)</td>
<td>8 (1.72)</td>
</tr>
</tbody>
</table>

Pattern mixture models for missing data problems in longitudinal settings.

4.3 Latent Class Models

Latent class analysis (Muthén et al., 2003) has been proposed to handle missing data and dropout process simultaneously. The underlying idea is to assume the data are generated from a mixture of different component models. The class membership is inferred from the data, and the pattern in the typical pattern mixture models is determined by the dropout process and known a priori. The dropout pattern can be extended from the two wave to multiple wave surveys and intermittent missing data.
One attrition indicator can be used for each follow-up wave, so that the dropout pattern is a multinomial vector consisting of all the attrition indicators. The well known latent class analysis refers to the case with a discrete number of subgroups, in which a latent categorical variable induces the within subject correlation.

Roy (2003) proposes a latent dropout class model for random effects of normal variables with many possible discrete dropout times following the convolution form
\[ f(Y, W) = f(W) \sum_S f(Y|S)f(S|W). \]
He assumes the distribution for the completed responses is a mixture over fewer number of latent class indicators \( S \) than the number of dropout patterns \( K \), and the class memberships are unknown. This combines pattern mixture models and latent class analysis. Roy (2003) considers the monotone missingness, and supposes that the latent variable \( S \) is independent of the other subject level random effects. The data are conditionally independent given the latent class, and the missing data are assumed MAR given the latent class. This assumption cannot be verified from the observed data.

Roy and Daniels (2008) extend the approach of Roy (2003) to generalized linear models with nonlinear link functions connecting the means \( E(Y_i|S_i) \) to covariates \( X_i \). They incorporate uncertainty in the number of classes through approximate Bayesian model averaging. Roy and Daniels (2008) assume in the ordinal regression model (actually, logistic link) that the intercept is monotonically changing across classes \( \lambda_{01} \leq \cdots \leq \lambda_{0,M-1} \) with the slope fixed as \( \lambda_1 \).

Lin et al. (2004) consider the other convolution form of conditional models to handle intermittent missingness; \( f(Y, W) = \sum_S f(Y|S)f(W|S)f(S) \). This avoids the sensitivity to misspecification of the direct dependence of the missingness pattern and the variables, while all the variables and the missingness pattern influence the choice of latent classes. The missingness process is assumed to be conditionally independent of the longitudinal outcomes given the latent classes. The conditional independence assumption is checked using a new noniterative approach (Bandeen-
Roche et al., 1997) without imputation. Beunckens et al. (2008) apply the similar
decomposition form but with a different link function- sequential ordinal regression
model- connecting the class membership to dropout pattern.

Dantan et al. (2008) compares the difference of pattern mixture models treating
dropout patterns as covariates for outcome variables \( Y|X = \sum_W [Y|W,X][W|X] \),
simple latent class model assuming dropout patterns are predictors for the class mem-
bership probabilities \( Y,W|X = \sum_S [Y|S,X][S|W][W] \) and joint latent class model
assuming the dropout patterns and the data jointly depend on latent classes but are
conditionally independent \( Y,W|X = \sum_S [Y|S,X][W|S,X][S] \). Class membership is
unknown, but the weight for latent classes depends on the dropout pattern. Hence,
the data are generated from a mixture of “latent patterns”, rather than the observed
dropout patterns. The latent pattern mixture models in the literature have a pre-
determined number of latent classes. These existing models are designed for only a
limited number of measures and can only be identified under the CI assumption.

4.4 Bayesian Latent Pattern Mixture Models

I extend the latent pattern mixture models to handle categorical data of high di-
mension beyond the context of continuous data of low dimension in the literature.
I propose a nonparametric Bayesian approach to let the data infer the number of
latent classes and incorporate the uncertainty. Two main procedures are proposed.

- **Dependent Bayesian latent pattern mixture (Dependent BLPM) model** has the
  convolution form

  \[
  f(Z, W) = f(W) \sum_s f(S|W)f(S|W).
  \]  
  \[ (4.12) \]

- **Joint Bayesian latent pattern mixture (Joint BLPM) model** has the convolution
Under the conditional independence (CI) assumption, given the latent class, the data $Z$ are independent of the dropout pattern $W$. Hence, we have

- Dependent BLPM model under CI assumption

\[
f(Z, W) = f(W) \sum_s f(Z|S) f(S|W). \tag{4.14}
\]

- Joint BLPM model under CI assumption

\[
f(Z, W) = \sum_s f(Z|S) f(W|S) f(S). \tag{4.15}
\]

4.4.1 Intuitive Compare of Dependent BLPM and Joint BLPM

Dependent BLPM treats the dropout pattern as a covariate for the class membership model. As described in Roy (2003), this is useful when a large number of dropout patterns are involved in the longitudinal data. The common pattern mixture models assume the subjects with the same dropout time follow the same distribution, which may be too strong or unrealistic in practice. Parameters may not be identifiable for some sparse patterns with few subjects. A smaller number of latent classes can avoid this problem. Different from the pattern mixture models, the relationship between the dropout patterns and class membership is not deterministic. Regularized prior distributions (e.g., DP prior) can be assumed on the class membership in the nonparametric Bayesian framework to induce sparsity and flexibility. Dependent stick-breaking process can be implemented when the allocation weights have covariates.
Roy (2003) only considered monotone missingness with observed dropout patterns. The marginal distribution of $W$ is ignored when estimating the class membership. When dropout patterns for some subjects are missing, dependent BLPM can do the imputation by drawing samples from the posterior distribution, but the posterior distribution for $W$ has to involve the likelihood for class membership. Joint BLPM model provides a more natural way to handle missing dropout patterns.

As described in Lin et al. (2004), joint BLPM model can handle arbitrary missing data patterns (intermittent and dropout) embodied by subjects’ visit process, which can occur irregularly or in continuous time. The joint BLPM model does not assume predefined discrete patterns of missing data, but uses latent classes to discover joint pattern of missingness and the data by themselves.

In my study, I have refreshment samples available with unknown attrition indicators. We can impute the missing values of $W$ straightforwardly conditional on the latent classes using Joint BLPM. Since the dropout pattern is not fully missing and we observe $W$ for the panel data, we can modify the dependent BLPM as well. For high dimensional data, I use the latent class representation to reduce dimension. I do not necessarily assume the number of the latent classes is less than the number of dropout patterns. For example, in the two wave panel, the attrition indicator is a binary variable. The dependent BLPM offers a more explicit representation to show whether the class memberships are different across the subgroups of individuals with different dropout patterns. Both dependent BLPM and joint BLPM can handle the multiple follow-up waves with intermittent missing data. If the number of the dropout patterns in the multiple wave study is large, the dependent BLPM can reduce the data dimension and regularize the patterns simultaneously.
4.5 BLPM under CI Assumption

4.5.1 Dependent BLPM

I propose the dependent BLPM with the missingness mechanism as a predictor for the class allocation probability. I use DP prior distributions on the class membership weights. That is, the weights in the DPM models depend on the missingness mechanism. Given the latent class, the variables are assumed as conditionally independent and generated from the corresponding item multinomial distributions. The generalized Dependent DP (DDP) mixture (Rodríguez and Dunson, 2011; Chung and Dunson, 2009) of products of multinomial distributions are used for imputation.

First I make the CI assumption that given latent classes, the missing values and the attrition are conditionally independent, using dependent BLPM illustrated in model (4.14). Comparing to literature on related models, I have the extra data source of refreshment samples available. I would like to compare the performances of using panel data only and combining panel and refreshment data under the CI assumption. All $X$, $Y_1$ and $Y_2$ influence the choice of latent classes and then depend on $W$. I do not distinguish the outcome or the demographic variables in terms of the missingness mechanism.

Rodríguez and Dunson (2011) develop a new class of nonparametric Bayesian models with probit stick-breaking process as priors where the weights of the process are constructed as probit transformations of normal random variables. I use the DDP where the beta distribution for the sticking variables $V_h$ in the DP are replaced by probit transformations of normal random variables with covariates $\Phi(\eta_h(x))$, so that $\pi_h(x) = \Phi(\eta_h(x)) \prod_{l \neq h} \{1 - \Phi(\eta_l(x))\}$. In Chung and Dunson (2009), $\eta_h(x)=\alpha_h + f_h(x)$. Rodríguez et al. (2010) propose a stochastic latent stick-breaking process with similar formalization. The probit stick-breaking prior can be specified as

$$V_h = \Phi(\mu_h), \quad \mu_h \sim N(\mu, \sigma^2). \quad (4.16)$$
If $\mu=0$ and $\sigma=1$, $V_h$ will follow a uniform distribution on $[0,1]$, which is equivalent to the DP with precision parameter $\alpha=1$. Rodríguez and Dunson (2011) investigate how the precision parameters $\mu$ and $\sigma$ control the structure of partitions generated by the model. The expected number of distinct clusters as sample size grows and the assignment of observations are controlled by $\mu$ and $\sigma$ for non-atomic base measures. The shrinkage effect of the model is determined by the prior distributions $N(\mu, \sigma^2)$ on $\mu_h$.

The resulting construction for the weights of the process is reminiscent of the continuation ratio probit model popular in survival analysis. For example, Albert and Chib (2001) considered the application of sequential ordinal modeling to survival data. Suppose that one observes independent observations $s_1, \ldots, s_n$, where each $s_i$ is an ordinal categorical response variable with $H$ possible values $\{1, \ldots, H\}$. In the sequential ordinal model, $s_i$ can take the value $h$ only after the levels $1, \ldots, h-1$, and the probability of stopping in level $h$ ($1 \leq h \leq H-1$) conditional on the event that the $h$th level is reached is given as $Pr(s_i=h|s_i \geq h, \gamma, \beta) = F(\gamma_h - X_i^T \beta)$, where $\gamma=(\gamma_1, \ldots, \gamma_{H-1})$ are unordered cutpoints and $X_i$ denotes the covariates. This is referred as the discrete-time hazard function. It follows that the probability of stopping at level $h$ is given by $Pr(s_i=h|\gamma, \beta) = F(\gamma_h - X_i^T \beta) \prod_{k=1}^{h-1} \{1 - F(\gamma_k - X_i^T \beta)\}$ for $h \leq H-1$ and $Pr(s_i=H|\gamma, \beta) = \prod_{k=1}^{H-1} \{1 - F(\gamma_k - X_i^T \delta)\}$.

I apply the probit stick-breaking construction for the DDP mixture of products of multinomial distributions assuming the number of classes and class membership are unknown and include the missingness mechanism as a covariate. Recall that $s_i$ denotes the latent class indicator. The model for the categorical variables and the attrition is as follows

$$Z_{ij} \sim \text{Multinomial}\{\{1, \ldots, d_j\}, \psi_{s_1}^{(j)}, \ldots, \psi_{s_d}^{(j)}\} \quad (4.17)$$

$$W \sim \text{Bernoulli}(\gamma_0). \quad (4.18)$$
The probit stick-breaking prior distributions is specified as

\[ s_i \sim \sum_{h=1}^{\infty} \pi_h(w_i) \]  
\[ \pi_h(w_i) = \Phi_h(w_i) \prod_{l<h}(1 - \Phi_l(w_i)) \]  
\[ \Phi_h(w_i) = \Phi(\beta_{0h} + \beta_1 w_i). \]

I assume the intercepts are varying across clusters and the slope is the same among clusters. Since the covariate \( w \) only has two possible values, I assume the same slope for simplicity.

The prior distribution on the item probability parameters \( \psi^{(j)}_h = (\psi^{(j)}_{h1}, \ldots, \psi^{(j)}_{hd_j}) \) is a flat Dirichlet distribution \((a_{j1}, \ldots, a_{jd_j})\), where \( d_j \) is the number of categories for variable \( j \). I set \( a_{j1} = \ldots = a_{jd_j} = 1 \). Suppose the prior distribution on the probability \( Pr(W=1) = \gamma_0 \) is Beta\((a_{\gamma_0}, b_{\gamma_0})\). The values of \((a_{\gamma_0}, b_{\gamma_0})\) are selected to make the prior carry little information, such as \( a_{\gamma_0} = b_{\gamma_0} = 1 \). The mean and variance in the normal distributions as hyper prior distributions for \( \beta \) determine the shrinkage effect. I choose independent prior distributions \( N(0,1) \) for each \( \beta_{0h} \) and \( \beta_1 \).

I introduce latent variables \( t_{ih} = \beta_{0h} + W_i \beta_1 + \epsilon_{ih} \) with \( \epsilon_{ih} \sim N(0,1) \) identically and independently. Then, \( t_{ih}|\beta \sim N(X_{ih}^{*}, \beta, 1) \), where \( X_{ih}^{*} = (0,0, \ldots, 1,0,0,W_i) \) with 1 in the \( h \)th column and \( \beta = (\beta_{0h}, h=1, \ldots, \infty, \beta_1) \). The individuals are assigned to class \( h \), that is, \( s_i = h \), if \( t_{i1} < 0, \ldots, t_{i,h-1} < 0 \) and \( t_{ih} \geq 0 \). Given the latent class indicator and the latent variables, the likelihood can be written as

\[ l(Y_1, Y_2, X, W|\theta, S_i, t_{ih}) = \prod_{i=1}^{N} f(Z_i|S_i, \Psi) f(S_i|W_i, t_{ih}, \beta) f(W_i|\gamma_0). \]

Neither \( \beta \) or \( \pi \) can be identified because of possible label switching.

I use blocked Gibbs sampler, truncating the stick breaking process at a large enough level \( H^* \) for the MCMC algorithm in posterior computation. The latent
variable representation (Albert and Chib, 2001) facilitates conjugate MCMC updating for the regression coefficients $\beta$. The steps for posterior updating are included in Appendix C.

4.5.2 Joint BLPM

Extending the case in Lin et al. (2004), I propose the joint BLPM model under the conditional independence assumption $f(Z,W)=\sum_S f(Z|S)f(W|S)f(S)$, shown as Model (4.15). Given the latent class, the missing data mechanism is independent of the collected variables, and the variables are mutually independent with each other. Here I treat the attrition indicator $W$ as another collected variable and impute its missing values for the refreshment samples with other data. I implement the joint BLPM model (4.15) as a straightforward extension of the DP mixture of products of the multinominal distributions. The summary of the model and prior distributions is as below

$$W \sim \text{Bernoulli}(\psi_{h1}^{(w)})$$

$$Z_{ij} \sim \text{Multinomial}\{1, \ldots, d_j\}, \psi_{s1,1}^{(j)}, \ldots, \psi_{sd_j}^{(j)}$$

$$s_i \sim \sum_{h=1}^{\infty} \pi_h$$

$$\pi_h = V_h \prod_{l<h} (1 - V_l)$$

$$V_h \sim \text{beta}(1, \alpha)$$

$$\psi_h^{(j)} = (\psi_{h1}^{(j)}, \ldots, \psi_{hd_j}^{(j)}) \sim \text{Dirichlet}(a_{j1}, \ldots, a_{jd_j})$$

$$\psi_{h1}^{(w)} \sim \text{Beta}(a_{w1}, a_{w0})$$

$$\alpha \sim \text{Gamma}(a_{\alpha}, b_{\alpha})$$

I use blocked Gibbs sampler with the truncation level $H^*$ for the posterior computation as in Appendix D.
4.5.3 Simulation Study

I use simulation examples to study the performance of the dependent BLPM and joint BLPM under the CI assumption on generated datasets under the CI assumption and generated datasets violating the CI assumption. For each case, I implement the procedures on the whole dataset (with both panel data and refreshment samples) and on the panel data only.

I. Dependent BLPM

Case 1: Clustered data under CI assumption

Suppose the sample size in the panel is $N_p=1500$, and the size for refreshment sample is $N_r=500$. The number of demographic variables $X$ is $q=50$, and the dimension of the outcome variables $Y_1$ and $Y_2$ in the two waves is $p_1=p_2=50$. Assume they are binary variables.

To generate data, first I simulate the attrition indicator variable $W$ from a Bernoulli distribution with probability $\gamma_0=.7$. Assume the data are generated from $nc_0=4$ different subpopulations. The weights depend on $W$: $\pi_h(w) = \Phi_h(w) \prod_{k<h}(1-\Phi_k(w))$ and $\Phi_h(w)=\Phi(\beta_{0h}+\beta_1w)$, for $h=2,\ldots, nc_0-1$; $\pi_1(w)=\Phi_1(w)$; and $\pi_{nc_0}(w) = 1 - \sum_{h=1}^{nc_0-1} \pi_h(w)$. Suppose the true value of $(\beta_{01},\ldots,\beta_{0,nc_0-1})$ is $[2,2,\ldots,2]$ and $\beta_1=-3$. The latent class indicators are generated from multinomial distributions with probability vector $\pi$ and sample size $N=N_p+N_r$. I assume the component specific probabilities $\psi_{hj}=\pi_h(1/h\pi_j+2\pi_j)/(d_j-1)$, for $j=1,\ldots, p$, $h=1,\ldots, nc_0$, and $c_j=1,\ldots, d_j-1$. Here, $\psi_{hdj}=1 - \sum_{c_j=1}^{d_j-1} \psi_{hcj}$, which change across clusters and variables. Both the panel data and refreshment samples are generated based on these latent class indicators and component specific probabilities.

I use a flat hyper prior distribution for $\gamma_0$ using a Beta distribution with $a_0=b_0=1/4$ and for $\psi$ using a Dirichlet distribution with $a_{j1}=\ldots=a_{jd_j}=1$. The prior distribu-
tion for the coefficients in the probit regression is a multivariate normal distribution $MVN_{nc}(0, I)$, which is equivalent for independent Beta$(1, 1)$ or Uniform$(0, 1)$ distributions.

The initial values for $\beta$ are arbitrarily assigned and all equal to 1. I use the marginal probabilities calculated from the data as starting values for $\psi$. I tried different initial values for the missing data; the outputs after convergence are the same. I recommend using draws from the marginal distribution based on observed data as the initial values. For illustration, I use the true values of the missing data as the starting points.

In each simulation, I run one MCMC chain with 50000 iterations and a thinning value of 50 until achieving convergence based on diagnostic checks on the weighted marginal quantities, which are not affected by label switching. As a general check to evaluate the approach, I collect the posterior samples for the mean of $Y_1$ and $Y_2$, the marginal probabilities that all the variables fall into the 1st category for the 1st individual, and the marginal probabilities that all the variables fall into the 1st category for the $Np$th individual. Their mean values versus the truth are shown in the top four plots of Figure 4.1. The 1st individual participated in both waves while the $Np$th individual dropped out after participating in wave 1. Then I implement the procedure only on the panel data, where we only need to impute missing values of $Y_2$. I collect the same quantities, ignoring $Y_1$, since it is not subject to attrition. Their mean values versus the truth are shown in the bottom four plots of Figure 4.1. As shown in Figure 4.1, both cases can generate plausible imputation values, while only using the panel data has more variability than using all the data sources.

Case 2: Clustered data violating CI assumption

I assume the component specific probabilities for data simulation depend on $w$ by
Figure 4.1: Simulation outputs in Case 1 of dependent BLPM under CI. Top four plots are outputs of all available data; bottom four plots are outputs of panel data only. The outputs from top left to bottom right in sequence are: mean of $Y_1$ vs the truth on all data, mean of $Y_2$ vs the truth on all data, $Pr(z_1=1)$ on all data, $Pr(z_{Np}=1)$ on all data, mean of $Y_2$ vs the truth only on panel, $Pr(z_1=1)$ only on panel, $Pr(z_{Np}=1)$ only on panel, and a comparison of $Pr(Z=1)$ for panel samples estimated on all data, only panel and the truth.
generating from

\[ \psi_{i,h,c,j}^{(j)} = w_i \cdot h \cdot j / (h \cdot j + 2 \cdot j + 2) / d_j + (1 - w_i) \cdot (1 - h \cdot j / (h \cdot j + 2 \cdot j + 2) / d_j), \]  

for \( i=1, \ldots, N, \ j=1, \ldots, p, \ h=1, \ldots, nc_0, \) and \( c_j=1, \ldots, d_j - 1. \) Here \( \psi_{i,h,d_j}^{(j)} = 1 - \sum_{c_j=1}^{d_j-1} \psi_{i,h,c_j}^{(j)}, \) which change across clusters, variables and individuals through \( w_i. \) The panel data and refreshment samples are generated based on these latent class indicators and component specific probabilities. The other settings are the same as Case 1 above.

The posterior mean for \( \gamma_0 \) is .6823, which is the same as the posterior mean value (.6822) for completed values of \( W \) after imputation and the truth (.6830). The outputs are shown in the top four plots in Figure 4.2. Then I implement the procedure only on the panel data, where we only need to impute missing values of \( Y_2. \) Using panel data only results in biased inference for \( Y_2 \) as shown in the bottom four plots of Figure 4.2. When the data violates the CI assumption, refreshment samples can adjust the inference and reduce the bias. The top four plots in Figure 4.2 illustrate the plausible results of imputation using all the data sources.

\((II.)\) Joint BLPM

Similarly, I use simulation examples to study the performance of the joint BLPM under the CI assumption on generated datasets under the CI assumption and generated datasets violating the CI assumption. Suppose the sample sizes \( N_p=2000, \) and \( N_r=1000. \) The numbers of variables are \( q=50 \) and \( p_1=p_2=50. \) First I assume there are \( nc_0=4 \) latent classes and generate the \( nc_0-1 \) stick-breaking variables from \( \text{Beta}(1,2). \) The weights for classes \( \pi \) can be calculated correspondingly and then the latent class indicator variable is drawn from a multinomial distribution with probability vector \( \pi \) and sample size \( N. \) I simulate \( W \) based on the latent class indicator with probability \( h/(h+1) \) for the four different clusters, for \( h=1, \ldots, nc_0. \)
Figure 4.2: Simulation outputs in Case 2 of dependent BLPM under CI. Top four plots are outputs of all available data; bottom four plots are outputs of panel data only. The outputs from top left to bottom right in sequence are: mean of $Y_1$ vs the truth on all data, mean of $Y_2$ vs the truth on all data, $Pr(z_1=1)$ on all data, $Pr(z_{Np}=1)$ on all data, mean of $Y_2$ vs the truth only on panel, $Pr(z_1=1)$ only on panel, $Pr(z_{Np}=1)$ only on panel, and a comparison of $Pr(Z=1)$ for panel samples estimated from all data, only panel and the truth.
I use flat hyper prior distribution for $\alpha$ using a Beta distribution with $a_0 = b_0 = 1/4$ and for $\psi$ using a Dirichlet distribution with $a_{j1} = \ldots = a_{jd_j} = 1$. For simplicity I truncate the blocked Gibbs sampler at the level $H^* = nc_0$, which is done to speed up the convergence. For a large truncation level, as long as there are enough iterations for convergence, the conclusion does not change. I run the MCMC chain for 50000 iterations with a thinning value of 50. After convergence check, I compare the performances of the procedure on both panel and refreshment samples and only on the panel data.

**Case 1: Clustered data under CI assumption**

The data $(X, Y_1, Y_2)$ in the panel and refreshment samples are simulated conditional on the latent class indicator from multinomial distributions with probability vector $(\psi_{hc_j}^j, c_j = 1, \ldots, d_j)$, where $\psi_{hc_j}^j = 0.5 \cdot h/d_j - 0.05$ if $1 \leq c_j \leq d_j - 1$ and $\psi_{hd_j}^j = 1 - \sum_{c_j = 1}^{d_j-1} \psi_{hc_j}^j$. I collect the posterior samples for mean of $Y_1, Y_2$ and $W$ and the marginal probabilities for all the variables with the 1st level (for binary variables, these are actually mean values), as the outputs for the implementation of the joint BLPM model on both the panel and refreshment samples. The comparison of their mean values versus the truth is shown in the top four plots in Figure 4.3. The posterior mean of $W$ is a bit different from the truth, but not much. The bottom four plots in Figure 4.3 present the outputs for the joint BLPM applied only on the panel data. They include the posterior samples of $\alpha$ versus the truth, the posterior samples of the mean value for $Y_2$ versus the truth, the marginal probabilities for all variables in the 1st category for the completed dataset in the last iteration from the chain and the marginal probabilities for all variables in the 1st category for the dropout subjects in the last iteration from the chain. The two marginal probabilities include the values from the imputation only using panel data, imputation using all the data and their true values. The two imputation procedures yield plausible results, which
could be improved with less bias by changing the prior settings or the increasing the number of iterations. There is no consistent pattern for the posterior mean samples to underestimate or overestimate the ratio of missingness after I run multiple chains with random initial settings.

**Case 2: Clustered data violating CI assumption**

Assume all the data still depends on W given latent classes. I do this by generating item specific probabilities using

\[
\psi_{i,h,c_j}^{(j)} = w_i \cdot h \cdot j / (h \cdot j + 2 \cdot j + 2) / d_j + (1 - w_i) \cdot (1 - h \cdot j / (h \cdot j + 2 \cdot j + 2)) / d_j, \tag{4.31}
\]

for \( i=1, \ldots, N, \ j=1, \ldots, p, \ h=1, \ldots, nc_0, \) and \( c_j=1, \ldots, d_j - 1. \) Here, \( \psi_{i,h,c_j}^{(j)} = 1 - \sum_{c_j=1}^{d_j-1} \psi_{i,h,c_j}^{(j)}. \) I collect the same quantities as in Case 1. However, only using the panel data for imputation yields biased estimation values related to the variables in \( Y_2. \) As shown in the bottom four plots of Figure 4.4, the mean estimates and the marginal probability distributions for completed \( Y_2 \) after imputation are different from the truth.

4.5.4 APYN Data Analysis

(I) Dependent BLPM

I implement the dependent BLPM under the CI assumption to the APYN study data. I consider 24 demographic variables in Table 4.4, 21 outcome measures in Table 4.3 from wave 1 and the same 21 outcome measures in Table 4.3 from wave 2.

I first calculate the marginal probabilities for the variables with missing values from their observed parts and use these marginal probability distributions to simulate the initial values for the missing data. For the missing values of \( Y_2 \) due to attrition in the panel, I use their marginal distributions in the refreshment and the marginal distributions estimated on the subjects stayed in wave 2 and borrow the
**Figure 4.3:** Simulation outputs in Case 1 of joint BLPM under CI. Top four plots are outputs of all available data; bottom four plots are outputs of panel data only. The outputs from top left to bottom right in sequence are: mean of $Y_1$ vs the truth on all data, mean of $Y_2$ vs the truth on all data, $Pr(W=1)$ vs the truth on all data, $Pr(Z=1)$ on all data, posterior samples of $\alpha$ versus the truth only on panel, mean of $Y_2$ vs the truth only on panel, $Pr(Z=1)$ only on panel, and a comparison of $Pr(Z=1)$ in one simulation for dropout subjects estimated from all data, only panel and the truth.
Figure 4.4: Simulation outputs in Case 2 of joint BLPM under CI. Top four plots are outputs of all available data; bottom four plots are outputs of panel data only. The outputs from top left to bottom right in sequence are: mean of $Y_1$ vs the truth on all data, mean of $Y_2$ vs the truth on all data, $Pr(W=1)$ vs the truth on all data, $Pr(Z=1)$ on all data, posterior samples of $\alpha$ versus the truth only on panel, mean of $Y_2$ vs the truth only on panel, $Pr(Z=1)$ only on panel, and a comparison of $Pr(Z=1)$ in one simulation for dropout subjects estimated from all data, only panel and the truth.
probability $Pr(W=1/0)$ based on the panel. Bayes rule is used to obtain the marginal probability distribution for $Y_2$ of dropout subjects. The corresponding initial values are generated based on the calculated probabilities. The initial values for $W$ is the refreshment samples are independently drawn from a Bernoulli distribution with success probability $N_{cp}/N_p=.6303$.

I use flat prior distribution Beta$(1,1,)$ on $\gamma_0$, Dirichlet distribution $(1,\ldots,1)$ on \(\Psi\), and $N(0,1)$ independently and identically on $\beta_1$ and on $\beta_{0h}$, with the shrinkage effect equivalent to Beta$(1,1)$ on $V_h$ in DP. I truncate the stick-breaking process as a large enough number $H^*=50$. The initial values of coefficients $\beta$ for the weights are set as 1. The starting point for the latent class indicator $s$ is drawn from a multinomial distribution with equal probability among $(1,\ldots, H^*)$ and sample size $N$. Based on the initial values of $W$, $s$ and $\beta$, I have the initial setting for $V_{si}(w_i)=\Phi(X^*_{i_{s_i}}\beta)$ and then $\pi_{s_i}(w_i)$ for each individual. The starting values for $\Psi$ are the marginal probabilities calculated from the initial completed datasets.

I run the MCMC chain for $T=100000$ iterations with a burn-in value of 20000 and a thinning value of 100. I collect the marginal probabilities for the variables of the 1st subject and the $N_p$th subject for convergence check. The trace plots for these marginal probabilities and the mean values for the variables show good mixing behavior and quick convergence. I also find that the marginal probabilities for the variables are the same for the 1st individual, who stayed in both waves of the panel, and for the $N_p$th subject, who dropped out of panel in wave 2.

The posterior mode of the number of distinct occupied clusters is 16, while the number across iterations is always less than 20. This shows that the truncation level at 50 is sufficient enough. The posterior samples for $\gamma_0$ converge fast and with a mean value .6304. The average for the posterior mean of imputed values of $W$ is .6305, which have good convergence performances. Both these two values are close to $Pr(W=1)=.6303$ in the panel. The average value for the posterior mean values of
$W$ in the refreshment sample is .6326. The ratio for the attrition is the same across the completed panel and the refreshment sample after imputation.

I collect $m=50$ multiple completed datasets from this single MCMC chain with a lag value of 1000. Figure 4.5 displays the marginal probabilities and MI confidence intervals for all the possible categories of $Y_2$, for the stayed and dropout individuals in the completed panel. The marginal probabilities for the two groups are essentially conditional probabilities depending on the attrition group. The mean estimates of two probabilities are similar to each other. This shows that the distribution of $Y_2$ is approximately independent of the attrition $W$, which is affected or in control by the CI assumption in the model, even though marginal dependence may exist under CI assumption.

I build analysis models $LV1 \sim X_1$, $LV3 \sim X_1$, $LV31 \sim X_1$, $CND1 \sim X_1$, $INT4-1 \sim X_2$ and $FAV1-4 \sim X_2$ on the multiple completed datasets after MI, where $X_1=$(age, edu, gender, race, income, marriage status) and $X_2=$(age, edu, gender, race, income, marriage status, partyID, ID1, REL3). The outputs for the mean estimates and confidence intervals ($\bar{q}_m \pm \sqrt{T_m}$) for the regression coefficients of these six models on the completed panel data and on the refreshment sample are shown in Figure 4.6. We can see that the estimates from the completed panel and from the refreshment sample are different from each other for some regression coefficients in the six models. The model with CI assumption restricts the correlation structure among these variables.

Figure 4.5 and Figure 4.6 show that the distribution of $Y_2$ is approximately independent of $W$ and the analysis on the completed panel is different from that on the refreshment sample. These facts suggest deficiencies in the performance of the dependent BLPM under the CI assumption for these data. The CI assumption restricts the model and does not fit the data well.
FIGURE 4.5: Comparison on marginal probabilities and 95% confidence intervals under MI for $Y_2$ between stayed and dropout samples in the completed panel using dependent BLPM under CI.

(II) Joint BLPM

I implement the joint BLPM under the CI assumption to the APYN study data using the same settings as in the data analysis for the dependent BLPM.

I use flat prior distribution $\alpha \sim \text{Gamma}(1/4, 1/4)$, and $\Psi \sim \text{Dirichlet}(1, \ldots, 1)$. I truncate the stick-breaking process as a large enough number $H^*=30$. The initial value of $\alpha$ is 1. The starting point for the latent class indicator $s$ is drawn from a multinomial distribution with equal probability from 1 to 10 and sample size $N$. The starting values for $\Psi$ are the marginal probabilities calculated from the initial completed datasets.
Figure 4.6: Comparison of coefficients in the six analysis models after MI using dependent BLPM under CI between completed panel data and refreshment samples. From top left to bottom right, the models in sequence are LV1~X1, LV3~X1, LV31~X1, CND1~X1, INT4-1~X2 and FAV1-4~X2.
I run the MCMC chain for $T=100000$ iterations with a burn in value of 20000 and a thinning value of 100. I collect the marginal probabilities for the variables for convergence check. The trace plots for these marginal probabilities and the mean values show good mixing behavior and quick convergence. The posterior samples for $\alpha$ converge fast with a mean value of 1.3061. The posterior mode of the number of distinct occupied clusters is 11, while the number across iterations is always less than or equal to 11. This shows that the truncation level at 30 is sufficient. The average for the posterior mean values of $W$ is .6032. This is close to the probability $Pr(W=1)=.6303$ in the panel. The average value for the posterior mean values of $W$ in the refreshment sample is .4433. The ratios for the attrition are different across the completed panel and the refreshment sample after imputation.

I collect $m=50$ multiple completed datasets from this MCMC chain with a lag value of 1000. I perform the same analysis after MI as that for the dependent BLPM, and add two more models $W \sim LV31$ and $W \sim LV3$ for the regression. The outputs are shown in Figure 4.7 and Figure 4.8. The marginal probabilities between the stayed and dropout individuals are close to each other. The analysis on the two regression models $W \sim LV31$ and $W \sim LV3$ is quite different between the completed panel data and the refreshment sample. The performances of the joint BLPM are similar to those under the dependent BLPM, where the CI assumptions seem to be too strong for the APYN data.

I obtain the mean estimates for $Pr(Y_2=1)$ and for the conditional probabilities $Pr(Y_2=1|W)$ based on the $m=50$ multiple completed datasets, where $W=0/1$, estimated on only the completed panel data and only the refreshment sample, as shown in Table 4.5. The marginal probabilities $Pr(Y_2=1)$ from the completed panel data try to match those from the refreshment sample. The conditional probabilities $Pr(Y_2=1|W)$ are similar with each other in terms $W=0/1$ in the completed panel data. Most variables, such as MD1, MD2, FAV1-1, FAV1-4, FAV1-9, FAV2-1, FAV2-
Figure 4.7: Comparison on marginal probabilities and 95% confidence intervals under MI for $Y_2$ between stayed and dropout samples in the completed panel using joint BLPM under CI.

2, VOT3, and INT4-1, have very different conditional probabilities $Pr(Y_2=1|W)$ in terms $W=0/1$ in the refreshment sample.

I compare these marginal, conditional and joint probabilities for $Y_2$ between the completed panel and refreshment sample based on multiple $m=50$ imputed datasets, illustrated in Figure 4.9. We can see that the CI assumption forces the conditional probabilities of $Y_2$ given $W$ to be the same in terms of $W=0/1$. The estimates for conditional and joint probabilities are quite different from each other between the completed panel and the refreshment sample. Figure 4.7, Figure 4.8, and Table 4.5 display that the distribution of $Y_2$ is approximately independent of $W$, and the analysis on the completed panel is different from that on the refreshment sample.
Figure 4.8: Comparison of coefficients in the eight analysis models after MI using joint BLPM under CI between completed panel data and refreshment samples. From top left to bottom right, the models in sequence are $W \sim LV31$, $W \sim LV3$, $LV1 \sim X_1$, $LV3 \sim X_1$, $LV31 \sim X_1$, $CND1 \sim X_1$, $INT4-1 \sim X_2$ and $FAV1-4 \sim X_2$. 
Figure 4.9: The marginal probability $Pr(Y_2=1)$, the conditional probability $Pr(Y_2=1|W)$ and the joint probability $Pr(Y_2=1, W)$ estimated from the completed panel (p) and refreshment (r) after MI using joint BLPM under CI; left 6 cases for $W=1$ and right 6 cases for $W=0$. 

123
These contradictions could reflect violations of the CI assumption.

In the simulation study for the BLPM models under the CI assumption, I implement a general check and find that the refreshment samples can help improve the estimates and reduce the bias due to attrition, compared to results under only panel data. The performance of BLPM models under the CI assumption in the APYN study shows that the CI assumption almost eliminates effects of the non-ignorable attrition, and the refreshment sample does not calibrate well or adjust the attrition effect. The assumptions of conditional independence between $X$ and $W$, and conditional independence between $Y_1$ and $W$ may affect the relationship between $Y_2$ and
W via the latent pattern structure. The marginal dependence between $Y_2$ and $W$ is not captured sufficiently under CI. The analysis after MI in completed panel is different from that on the refreshment sample.

4.6 Release the CI Assumption for BLPM

4.6.1 BLPM Model Specification

The data analysis in Section 4.5.4 suggests that the CI assumption is too strong. We can release the CI assumption when refreshment samples are available. Even given the latent class, the missing values of $Y_2$ still depend on missingness mechanism. This leads to the dependent BLPM and joint BLPM models that I now describe. In these models, $Y_1$ remains conditionally independent with $W$ given latent classes for identification.
The dependent BLPM model is specified as

\[ W \sim \text{Bernoulli}(\gamma_0) \]

\[ s_i \sim \sum_{h=1}^{\infty} \pi_h(w_i) \]

\[ \pi_h(w_i) = \Phi_h(w_i) \prod_{l<h} (1 - \Phi_l(w_i)) \]

\[ \Phi_h(w_i) = \Phi(\beta_{0h} + \beta_1 w_i) \]

\[
(Y_{i,2j}|w_i=1, -) \sim \text{Multinomial}\{\{1, \ldots, d_j\}, \psi_{s,1}^{(1j)}, \ldots, \psi_{s,d_j}^{(1j)}\}
\]

\[
(Y_{i,2j}|w_i=0, -) \sim \text{Multinomial}\{\{1, \ldots, d_j\}, \psi_{s,1}^{(0j)}, \ldots, \psi_{s,d_j}^{(0j)}\}
\]

\[
(Y_{i,1j}|-) \sim \text{Multinomial}\{\{1, \ldots, d_j\}, \psi_{s,1}^{(j)}, \ldots, \psi_{s,d_j}^{(j)}\}
\]

\[
(X_{i,j}|-, -) \sim \text{Multinomial}\{\{1, \ldots, d_j\}, \psi_{s,1}^{(j)}, \ldots, \psi_{s,d_j}^{(j)}\}
\]

\[
\psi_{h}^{(1j)} = (\psi_{h1}^{(1j)}, \ldots, \psi_{hd_j}^{(1j)}) \sim \text{Dirichlet}(a_{j1}, \ldots, a_{jd_j})
\]

\[
\psi_{h}^{(0j)} = (\psi_{h1}^{(0j)}, \ldots, \psi_{hd_j}^{(0j)}) \sim \text{Dirichlet}(b_{j1}, \ldots, b_{jd_j})
\]

\[ \gamma_0 \sim \text{Beta}(a_{\gamma_0}, b_{\gamma_0}) \]

\[ \beta = (\beta_{0h}, h=1, \ldots, \infty, \beta_1)' \sim \text{MVN}(0, \Sigma_{\beta}) \]

Data augmented posterior computation (blocked sampler with truncation level \( H^* \)) is included in Appendix E.

The joint BLPM model is specified as
\[
W \sim \text{Bernoulli}(\psi_{s_1}^{(w)})
\]
\[
s_i \sim \sum_{h=1}^{\infty} \pi_h
\]
\[
\pi_h = V_h \prod_{l<h} (1 - V_l)
\]
\[
V_h \sim \text{Beta}(1, \alpha)
\]
\[
(Y_{i,2j}|w_i=1, -) \sim \text{Multinomial}\{1, \ldots, d_j\}, \psi_{s_1}^{(1j)}, \ldots, \psi_{s_id_j}^{(1j)}
\]
\[
(Y_{i,2j}|w_i=0, -) \sim \text{Multinomial}\{1, \ldots, d_j\}, \psi_{s_1}^{(0j)}, \ldots, \psi_{s_id_j}^{(0j)}
\]
\[
(Y_{i,1j}|-) \sim \text{Multinomial}\{1, \ldots, d_j\}, \psi_{s_1}^{(j)}, \ldots, \psi_{s_id_j}^{(j)}
\]
\[
(X_{i,j}, -) \sim \text{Multinomial}\{1, \ldots, d_j\}, \psi_{s_1}^{(j)}, \ldots, \psi_{s_id_j}^{(j)}
\]
\[
\psi_h^{(1j)} = (\psi_{h_1}^{(1j)}, \ldots, \psi_{h_d_j}^{(1j)}) \sim \text{Dirichlet}(a_{j1}, \ldots, a_{jd_j})
\]
\[
\psi_h^{(0j)} = (\psi_{h_1}^{(0j)}, \ldots, \psi_{h_d_j}^{(0j)}) \sim \text{Dirichlet}(b_{j1}, \ldots, b_{jd_j})
\]
\[
\psi_h^{(j)} = (\psi_{h_1}^{(1j)}, \ldots, \psi_{h_d_j}^{(1j)}) \sim \text{Dirichlet}(c_{j1}, \ldots, c_{jd_j})
\]
\[
\psi_h^{(w)} \sim \text{Beta}(a_{w1}, a_{w0})
\]
\[
\alpha \sim \text{Gamma}(a_\alpha, b_\alpha).
\]

I use blocked Gibbs sampler with the truncation level \(H^*\) for the posterior computation as shown in Appendix F.

Because of the similar performance of dependent BLPM and joint BLPM in the data analysis in Section 4.5.4, I implement the joint BLPM model releasing the CI assumption between \(Y_2\) and \(W\) via simulation study and then apply the joint BLPM to the APYN panel data.
4.6.2 Simulation Study

Suppose the sample sizes $N_p = 2000$ and $N_r = 1000$. The numbers of binary variables are $q = 50$ and $p_1 = p_2 = 50$. First I assume there are $nc_0 = 4$ latent classes and generate the $nc_0 - 1$ stick-breaking variables from Beta$(1, 2)$. The weights for classes $\pi$ can be calculated correspondingly as $(.6239, .0558, .0131, .3073)$ and then the latent class indicator variable is drawn from a multinomial distribution with probability vector $\pi$ and sample size $N = 3000$. I simulate $W$ based on the latent class indicator with probability $h/(h + 1)$ for the four different clusters, for $h = 1, \ldots, nc_0$.

Assume the variables of $X$ and $Y_1$ are conditionally independent of $W$ given latent classes. The item specific probabilities are

$$\psi_{i, h, c_j}^{(j)} = \frac{h \cdot j \cdot (h \cdot j + 2 \cdot j + 2)}{d_j},$$

(4.32)

for $i = 1, \ldots, N, j = 1, \ldots, q + p_1, h = 1, \ldots, nc_0$, and $c_j = 1, \ldots, d_j - 1$. Here, $\psi_{i, h, d_j}^{(j)} = 1 - \frac{1}{\sum_{c_j=1}^{d_j-1} \psi_{i, h, c_j}^{(j)}}$.

Assume all the variables of $Y_2$ still depend on $W$ given latent classes. The item specific probabilities for $Y_2$ are assumed as

$$\psi_{i, h, c_j}^{(j)} = w_i \cdot h \cdot j \cdot (h \cdot j + 2 \cdot j + 2)/d_j + (1 - w_i) \cdot (1 - h \cdot j \cdot (h \cdot j + 2 \cdot j + 2))/d_j,$$

(4.33)

for $i = 1, \ldots, N, j = q + p_1 + 1, \ldots, p, h = 1, \ldots, nc_0$, and $c_j = 1, \ldots, d_j - 1$. Hence, $\psi_{i, h, d_j}^{(j)} = 1 - \frac{1}{\sum_{c_j=1}^{d_j-1} \psi_{i, h, c_j}^{(j)}}$.

I use flat hyper prior distribution for $\alpha$ using a Beta distribution with $a_0 = b_0 = 1/4$ and for $\psi$ using a Dirichlet distribution with $a_j = \ldots = a_{jd_j} = b_j = \ldots = b_{jd_j} = 1$. For simplicity I truncate the blocked Gibbs sampler at the level $H^* = 10$. The initial values for $V_h$ are simulated independently from Beta$(1, 2)$ and then yield the initial weight values $\pi_0$. The starting point for the latent class indicator $s$ is drawn from a multinomial distribution with probability vector $\pi_0$ and sample size $N$. The starting
values for $\Psi$ are also set as the marginal probabilities obtained from the initial completed dataset. Here, for simple illustration, I choose the truth the starting points for the missing values.

I run the MCMC chains for 10000 iterations with a burn-in value of 3000 and a thinning value of 7. The posterior mode for the number of occupied components is 7 and the posterior mean for $\alpha$ is 1 (this is different from the truth since I truncate the blocked sampler at a higher level than the truth). I collect the posterior samples for mean of $Y_1$, $Y_2$ and $W$, posterior samples of $\alpha$ and the marginal probabilities for all the variables within the 1st level for the stayed and dropout individuals. These quantities not affected by label switching demonstrate good convergence behavior.

The comparison of the estimated values versus the truth is shown in the top four plots in Figure 4.10. The posterior mean values of $Y_1$, $Y_2$ and $W$ are close to the truth. The mean value of $Pr(W=1)$ in the refreshment samples is .6950, which is the same as in the initial completed datasets. The bottom two plots in Figure 4.10 present the outputs for marginal probabilities under the joint BLPM model. They are close to the true values. The bias can be reduced by more similar structure for the clustering even under marginal probabilities as initial values. This shows that the joint BLPM model can be identified when the panel and refreshment samples are generated from the same population.

4.6.3 APYN Data Analysis

I implement the joint BLPM without the CI assumption for $Y_2$ to the APYN study data. The initial setting for the MCMC chain is the same as the Section 4.5.4. The initial value of $\alpha$ is 1. I use a flat prior distribution $\text{Gamma}(1/4, 1/4)$ on $\alpha$ and a Dirichlet distribution $(1, \ldots, 1)$ on $\Psi$. I truncate the stick-breaking process as a large enough number $H^* = 20$. The starting point for the latent class indicator $s$ is drawn from a multinomial distribution with probability vector $\pi_0$ and sample size $N$. 

129
Figure 4.10: Simulation outputs using joint BLPM: from top left to bottom right in sequence are, posterior mean estimates of $Y_1$, $Y_2$ and $W$, posterior samples of $\alpha$ and marginal probabilities of all variables within the 1st level for stayed and dropout individuals, compared to the truth.

130
The starting values for $\Psi$ are the marginal probabilities calculated from the initial completed datasets.

I run the MCMC chain for $T=100000$ iterations with a burn in value of 30000 and a thinning value of 70. I collect the marginal probabilities for the variables, which are weighted average values across all clusters, for convergence check. The trace plots for these marginal probabilities and the mean values for the variables show good mixing behavior and quick convergence.

The posterior samples for $\alpha$ converge fast and with a mean value .9800. The posterior mode of the number of distinct occupied clusters is 8, while the number across iterations is always less than or equal to 11. This shows that the truncation level at 20 is sufficient enough. The average value for the posterior mean values of $W$ in the refreshment sample is .8202, which have good convergence performance. This is a bit away from the probability $Pr(W=1)=.6303$ in the completed panel. The ratios for the attrition are different across the panel and the refreshment sample after imputation.

I collect $m=50$ multiple completed datasets from this MCMC chain with a lag value of 1000. I build the same analysis models as in Section 4.5.4: $W \sim LV31$, $W \sim LV3$, $LV1 \sim X_1$, $LV3 \sim X_1$, $LV31 \sim X_1$, $CND1 \sim X_1$, $INT4-1 \sim X_2$ and $FAV1-4 \sim X_2$. Figure 4.11 compare the mean estimates and the 95% confidence intervals under MI on the panel data and on the refreshment sample. The estimates are close, and the confidence intervals overlap with each other for most quantities. The longer confidence intervals are due to the uncertainty induced by imputation. The analysis separately on the panel and the refreshment sample has the similar performance.

The outputs for the comparison of the marginal probability distributions for $Y_2$ of the stayed individuals and the dropout in the completed panel data are shown in Figure 4.12. The posterior mean for the probabilities are different between the stayed and the dropout individuals for most variables, and some variables have very
significant difference, such as MD1, FAV1-1, FAV2-2, LV1, and LV31, etc. This can also be seen in Table 4.6. The distributions of $Y_2$ depend on $W$. I compare the marginal probabilities $Pr(Y_2=1)$, the conditional probabilities $Pr(Y_2=1|W)$ and the joint probabilities $Pr(Y_2=1, W)$ estimated from the completed panel and refreshment sample under MI using joint BLPM, as shown in Figure 4.13 and Table 4.6. The conditional probabilities $Pr(Y_2=1|W=1)$ which are observed in the panel, have similar estimates in the completed panel and the refreshment samples. The joint probabilities $Pr(Y_2=1, W=0)$ and $Pr(Y_2=1, W=1)$ also have similar values in the completed panel and the refreshment samples as shown in Table 4.7. The marginal probabilities $Pr(Y_2=1)$ from the panel are close to those from the refreshment samples. Ideally, if the refreshment samples are randomly from the population as the panel data, all the estimation for these quantities should be approximately the same. The estimate of $Pr(W=1)$ in the refreshment sample (.82) is larger than those in the panel (.63), and this will cause estimation difference for the conditional and joint probabilities of $Y_2$ and $W$.

I am also interested in the persistency of the wave 2 and wave 1 outcome measures and construct regression analysis to study their mutual dependency across panel. The measures in wave 2 are treated as outcome variables, and the corresponding measure in wave 1 are treated as predictors. I build the multinominal logistic and logistic regression models for these 21 measure pairs and estimate the coefficients. The estimates for the regression coefficients under MI on the $m=50$ completed datasets are shown in Figure 4.14. Most values for the coefficients are obviously away from 0 and have relatively large absolute values. This demonstrates significant persistency among the outcome measure pairs between wave 1 and wave 2. I compare the persistency on the completed datasets and on the samples who stayed in the completed panel. Figure 4.14 shows that the estimates for the regression coefficients are different between these two data samples.
Figure 4.11: Comparison of estimates and 95% confidence intervals for coefficients in the eight analysis models after MI using joint BLPM between completed panel data and refreshment samples for APYN study: from top left to bottom right, the models in sequence are $W \sim LV31$, $W \sim LV3$, $LV1 \sim X_1$, $LV3 \sim X_1$, $LV31 \sim X_1$, $CND1 \sim X_1$, $INT4-1 \sim X_2$ and $FAV1-4 \sim X_2$. 

133
Based on the multiply imputed datasets and the correspondingly collected attrition $W$, I build a selection model for $W$ with a logit link including the main effects of $X_2$ (age, edu, gender, race, income, marriage status, partyID, ID1, REL3), $Y_1$ (FAV1-1, FAV1-9, LV1, LV3, CND1, and LV31) and $Y_2$ (FAV1-1, FAV1-9, LV1, LV3, CND1, and LV31), two-way and three-way interaction terms among $Y_2$ (FAV1-1 * FAV1-9, LV1 * LV3, LV3 * CND1, LV3 * LV31, CND1 * LV31, CND1 * LV1 and CND1*LV1*LV31 ) variables. The fitting has extremely small p-values for these high-way interactions and illustrates that these high-way interaction terms are significant predictors for $W$. The estimates for the regression coefficients related to these interaction terms have values away from 0, as shown in Figure 4.15. This
Figure 4.13: The marginal probability $Pr(Y_2=1)$, the conditional probability $Pr(Y_2=1|W)$ and the joint probability $Pr(Y_2=1, W)$ estimated from the completed panel and refreshment sample under MI using joint BLPM for APYN study; left 6 cases are for $W=1$ and right 6 cases are for $W=0$. 
Table 4.6: The marginal probability $Pr(Y_2=1)$ and the conditional probability $Pr(Y_2=1 \mid W)$ for all measures in wave 2 after imputation, for completed panel data (p), refreshment sample (r), and all data (all) for the APYN study after MI using joint BLPM without CI assumption for $Y_2$.

|       | Y     | (Y|W=1) | (Y|W=0) |
|-------|-------|--------|--------|
|       | p     | r      | p     | r     | all   | p     | r     | all   |
| MD1   | .228  | .136   | .162  | .156  | .156  | .341  | .156  | .328  |
| MD2   | .192  | .188   | .169  | .177  | .172  | .229  | .140  | .227  |
| FAV1-1| .102  | .063   | .056  | .057  | .057  | .179  | .082  | .172  |
| FAV1-4| .313  | .342   | .289  | .272  | .294  | .353  | .311  | .362  |
| FAV2-1| .199  | .219   | .172  | .159  | .176  | .244  | .133  | .251  |
| FAV2-2| .129  | .111   | .086  | .112  | .091  | .201  | .111  | .194  |
| LV1   | .804  | .892   | .903  | .872  | .906  | .636  | .923  | .646  |
| LV3   | .332  | .356   | .318  | .297  | .322  | .356  | .383  | .360  |
| CND1  | .641  | .726   | .652  | .680  | .666  | .623  | .729  | .629  |
| LV31  | .159  | .106   | .102  | .131  | .098  | .256  | .075  | .254  |
| VOT3  | .529  | .562   | .494  | .482  | .500  | .590  | .488  | .599  |
| INT4-1| .733  | .772   | .775  | .782  | .772  | .660  | .794  | .673  |
| INT4-2| .525  | .537   | .542  | .590  | .540  | .496  | .492  | .502  |
| INT4-3| .566  | .607   | .556  | .553  | .565  | .582  | .569  | .585  |
| INT4-4| .843  | .909   | .887  | .848  | .892  | .770  | .895  | .778  |
| INT4-5| .798  | .835   | .827  | .812  | .829  | .748  | .752  | .754  |
| INT4-6| .798  | .889   | .856  | .872  | .862  | .699  | .829  | .713  |
| INT4-7| .529  | .533   | .516  | .518  | .513  | .552  | .508  | .562  |
| INT4-8| .801  | .856   | .839  | .871  | .842  | .736  | .841  | .746  |
| INT4-9| .767  | .807   | .775  | .778  | .779  | .755  | .806  | .761  |

demonstrates the complicated dependency structure between the attrition $W$ and the outcome measures, especially in $Y_2$.

4.7 Conclusions

Starting from pattern mixture models, I handle the item nonresponse and non-ignorable attrition simultaneously. I introduce latent classes and propose fully Bayesian approaches for imputation with flexible modeling property and efficient computation algorithms. These procedures generate plausible imputation values for the collected variables in the APYN study. When the refreshment samples are available, the com-
Table 4.7: The joint probability $Pr(Y_2=1, W)$ for all measures in wave 2 after imputation, for completed panel data (p), refreshment sample (r), and all data (all) for the APYN study after MI using joint BLPM without CI assumption for $Y_2$ for APYN study.

<table>
<thead>
<tr>
<th></th>
<th>$(Y, W=1)$</th>
<th></th>
<th>$(Y, W=0)$</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>p</td>
<td>r</td>
<td>all</td>
<td>p</td>
</tr>
<tr>
<td>MD1</td>
<td>.102</td>
<td>.103</td>
<td>.102</td>
<td>.126</td>
</tr>
<tr>
<td>MD2</td>
<td>.107</td>
<td>.152</td>
<td>.113</td>
<td>.085</td>
</tr>
<tr>
<td>FAV1-1</td>
<td>.035</td>
<td>.049</td>
<td>.037</td>
<td>.066</td>
</tr>
<tr>
<td>FAV1-4</td>
<td>.182</td>
<td>.255</td>
<td>.193</td>
<td>.130</td>
</tr>
<tr>
<td>FAV1-9</td>
<td>.115</td>
<td>.120</td>
<td>.116</td>
<td>.095</td>
</tr>
<tr>
<td>FAV2-1</td>
<td>.109</td>
<td>.157</td>
<td>.116</td>
<td>.090</td>
</tr>
<tr>
<td>FAV2-2</td>
<td>.054</td>
<td>.091</td>
<td>.060</td>
<td>.074</td>
</tr>
<tr>
<td>LV1</td>
<td>.569</td>
<td>.752</td>
<td>.596</td>
<td>.235</td>
</tr>
<tr>
<td>LV3</td>
<td>.200</td>
<td>.280</td>
<td>.212</td>
<td>.132</td>
</tr>
<tr>
<td>CND1</td>
<td>.411</td>
<td>.596</td>
<td>.438</td>
<td>.230</td>
</tr>
<tr>
<td>LV31</td>
<td>.065</td>
<td>.064</td>
<td>.064</td>
<td>.095</td>
</tr>
<tr>
<td>VOT3</td>
<td>.311</td>
<td>.434</td>
<td>.329</td>
<td>.218</td>
</tr>
<tr>
<td>INT4-1</td>
<td>.489</td>
<td>.619</td>
<td>.508</td>
<td>.244</td>
</tr>
<tr>
<td>INT4-2</td>
<td>.342</td>
<td>.431</td>
<td>.355</td>
<td>.183</td>
</tr>
<tr>
<td>INT4-3</td>
<td>.351</td>
<td>.493</td>
<td>.371</td>
<td>.215</td>
</tr>
<tr>
<td>INT4-4</td>
<td>.559</td>
<td>.748</td>
<td>.586</td>
<td>.285</td>
</tr>
<tr>
<td>INT4-5</td>
<td>.521</td>
<td>.683</td>
<td>.545</td>
<td>.277</td>
</tr>
<tr>
<td>INT4-6</td>
<td>.539</td>
<td>.727</td>
<td>.567</td>
<td>.258</td>
</tr>
<tr>
<td>INT4-7</td>
<td>.325</td>
<td>.407</td>
<td>.337</td>
<td>.204</td>
</tr>
<tr>
<td>INT4-8</td>
<td>.529</td>
<td>.698</td>
<td>.553</td>
<td>.272</td>
</tr>
<tr>
<td>INT4-9</td>
<td>.488</td>
<td>.653</td>
<td>.512</td>
<td>.279</td>
</tr>
</tbody>
</table>

The common conditional independence assumption of $Y_2$ and $W$ given latent class can be checked. The refreshment samples can calibrate the inference if the conditional independence assumption is violated. If $Y_2$ still depends on $W$ given latent class, the dependent and joint Bayesian latent pattern mixture models releasing the conditional independence assumption can be identified with refreshment samples. However, the benefits of refreshment samples are based on the assumption of the same structure for the panel data and refreshment samples. The refreshment samples should come from the same population as the completed panel data, as a random representative.
Figure 4.14: Persistency check between wave 1 and wave 2 measures: estimates and 95% confidence intervals of the regression coefficients based on completed datasets after MI where $Y_2$ are outcomes and $Y_1$ are predictors in APYN study: comparison between the multiple imputed all samples and samples stayed in panel.
Figure 4.15: Estimates and 95% confidence intervals for regression coefficients in the constructed selection model for \( W \) with high-way interaction terms used for analysis based on completed datasets after MI using joint BLPM in APYN study: from left to right are for main effects of \( X \), \( Y_1 \) and \( Y_2 \), two-way interactions among \( Y_2 \) and three-way interaction inside \( Y_2 \).
The work in this thesis can be extended for several different directions. The first generalization is to handle mixed type of data. In this thesis, I focus on unordered categorical data of high dimension. It would be useful to extend the latent class models in Chapter 2 for mix of data types, such as continuous and ordered categorical variables. The conditional independence assumption will facilitate this extension since independent distributions can be fit on the corresponding type of variables. Particularly, for the imputation of background variables in Chapter 2, I will include the students’ proficiency score for imputation, which is usually assumed as a continuous outcome variable. Psychometricians are interested in the relationship between the student proficiency and the background information to make better decisions about the assessment. Being able to deal with missing data for various types of variables, the latent class analysis can become a general and flexible imputation engine.

Another meaningful extension is to evaluate the performance of shrinkage prior distributions on the regression coefficients for the selection model in Chapter 3. When the goal is imputation, how to balance congeniality and sparsity of the model under multiple imputation is a crucial issue. In order for better mixing behavior of the
MCMC chain and better control of multiple co-linearity or possible complete and quasi-complete separation problems, shrinkage prior distributions are useful. The prediction error under shrinkage should be compared to the case without shrinkage effect. Cross validation check is under my consideration here.

One fundamental ongoing work is to provide theoretical support for the identification of the Bayesian latent pattern mixture models. When refreshment samples are available, the Bayesian latent pattern mixture models under the conditional independence assumption are over-identified and the assumption can be released. However, only one part of the two wave outcome variables can be supposed as dependent on the attrition given the latent classes. I assume $Y_2$ still depends on $W$ in Chapter 4. The preliminary theoretical proof shows that this model can be identified, and I need more elegant theories based on latent class structure.

Currently I consider a two wave panel with one refreshment sample. I would like to extend the latent class models for all the 11 wave and 3 refreshment samples for the APYN study and fully utilize the panel data across time. This becomes more demanding in terms of the identification and dimension reduction problems. For three wave panel studies with two refreshment samples, theoretical work has been done on the identification issue for selection models for different missingness patterns and different follow up behaviors for the refreshment samples. These mechanisms include drop out and intermittent missingness. The refreshment samples can be followed up in later waves or as an isolated sample in one wave. I would like to extend the Bayesian latent pattern mixture models for the more complicated scenario and handle the high dimensional data and missing data mechanism jointly and flexibly.

Another interesting extension is to apply these models for data confidentiality protection. The government and statistical organizations emphasize confidentiality protection for data disclosure. The methods of replacing the sensitive information are essentially generalizations of missing data methodology. I would like to adjust and
implement the latent class models for categorical data with confidentiality. When the true data are available, we have more information for imputation modeling and can construct more plausible replacement. However, the inference demands adjustments with carefulness, since typically the data released for analysis are only part of those for imputation. Furthermore, multiple imputation can handle the synthetic confidential data and missing data simultaneously. Handling incomplete and confidential data of high dimension simultaneously is a useful and attractive research direction.
Appendix A

Appendix to Chapter 2 - Hierarchical DP Mixture of Products of Multinomial Distributions

A.0.1 Model Specification

In addition to student background variables, the educational surveys also collect background information for teachers and schools. Using this related information motivates construction of multilevel models. To accommodate the complex population structure of the dataset, a hierarchical version of the latent class analysis could be applied. As to international education surveys, it is generally believed that students from different countries have systematic differences on background information.

Take different countries as hyper-groups, and suppose latent classes can differ in mixing proportions across groups. The mixing component proportions $\pi$ are dependent on countries. However, the component parameters of all variables are the same across countries. They only differ across the allocation probabilities of latent classes. The hierarchical latent class analysis with a pre-determined number of classes is a special case of the hierarchical general diagnostic model (Von Davier, 2010).

Using a Bayesian nonparametric prior, I would like to keep the flexible and
sparse property of the DPMPM and take the hierarchical structure difference into account. I propose a hierarchical Dirichlet process mixture of products of multinomial (HDPMPM) models to impute missing values in groups of categorical variables. The basic framework is that we collect groups of data, where each individual within a group is a realization from DPMPM distributions. The mixture components $\Phi$ are shared between groups and the allocation probabilities $\pi$ are different across groups.

The Hierarchical Dirichlet process (HDP) (Teh et al., 2006) prior is specified as $(P_g|\alpha, P_0) \sim DP(\alpha, P_0)$ and $(P_0|\gamma, P_{00}) \sim DP(\gamma, P_{00})$, for each group index $g$. $P_{00}$ is the base measure for the component parameters. One nice property of HDP is that mixture models in the different groups necessarily share mixture components. Each individual within one group is a realization from a mixture model, while between groups individuals share common mixture components. In this case, within each country, the observation for each student is a draw from a mixture distribution. The mixture components are shared among different countries. A simple interpretation is from the Chinese restaurant process perspective; we have several restaurants, and the tables inside these different restaurants share common dishes.

Denote $X_{gij}$ as the category of variable $j$ for individual $i$ from group $g$, for $g = 1, \ldots, G$, $i = 1, \ldots, n_g$ and $j = 1, \ldots, p$, where $n_g$ is the number of individuals in Group $g$. Let $\{s_{gi}\}$ be the latent class indicator. Let $\pi_g = \{\pi_{g1}, \ldots, \pi_{g\infty}\}$ be defined as the allocation probabilities to latent classes in Group $g$, $\pi_{gh} = P r(s_{gi} = h)$. Let $\gamma = \{\gamma_1, \ldots, \gamma_{\infty}\}$ denote the global weights or allocation probabilities to latent classes. I apply the stick-breaking representation of DP to construct the HDPMPM
model as follows:

\[
X_{gij} \sim \text{Multinomial}\{1, \ldots, d_j\}, \psi_{s_{gi1}}, \ldots, \psi_{s_{gid_j}}
\]  
\[\text{(A.1)}\]

\[
s_{gi} \sim \text{Multinomial}\{1, \ldots, \infty\}, \pi_g = \{\pi_{g1}, \ldots, \pi_{g\infty}\}
\]  
\[\text{(A.2)}\]

\[
\pi_g \sim \text{Dirichlet}(\alpha_2, \gamma = \{\gamma_1, \ldots, \gamma_\infty\})
\]  
\[\text{(A.3)}\]

\[
\gamma_h = V_h \prod_{l < h} (1 - V_l)
\]  
\[\text{(A.4)}\]

\[
V_h \sim \text{Beta}(1, \alpha_1)
\]  
\[\text{(A.5)}\]

\[
\psi_{h}^{(j)} \sim \text{Dirichlet}(a_{j1}, \ldots, a_{jc_j}).
\]  
\[\text{(A.6)}\]

Equivalently, the HDPMPM model can be represented as

\[
X_{gij} \sim \text{Multinomial}\{1, \ldots, d_j\}, \psi_{s_{gi1}}, \ldots, \psi_{s_{gid_j}}
\]  
\[\text{(A.7)}\]

\[
s_{gi} \sim \text{Multinomial}\{1, \ldots, \infty\}, \pi_g = \{\pi_{g1}, \ldots, \pi_{g\infty}\}
\]  
\[\text{(A.8)}\]

\[
\pi_{gh} = \pi_{gh}' \prod_{l < h} (1 - \pi_{gl}')
\]  
\[\text{(A.9)}\]

\[
\pi_{gh}' \sim \text{Beta}(\alpha_2 \gamma_h, \alpha_2 (1 - \sum_{i=1}^{h} \gamma_i))
\]  
\[\text{(A.10)}\]

\[
\gamma_h = V_h \prod_{l < h} (1 - V_l)
\]  
\[\text{(A.11)}\]

\[
V_h \sim \text{Beta}(1, \alpha_1)
\]  
\[\text{(A.12)}\]

\[
\psi_{h}^{(j)} \sim \text{Dirichlet}(a_{j1}, \ldots, a_{jc_j}).
\]  
\[\text{(A.13)}\]

Given \(\Psi, \pi\), the joint likelihood is

\[
\prod_{g=1}^{G} \prod_{i=1}^{n_g} \left\{ \sum_{h=1}^{\infty} \prod_{j=1}^{p} \prod_{l=1}^{d_j} (\psi_{h}^{(j)}) f(X_{gij} = l) \right\}.
\]  
\[\text{(A.14)}\]

### A.0.2 Posterior Computation for HDPMPM

Truncating the stick-breaking process for \(\gamma\) at \(H^*\), I propose a blocked sampler algorithm for the posterior computation. As a Chinese restaurant franchise, let \(n_{gth}\)
be the number of customers in restaurant $g$ at table; let $m_{gh}$ be the number of tables in restaurant $g$ taking dish $h$. Denote $n_{g-h} = \sum_{t=1}^{m_{gh}} n_{gth}$ and $m_h = \sum_{g=1}^{G} m_{gh}$. The steps for posterior computations are specified as following:

**Step 1:** Update $s_{gi} \in \{1, \ldots, H^*\}$ from multinomial distribution with probabilities:

$$Pr(s_{gi} = h|\cdot) = \frac{\pi_{gh} \prod_{j=1}^{p} \psi_{hX_{gij}}^{(j)}}{\sum_{k=1}^{H^*} \pi_{gk} \prod_{j=1}^{p} \psi_{kX_{gij}}^{(j)}},$$

where $X_{gij} \in X_{com}$ and $X_{com} = \{X_{obs}, X_{mis}\}$, for $g = 1, \ldots, G$, $i = 1, \ldots, n_g$ and $j = 1, \ldots, p$.

**Step 2:** Update $V_h$ from conjugate Beta distributions, for $h = 1, \ldots, H^*$:

$$(V_h|\cdot) \sim \text{Beta}(1 + m_h, \alpha_1 + \sum_{k=h+1}^{H^*} m_k),$$

then directly calculate $\gamma_h$ from $\gamma_h = V_h \prod_{k<h}(1 - V_k)$.

**Step 3:** Update $\pi'_{gh}$ from conjugate Beta distributions, for $h = 1, \ldots, H^*$ and $g = 1, \ldots, G$:

$$(\pi_{gh}'|\cdot) \sim \text{Beta}(\alpha_2 \gamma_h + n_{g-h}, \alpha_2(1 - \sum_{k=1}^{h} \gamma_k) + \sum_{k=h+1}^{H^*} n_{g-k}),$$

then directly calculate $\pi_{gh}$ from $\pi_{gh} = \pi'_{gh} \prod_{k<h}(1 - \pi'_{gk})$.

**Step 4:** Update $\psi_{h}^{(j)} = (\psi_{h1}^{(j)}, \ldots, \psi_{hd_{j}}^{(j)})$ from conjugate Dirichlet distributions:

$$(\psi_{h}^{(j)}|\cdot) \sim \text{Dirichlet}(a_{j1} + \sum_{g:i=s_{gi}=h} I(X_{gij} = 1), \ldots, a_{jd_{j}} + \sum_{g:i=s_{gi}=h} I(X_{gij} = d_{j})).$$

**Step 5:** Update $m_{gh}$

$$Pr(m_{gh} = m|\cdot) = \frac{\Gamma(\alpha_2 \gamma_h)}{\Gamma(\alpha_2 \gamma_h + n_{g-h})} S(n_{g-h}, m) (\alpha_2 \gamma_h)^m.$$
where \( S(n+1, m) = S(n, m-1) + nS(n, m) \), \( S(0, 0) = S(1, 1) = 1 \), \( S(n, 0) = 0 \) for \( n > 0 \) and \( S(n, m) = 0 \) for \( m > n \).

**Step 6:** Update \( X_{mis} \) from conjugate multinomial distributions. Draw new values for those \( X_{ij} \)’s with \( (i, j) \in \{(i, j) : r_{ij} = 0\} \),

\[
(X_{gij}|\cdot) \sim \text{multinomial}(\{1, \ldots, d_j\}, \psi_{sj1}^{(j)}, \ldots, \psi_{sijd_j}^{(j)}).
\]
Appendix B

Appendix to Chapter 4 - Proof for Theorem 1

Rewrite the first four conditions

\[ \beta_0^{**} + \beta_1^{**} + \beta_2^{**} = f^{-1}(q_{11}) \quad (B.1) \]
\[ \beta_0^{**} + \beta_2^{**} = f^{-1}(\hat{q}_{10}) \quad (B.2) \]
\[ \beta_0^{**} + \beta_1^{**} = f^{-1}(q_{01}) \quad (B.3) \]
\[ \beta_0^{**} = f^{-1}(q_{00}). \quad (B.4) \]

Eliminate \((\beta_0^{**}, \beta_1^{**}, \beta_2^{**})\) and we have a restriction

\[ h(\hat{q}_{10}, \hat{q}_{00}) = f^{-1}(q_{01}) + f^{-1}(q_{10}) - f^{-1}(q_{00}) - f^{-1}(q_{11}) = 0. \quad (B.5) \]

Because of continuity of \(h(\cdot, \cdot)\) and because \(h(\cdot, \cdot)\) is increasing in \(q_{10}\) and decreasing in \(q_{00}\), and we have the fixed condition \(h(\hat{q}_{10}, \hat{q}_{00}) = 0\), if we let \(q_{00}\) increases, then \(q_{10}\) will decrease. This restriction defines an implicit function \(q_{10} = \bar{Q}_{10}(q_{00})\) with the following properties

\[ \frac{\partial \bar{Q}_{10}}{\partial x}(x) > 0 \quad (B.6) \]
\[ \lim_{x\to 0} \bar{Q}_{10}(x) = 0 \quad (B.7) \]
\[ \lim_{x\to 1} \bar{Q}_{10}(x) = 1. \quad (B.8) \]
The restriction (4.11) defines a function

\[ \tilde{Q}_{10}(X) = \frac{q_{00}r_{00} + q_{10}r_{10} - xr_{00}}{r_{10}}, \]  

(B.9)

with properties

\[ \frac{\partial \tilde{Q}_{10}}{\partial x}(x) < 0 \]  

(B.10)

\[ \tilde{Q}_{10}(0) > 0 \]  

(B.11)

\[ \tilde{Q}_{10}(1) < 1. \]  

(B.12)

Hence, there is a unique value \( \hat{q}_{00} \) solving \( \tilde{Q}_{10}(x) = \tilde{Q}_{10}(x) \) and \( \hat{q}_{10} = \tilde{Q}_{10}(\hat{q}_{00}) \). The proposed model (4.10) in Chapter 4 can be identified. Similar to Hirano et al. (2001), the model and theories can extend to joint distributions of any data types.
Appendix C

Appendix to Chapter 4 - Posterior Computation
Algorithm for Dependent BLPM under CI
Assumption

**Step 1:** Update \( s_i \in \{1, \ldots, H^s\} \) from multinomial distribution with probabilities

\[
\Pr(s_i = h | -) = \frac{\pi_h(w_i) \prod_{j=1}^{p} \psi(j)_{hZ_{ij}}}{\sum_{k=1}^{H^s} \pi_k(w_i) \prod_{j=1}^{p} \psi(j)_{kZ_{ij}}},
\]

where \( Z_{ij} \in Z_{\text{com}} = \{X_{\text{mis}}, X_{\text{obs}}, Y_{1\text{mis}}, Y_{1\text{obs}}, Y_{2\text{obs}}, Y_{2\text{mis}}\} \), for \( i=1, \ldots, N \) and \( j=1, \ldots, p \).

**Step 2:** Update \( t_{ih} \), where \( TN_{[a,b]}(\mu, \sigma^2) \) represents truncated normal distribution with mean \( \mu \) and variance \( \sigma^2 \) at the interval \((a, b)\).

- if \( s_i = 1 \), \( (t_{i1} | -) \sim TN_{[0, \infty]}(X_{1i}^{s'}, \beta, 1) \);
- if \( s_i = h \) and \( 2 \leq h \leq H^s - 1 \), draw \( \{t_{ik}\}_{k=1}^{h-1} \) independently and identically from \( TN_{(-\infty, 0]}(X_{ik}^{s'}, \beta, 1) \) and \( t_{ih} \) from \( TN_{[0, \infty]}(X_{ih}^{s'}, \beta, 1) \);
- if \( s_i = H^s \), generate \( \{t_{ik}\}_{k=1}^{H^s-1} \) independently from \( TN_{(-\infty, 0]}(X_{ik}^{s'}, \beta, 1) \).
Step 3: Update $\beta=(\beta_{01},\ldots,\beta_{0H^*+1},\beta_1)'$ under prior $N_{H^*}(0,\Sigma_0)$. Let $\tilde{t}_i=(t_{i1},\ldots,t_{ih_i})$ be the latent variables corresponding to $s_i=h$ where $h_i=\min(h,H^*-1)$ and $M^*_i$ denote the covariate matrix for individual $i$ consisting $h_i$ rows $X^*_{ih},\ldots,X^*_{ih}$. The length of $X^*_{ih}$ is $H^*$, and then the $M^*_i$ is a $h_i \times H^*$ covariate matrix.

$$(\beta-) \sim MVN^*_H\left(\hat{\beta},(\Sigma_0^{-1} + \sum_{i=1}^{N} M^*_i M^*_i)^{-1}\right),$$

where $\hat{\beta}=(\Sigma_0^{-1} + \sum_{i=1}^{N} M^*_i M^*_i)^{-1} \sum_{i=1}^{N} M^*_i \tilde{t}_i$. Hence, for $2 \leq h \leq H^*-1$, $\pi_h(w_i)=\Phi(X^*_{ih}\beta) \prod_{i<h} (1-\Phi(X^*_{ih}\beta))$, and $\pi_1(w_i)=\Phi(X^*_1\beta)$. Then $\pi_{H^*}(w_i)=1-\sum_{k=1}^{H^*-1} \pi_k(w_i)$.

Step 4: Update $\psi^{(j)}_h=(\psi^{(j)}_{h1},\ldots,\psi^{(j)}_{hd_j})$ from conjugate Dirichlet distributions:

$$(\psi^{(j)}_h-) \sim \text{Dirichlet}(a_{j1} + \sum_{i:s_i=h} I(Z_{ij}=1),\ldots,a_{jd_j} + \sum_{i:s_i=h} I(Z_{ij}=d_j)).$$

Step 5: Update $\gamma_0$ from conjugate Beta distributions:

$$(\gamma_0-) \sim \text{Beta}(a_{\gamma_0} + \sum W_i, b_{\gamma_0} + N - \sum W_i).$$

Step 6: Update $Z_{mis}$ in the panel and refreshment samples. For $\{(i,j) : Z_{ij} \in Z_{mis}\}$,

$$(Z_{i,j}-) \sim \text{Multinomial}\{\{1,\ldots, d_j\}, \psi^{(j)}_{s_{i1}},\ldots,\psi^{(j)}_{s_{id_j}}\}.$$.

Step 7: Update $W_{mis}$ in refreshment samples: $S$ depends on $W$

$$(W_i-) \sim \text{Bernoulli}(Pr(W_i=1|S,\beta,\gamma)),$$

where $Pr(W_i=1|S,\beta,\gamma) = \frac{Pr(S|W_i=1,\beta)Pr(W_i=1|\gamma)}{\sum_{w_i\in\{0,1\}}Pr(S|W_i=w,\beta)Pr(W_i=w|\gamma)}$. Based on new $W_{mis}$, recall to update $X^*_{ih}=(0,0,\ldots,1,0,0,W_i)$ for $h=1,\ldots,H^*$. 

151
Appendix D

Appendix to Chapter 4 - Posterior Computation
Algorithm for Joint BLPM under CI Assumption

Step 1: Update $s_i \in \{1, \ldots, H^*\}$ from multinomial distribution with probabilities

$$
\Pr(s_i = h|\cdot) = \frac{\pi_h \prod_{j=1}^{p} \psi_{hZ_{ij}}^{(j)} \psi_{hW_i}^{(w)}}{\sum_{k=1}^{H^*} \pi_k \prod_{j=1}^{p} \psi_{kZ_{ij}}^{(j)} \psi_{kW_i}^{(w)}},
$$

where $Z_{ij} \in Z_{com} = \{X_{mis}, X_{obs}, Y_{1mis}, Y_{1obs}, Y_{2obs}, Y_{2mis}\}$, for $i=1, \ldots, N$ and $j=1, \ldots, p$.

Step 2: Update $V_h$ from conjugate Beta distributions, for $h = 1, \ldots, H^*$

$$(V_h|\cdot) \sim \text{Beta}(1 + n_h, \alpha + \sum_{k=h+1}^{H^*} n_k),$$

where $n_h = \sum_{i=1}^{N} I(s_i = h)$, and calculate $\pi_h$ from $\pi_h = V_h \prod_{k<h} (1 - V_k)$.

Step 3a: Update $\psi_{h}^{(j)} = (\psi_{h1}^{(j)}, \ldots, \psi_{hd_j}^{(j)})$ from conjugate Dirichlet distributions

$$(\psi_{h}^{(j)}|\cdot) \sim \text{Dirichlet}(a_{j1} + \sum_{i:s_i=h} I(Z_{ij} = 1), \ldots, a_{jd_j} + \sum_{i:s_i=h} I(Z_{ij} = d_j)).$$
Step 3b: Update \( \psi_{h_1}^{(w)} \) from conjugate Beta distributions

\[
(\psi_{h_1}^{(w)} | -) \sim \text{Beta}(a_{w_1} + \sum_{i:s_i=h} I(W_i = 1), a_{w_0} + \sum_{i:s_i=h} I(W_i = 1)).
\]

Step 4: Update \( \alpha \) from conjugate Gamma distributions

\[
(\alpha | -) \sim \text{Gamma}(a_{\alpha} + H^* - 1, b_{\alpha} - \log \pi_{H^*}).
\]

Step 5: Update \( Z_{mis} \) in the panel and refreshment samples. For \( \{ (i, j) : Z_{i,j} \in Z_{mis} \} \)

\[
(Z_{i,j} | -) \sim \text{Multinomial}(|1, \ldots, d_j}, \psi_{s_1}, \ldots, \psi_{s_{d_j}}\).
\]

Step 6: Update \( W_{mis} \) in refreshment samples

\[
(W_i | -) \sim \text{Bernoulli}(\psi_{s_1}^{(w)}).
\]
Appendix E

Appendix to Chapter 4 - Posterior Computation
Algorithm for Dependent BLPM

Step 1: Update \( s_i \in \{1, \ldots, H^s\} \) from multinomial distribution with probabilities

\[
\text{Pr}(s_i = h | \cdot) = \frac{\pi_h(w_i) \prod_{j=1}^{p} \psi_{hZ_{ij}}^{(w_{ij})} \psi_{hZ_{ij}}^{(j)}}{\sum_{k=1}^{H^s} \pi_k(w_i) \prod_{j=1}^{p} \psi_{kZ_{ij}}^{(w_{ij})} \psi_{kZ_{ij}}^{(j)}},
\]

where \( Z_{ij} \in Z_{com} = \{X_{mis}, X_{obs}, Y_{1mis}, Y_{1obs}, Y_{2obs}, Y_{2mis}\} \), for \( i=1, \ldots, N \) and \( j=1, \ldots, p \).

Step 2: Update \( t_{ih} \), where \( TN_{(a,b)}(\mu, \sigma^2) \) represents truncated normal distribution with mean \( \mu \) and variance \( \sigma^2 \) at the interval \((a, b)\).

- if \( s_i = 1 \), \( (t_{i1}|\cdot) \sim TN_{[0,\infty)}(X_{1i}^w, \beta, 1) \);
- if \( s_i = h \) and \( 2 \leq h \leq H^s - 1 \), draw \( \{t_{ik}\}_{k=1}^{h-1} \) independently and identically from \( TN_{(-\infty,0)}(X_{ik}^w, \beta, 1) \) and \( t_{ih} \) from \( TN_{[0,\infty)}(X_{ih}^w, \beta, 1) \);
- if \( s_i = H^s \), generate \( \{t_{ik}\}_{k=1}^{H^s-1} \) independently from \( TN_{(-\infty,0)}(X_{ik}^w, \beta, 1) \).
Step 3: Update $\beta=(\beta_{01}, \ldots, \beta_{0h^*-1}, \beta_1)'$ under prior $N_{H^*}(0, \Sigma_0)$. Let $\tilde{t}_i=(t_{i1}, \ldots, t_{ih_i})$ be the latent variables corresponding to $s_i=h$ where $h_i=\min(h, H^*-1)$ and $M_i^*$ denote the covariate matrix for individual $i$ consisting $h_i$ rows $X_{i1}^{s'_i}, \ldots, X_{ih_i}^{s'_i}$. The length of $X_{ih_i}^{s'_i}$ is $H^*$, and then the $M_i^*$ is a $h_i \times H^*$ covariate matrix.

$$\begin{align*}
(\beta|\cdot) & \sim MVN_{H^*}^\pi(\hat{\beta}, (\Sigma_0^{-1} + \sum_{i=1}^{N} M_i^* M_i^*)^{-1}),
\end{align*}$$

where $\hat{\beta}=(\Sigma_0^{-1} + \sum_{i=1}^{N} M_i^* M_i^*)^{-1} \sum_{i=1}^{N} M_i^* \tilde{t}_i$. Hence, for $2 \leq h \leq H^*-1$, $\pi_h(w_i)=\Phi(X_{ih_i}^{s'_i} \beta) \prod_{l<h}(1-\Phi(X_{il}^{s'_i} \beta))$, and $\pi_1(w_i)=\Phi(X_{i1}^{s'_i} \beta)$. Then $\pi_{H^*}(w_i)=1-\sum_{k=1}^{H^*-1} \pi_k(w_i)$.

Step 4a: Update $\psi_{h}^{(1j)}=(\psi_{h1}^{(1j)}, \ldots, \psi_{hdj}^{(1j)})$ from conjugate Dirichlet distributions

$$\begin{align*}
(\psi_{h}^{(1j)}|\cdot) & \sim \text{Dirichlet}(a_{j1} + \sum_{i:s_i=h, w_i=1} I(Y_{i2j}=1), \ldots, a_{jdj} + \sum_{i:s_i=h, w_i=1} I(Y_{i2j}=d_j)).
\end{align*}$$

Step 4b: Update $\psi_{h}^{(0j)}=(\psi_{h1}^{(0j)}, \ldots, \psi_{hdj}^{(0j)})$ from conjugate Dirichlet distributions

$$\begin{align*}
(\psi_{h}^{(0j)}|\cdot) & \sim \text{Dirichlet}(b_{j1} + \sum_{i:s_i=h, w_i=0} I(Y_{i2j}=1), \ldots, b_{jdj} + \sum_{i:s_i=h, w_i=0} I(Y_{i2j}=d_j)).
\end{align*}$$

Step 4c: Update $\psi_{h}^{(j)}$ from conjugate Dirichlet distributions

$$\begin{align*}
(\psi_{h}^{(j)}|\cdot) & \sim \text{Dirichlet}(b_{j1} + \sum_{i:s_i=h, \tilde{Z}_{ij}=1} I(\tilde{Z}_{ij}=1), \ldots, b_{jdj} + \sum_{i:s_i=h} I(\tilde{Z}_{ij}=d_j)),
\end{align*}$$

where $\tilde{Z}_{ij}=(X_{ij}, Y_{1ij})$ include variables of $X$ and $Y_1$ and do not include the variables of $Y_2$.

Step 5: Update $\gamma_0$ from conjugate Beta distributions

$$\begin{align*}
(\gamma_0|\cdot) & \sim \text{Beta}(a_{\gamma_0} + \sum W_i, b_{\gamma_0} + N - \sum W_i).
\end{align*}$$
**Step 6a:** Update $Y_{2\text{mis}}$ in the panel and refreshment samples. For $\{(i, j) : Y_{i,j} \in Y_{2\text{mis}}\}$

$$(Y_{i,j}|-) \sim \text{Multinomial}\{\{1, \ldots, d_j\}, \psi_{s_i1}^{(w_j)}, \ldots, \psi_{s_i d_j}^{(w_j)}\}.$$  

**Step 6b:** Update $Y_{1\text{mis}}$ in the panel and refreshment samples. For $\{(i, j) : Y_{i,j} \in Y_{1\text{mis}}\}$

$$(Y_{i,j}|-) \sim \text{Multinomial}\{\{1, \ldots, d_j\}, \psi_{s_i1}^{(j)}, \ldots, \psi_{s_i d_j}^{(j)}\}.$$  

**Step 7:** Update $W_{\text{mis}}$ in refreshment samples: both $S$ and $Y_2$ depend on $W$

$$(W_i|-) \sim \text{Bernoulli}(Pr(W_i=1|Y_{i,2-}, S_i, \Psi, \beta, \gamma)),$$

where $Pr(W_i=1|S_i, \beta, \gamma) = \frac{f(Y_{i,2-}, S, W_i=1, \Psi)|Pr(W_i=1|\gamma)}{\sum_{w \in \{0,1\}} f(S, W_i=w, \Psi)|Pr(W_i=w|\gamma)}$. Based on new $W_{\text{mis}}$, recall to update $X_{ih}^{*}= (0, 0, \ldots, 1, 0, 0, W_i)$ for $h=1, \ldots, H^*$.  

156
Appendix F

Appendix to Chapter 4 - Posterior Computation
Algorithm for Joint BLPM

Step 1: Update \( s_i \in \{1, \ldots, H\} \) from multinomial distribution with probabilities

\[
Pr(s_i=h|--) = \frac{\pi_h \prod_{j=1}^{p} \psi_{hZ_{ij}}^{(w)} \psi_{hW_i}^{(u)}}{\sum_{k=1}^{H} \pi_k \prod_{j=1}^{p} \psi_{kZ_{ij}}^{(w)} \psi_{kW_i}^{(u)}},
\]

where \( Z_{ij} \in Z_{com} = \{X_{mis}, X_{obs}, Y_{1mis}, Y_{1obs}, Y_{2obs}, Y_{2mis}\} \), for \( i=1, \ldots, N \) and \( j=1, \ldots, p \).

Step 2: Update \( V_h \) from conjugate Beta distributions, for \( h=1, \ldots, H \)

\[
(V_h|--) \sim \text{Beta}(1 + n_h, \alpha + \sum_{k=h+1}^{H} n_k),
\]

where \( n_h = \sum_{i=1}^{N} I(s_i=h) \), and calculate \( \pi_h \) from \( \pi_h = V_h \prod_{k<h} (1 - V_k) \).

Step 3a: Update \( \psi_h^{(1j)}=(\psi_{h1}^{(1j)}, \ldots, \psi_{hd_j}^{(1j)}) \) from conjugate Dirichlet distributions

\[
(\psi_h^{(1j)}|--) \sim \text{Dirichlet}(a_{j1} + \sum_{i:s_i=h,w_i=1} I(Y_{i2j}=1), \ldots, a_{jd_j} + \sum_{i:s_i=h,w_i=1} I(Y_{i2j}=d_j)).
\]
Step 3b: Update $\psi_h^{(0j)} = (\psi_h^{(0j)}, \ldots, \psi_h^{(0j)})$ from conjugate Dirichlet distributions

$$(\psi_h^{(0j)}|\cdot) \sim \text{Dirichlet}(b_j + \sum_{i:s_i=h, w_i=0} I(Y_{i2j}=1), \ldots, b_{jd_j} + \sum_{i:s_i=h, w_i=0} I(Y_{i2j}=d_j)).$$

Step 3c: Update $\psi_h^{(j)}$ from conjugate Dirichlet distributions

$$(\psi_h^{(j)}|\cdot) \sim \text{Dirichlet}(b_j + \sum_{i:s_i=h} I(\tilde{Z}_{ij}=1), \ldots, b_{jd_j} + \sum_{i:s_i=h} I(\tilde{Z}_{ij}=d_j)), $$

here $\tilde{Z}_{ij}=(X_{ij}, Y_{1ij})$ includes $X$ and $Y_1$ but no $Y_2$.

Step 3d: Update $\psi_{h1}^{(w)}$ from conjugate Beta distributions

$$(\psi_{h1}^{(w)}|\cdot) \sim \text{Beta}(a_{w1} + \sum_{i:s_i=h} I(W_i=1), a_{w0} + \sum_{i:s_i=h} I(W_i=0)).$$

Step 4: Update $\alpha$ from conjugate Gamma distributions

$$(\alpha|\cdot) \sim \Gamma(a_\alpha + H - 1, b_\alpha - \log \pi_H).$$

Step 5a: Update $Y_{2\text{mis}}$ in the panel and refreshment samples. For $\{(i, j) : Y_{i,j} \in Y_{2\text{mis}}\}$

$$(Y_{i,j}|\cdot) \sim \text{Multinomial}\{\{1, \ldots, d_j\}, \psi_{s_{i1}}^{(w_{ij})}, \ldots, \psi_{s_{jd_j}}^{(w_{ij})}\}.$$

Step 5b: Update $Y_{1\text{mis}}$ in the panel and refreshment samples. For $\{(i, j) : Y_{i,j} \in Y_{1\text{mis}}\}$

$$(Y_{i,j}|\cdot) \sim \text{Multinomial}\{\{1, \ldots, d_j\}, \psi_{s_{i1}}^{(j)}, \ldots, \psi_{s_{jd_j}}^{(j)}\}.$$ 

Step 6: Update $W_{\text{mis}}$ in refreshment samples: both $S$ and $Y_2$ depend on $W$

$$(W_i|\cdot) \sim \text{Bernoulli}(\Pr(W_i=1|Y_{i,2-}, S_i, \Psi)), $$

where

$$Pr(W_i=1|Z_i, S_i, \Psi) = \frac{\sum_{w \in \{0, 1\}} f(Y_{i,2-}|S_i, W_i=1, \Psi) Pr(W_i=1|S_i, \Psi) f(S_i|\alpha)}{\sum_{w \in \{0, 1\}} f(Y_{i,2-}|S_i, W_i=w, \Psi) Pr(W_i=w|S_i, \Psi) f(S_i|\alpha)}$$

$$= \frac{f(Y_{i,2-}|S_i, W_i=1, \Psi) Pr(W_i=1|S_i, \Psi)}{\sum_{w \in \{0, 1\}} f(Y_{i,2-}|S_i, W_i=w, \Psi) Pr(W_i=w|S_i, \Psi).}$$

158
Bibliography


Formann, A. K. (2007), “Mixture analysis of multivariate categorical data with co-
variates and missing entries,” *Computational Statistics and Data Analysis*, 51,
5236–5246.

Tech. rep., Chestnut Hill, MA: TIMSS & PIRLS International Study Center,
Boston College.

spatial modeling with Dirichlet process mixing,” *Journal of the American Statis-
tical Association*, 100, 1021–1035.

Gelman, A. (2007), “Struggles with survey weighting and regression modeling,” *Sta-

“Multiple imputation for model checking: Completed-Data plots with missing and
latent data,” *Biometrics*, 61, 74–85.

default prior distribution for logistic and other regression models,” *The Annals of

models,” *Journal of Mathematical Psychology*, 56, 1–12.


mixture modeling with nonignorable nonresponse,” in *Drawing Inferences from

for multivariate data with small sample size,” in *Statistical Strategies for Small


large-scale complex survey: A guide,” *Statistical Methods in Medical Research*, 19,
653–670.


Reiter, J. P. (2008), “Multiple imputation when records used for imputation are not used or disseminated for analysis,” *Biometrika*, 95, 933–946.


Biography

Yajuan Si was born on February 29, 1984 in Hebei, China. She received her Bachelor’s degree in Statistics (minor on Actuarial Science and Risk Management) in School of Statistics at Renmin University of China in 2008. Yajuan came to Durham, NC in August 2008 to pursue a Ph.D in statistics in the Department of Statistical Science at Duke University. In 2011, she earned a Master’s degree in Statistical Science en route to her Ph.D.