Systematically Missing Subject-Level Data in Longitudinal Research Synthesis

Dissertation

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of The Ohio State University

By

David Kline, M.S.

Graduate Program in Biostatistics

The Ohio State University

2015

Dissertation Committee:

Dr. Rebecca Andridge, Co-Advisor
Dr. Eloise Kaizar, Co-Advisor
Dr. Bo Lu
Abstract

When conducting research synthesis, the collection of studies that will be combined often do not measure the same set of variables, which creates missing data. When the studies to combine are longitudinal, missing data can occur on either the observation-level (time-varying) or the subject-level (non-time-varying). Traditionally, the focus of missing data methods for longitudinal data has been on missing observation-level variables. In this dissertation, we focus on missing subject-level variables where few methods have been developed or compared.

We compare two multiple imputation approaches that have been proposed for missing subject-level data in single longitudinal studies: a joint modeling approach and a sequential conditional modeling approach. Based on analytical and empirical results for the case when all variables are normally distributed, we find the joint modeling approach to be preferable to the sequential conditional approach except when the covariance structure of the repeated outcome for each individual has homogeneous variance and exchangeable correlation. Specifically, the regression coefficient estimates from an analysis incorporating imputed values based on the sequential conditional method are attenuated and less efficient than those from the joint method.

Based on this preference, we develop a new joint model for multiple imputation of missing subject-level variables that models subject- and observation-level variables
with distributions in the exponential family. Our model is built within the gen-
eralized linear models framework and uses normally distributed latent variables to
account for dependence on both the subject- and observation-levels. When compared
via simulation, the performance of our model is similar to or better than existing ap-
proaches for imputing missing subject-level variables with normal, Bernoulli, Poisson,
and multinomial distributions. We illustrate our method by applying it to combine
two longitudinal studies on the psychological and social effects of pediatric traumatic
brain injury that have systematically missing subject-level data.
To my wife Jennifer and my parents Randy and Deb.
Acknowledgments

First, I would like to thank my advisors, Dr. Eloise Kaizar and Dr. Rebecca Andridge. I am truly grateful for all of the time and effort that they have put into mentoring me over the past four years. I have learned a tremendous amount that I will certainly apply throughout my career. I would also like to acknowledge Dr. Bo Lu and Dr. Elizabeth Stasny for their valuable insights and service on my committee.

Thank you to my friends and family for all of their love and support over the past five years. I am certain that I would not be where I am today without them. I would especially like to thank my wife, Jennifer, for her patience and willingness to walk with me through the highs and lows of my graduate school experience.
Vita


Education

2010 ................................. B.A. Mathematics,
Messiah College
Mechanicsburg, PA

2012 ................................. M.S. Statistics,
The Ohio State University
Columbus, OH

Professional Experience

2009-2010 ............................. Statistical Analyst Intern,
Department of Public Health Sciences
Pennsylvania State University
Hershey, PA

2011-2014 ............................. Graduate Teaching Assistant,
Department of Statistics and Division
of Biostatistics
The Ohio State University
Columbus, OH

2014-2015 ............................. Graduate Research Assistant,
Department of Statistics and Division
of Biostatistics
The Ohio State University
Columbus, OH
Publications

Research Publications


Fields of Study

Major Field: Biostatistics
# Table of Contents

Abstract . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . ii

Dedication . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . iv

Acknowledgments . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . v

Vita . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . vi

List of Tables . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . xi

List of Figures . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . xiii

1. Motivation . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 1
   1.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 1
   1.2 Pediatric TBI Studies to be Combined . . . . . . . . . . . . . . . . . 3
   1.3 Preliminary Method for Meta-Analysis of TBI Studies . . . . . . . 5
   1.4 Results of Preliminary Meta-Analysis . . . . . . . . . . . . . . . . . 7
   1.5 Analysis of Each Study Separately . . . . . . . . . . . . . . . . . . . 10
   1.6 Missing Data . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 10

2. Literature Review . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 13
   2.1 Longitudinal Data . . . . . . . . . . . . . . . . . . . . . . . . . . . . 13
      2.1.1 Mixed Models . . . . . . . . . . . . . . . . . . . . . . . . . . 14
      2.1.2 Generalized Estimating Equations . . . . . . . . . . . . . . . 15
      2.1.3 Joint Modeling . . . . . . . . . . . . . . . . . . . . . . . . . . 16
   2.2 Research Synthesis . . . . . . . . . . . . . . . . . . . . . . . . . . . 19
      2.2.1 Types of Data . . . . . . . . . . . . . . . . . . . . . . . . . . 19
      2.2.2 Methods for Summary Statistics . . . . . . . . . . . . . . . . 21
      2.2.3 Methods for Individual Patient Data . . . . . . . . . . . . . 22
2.2.4 Longitudinal Methods for Meta-Analysis . . . . . . . . . . . . . . . . 23
2.3 Missing Data . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 27
  2.3.1 Types of Missing Data . . . . . . . . . . . . . . . . . . . . . . 27
  2.3.2 General Methods for Analysis With Missing Data . . . . . . . . 32
  2.3.3 Model-based Multiple Imputation . . . . . . . . . . . . . . . . . 35
2.4 Clustered Missing Data . . . . . . . . . . . . . . . . . . . . . . . . . 40
   2.4.1 Multiple Imputation for Clustered Data . . . . . . . . . . . . . 40
   2.4.2 Missing Data in Research Synthesis . . . . . . . . . . . . . . . 50
2.5 Summary . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 53
3. Comparing Multiple Imputation Methods for Systematically Missing Subject-
   Level Data . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 54
  3.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 54
  3.2 Multiple Imputation Models . . . . . . . . . . . . . . . . . . . . . . . 56
  3.3 Comparisons Between Imputation Distributions . . . . . . . . . . . . . 62
    3.3.1 Imputation Distributions . . . . . . . . . . . . . . . . . . . . 62
    3.3.2 Comparing Means . . . . . . . . . . . . . . . . . . . . . . . . 64
    3.3.3 Comparing Variances . . . . . . . . . . . . . . . . . . . . . . . 66
  3.4 Simulation to Assess Impact on Estimation . . . . . . . . . . . . . . . 67
    3.4.1 Data Generation . . . . . . . . . . . . . . . . . . . . . . . . . 67
    3.4.2 Imputation and Analysis Procedures . . . . . . . . . . . . . . 69
    3.4.3 Performance Metrics . . . . . . . . . . . . . . . . . . . . . . . 70
    3.4.4 Results . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 71
  3.5 Comparison of Results for TBI Data Example . . . . . . . . . . . . . . 73
  3.6 Discussion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 77
4. Joint Two-Level Model for Variables of Mixed Type Using Latent Variables 84
  4.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 84
  4.2 Joint Two-Level Model . . . . . . . . . . . . . . . . . . . . . . . . . 85
    4.2.1 Latent Variables . . . . . . . . . . . . . . . . . . . . . . . . . 86
    4.2.2 Observation-Level Sub-Model . . . . . . . . . . . . . . . . . . 89
    4.2.3 Subject-Level Sub-Model . . . . . . . . . . . . . . . . . . . . 90
    4.2.4 Example . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 91
    4.2.5 Congeniality . . . . . . . . . . . . . . . . . . . . . . . . . . . 92
    4.2.6 Prior Distributions . . . . . . . . . . . . . . . . . . . . . . . . 94
  4.3 Simulation Study to Assess Use in Multiple Imputation . . . . . . . . . 95
    4.3.1 Data Generation . . . . . . . . . . . . . . . . . . . . . . . . . 95
    4.3.2 Analysis Methods . . . . . . . . . . . . . . . . . . . . . . . . . 101
    4.3.3 Implementation of Imputation Procedures . . . . . . . . . . . 102
    4.3.4 Performance Metrics . . . . . . . . . . . . . . . . . . . . . . . 112
4.3.5 Results .................................................. 112
4.4 TBI Data Analysis .......................................... 126
4.5 Choosing the Number of Latent Variables ............... 131
4.6 Discussion ................................................ 136

5. Summary and Future Work .................................... 142
   5.1 Comparing Methods for Multiple Imputation ............. 142
   5.2 A Novel Two-Level Joint Imputation Model ............... 144
   5.3 Future Work ............................................ 145
       5.3.1 Further Development of the Joint Model ............ 145
       5.3.2 Extension to Model Heterogeneity .................. 147

References ..................................................... 150


List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>TBI Data Summary</td>
</tr>
<tr>
<td>1.2</td>
<td>Preliminary Analysis Results</td>
</tr>
<tr>
<td>1.3</td>
<td>TBI Missing Data</td>
</tr>
<tr>
<td>3.1</td>
<td>Simulation Correlation and Variance Parameter Settings</td>
</tr>
<tr>
<td>3.2</td>
<td>Simulation Results</td>
</tr>
<tr>
<td>3.3</td>
<td>Relative Estimated Efficiencies</td>
</tr>
<tr>
<td>3.4</td>
<td>Estimated Correlation and Variance Parameters</td>
</tr>
<tr>
<td>3.5</td>
<td>TBI Application Results</td>
</tr>
<tr>
<td>4.1</td>
<td>Simulation Covariance Parameter Settings</td>
</tr>
<tr>
<td>4.2</td>
<td>Simulation Results: Normally Distributed Covariate</td>
</tr>
<tr>
<td>4.3</td>
<td>Simulation Results: Binary Covariate Case 1</td>
</tr>
<tr>
<td>4.4</td>
<td>Simulation Results: Binary Covariate Case 2</td>
</tr>
<tr>
<td>4.5</td>
<td>Simulation Results: Poisson Covariate</td>
</tr>
<tr>
<td>4.6</td>
<td>Simulation Results: Ordinal Covariate</td>
</tr>
<tr>
<td>4.7</td>
<td>Simulation Results: Categorical Covariate</td>
</tr>
</tbody>
</table>
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Diagram of Missing Data Patterns</td>
<td>28</td>
</tr>
<tr>
<td>3.1</td>
<td>Average Confidence Interval Width and Coverage</td>
<td>71</td>
</tr>
<tr>
<td>4.1</td>
<td>Joint Latent Variable Model Diagram</td>
<td>87</td>
</tr>
</tbody>
</table>
Chapter 1: Motivation

1.1 Introduction

The consequences of traumatic brain injuries (TBI) have been well publicized and garnered a lot of attention over the past few years; however, there is still much to learn about both the short and long term effects of TBI. This is especially the case for injuries that occur in adolescents and children. While researchers have made substantial progress on understanding the neurological impacts of such injuries, little has been done to examine the social and psychological impacts of TBI on both the affected children and their families. There have been several longitudinal studies conducted to investigate this issue, but they have been relatively small and narrow in their scope (Stancin et al., 2008; Wade et al., 2006, 2002; Yeates et al., 2012).

It is particularly difficult for studies of this nature to recruit a sufficient number of patients and keep them enrolled in the study long enough to learn about long term effects of the injury. Other issues, such as the timing of enrollment and the age of the patients, can also hinder recruitment. These studies are also quite costly to administer due to the need for long term follow-up and the detailed level of behavioral and psychological information that must be collected. These challenges make it very difficult to draw broad conclusions about the long term effects of TBI and how those
effects may differ for children of different ages at injury. However, these are precisely the questions that are of interest to researchers, so it is beneficial to consider combining the studies on pediatric TBI. Statistical methods have been developed to combine multiple studies, commonly referred to as research synthesis, or meta-analysis.

Difficulty with enrollment is just one example of a situation where it can be advantageous to examine collections of studies. In another example, Higgins et al. (2001) conducted a meta-analysis of small to moderate sized clinical trials to examine the dosage effect of the drug tacrine for treating Alzheimer’s Disease. In this instance, each study administered different dosages of the drug. It can also be beneficial to combine studies to synthesize the current state of research, as in a systematic review (Deeks et al., 2011). Sometimes different studies may address slightly different questions and combining studies might help answer a unified question of interest (Lumley, 2002).

The goal of our research is to improve the methodology for combining information from collections of studies, specifically targeting the handling of missing data. This can be particularly challenging when combining different studies because there are two distinct types of missingness. There is missing data within each study, such as loss to follow-up, which in the context of meta-analysis is referred to as sporadic missingness (Burgess et al., 2013). In addition, some data may be missing because the studies were not designed to collect the exact same information. This is referred to as systematic missingness (Fibrinogen Studies Commission, 2009; Resche-Rigon et al., 2013). A variety of methods have been developed for sporadic missing data (Little and Rubin, 2002), but the treatment of systematic missing data is not yet fully developed in the literature. Our research aims to develop methods for handling systematically
missing data when combining longitudinal studies. This involves building on research synthesis methodology developed for longitudinal data analysis and for missing data, each of which we introduce in Chapter 2.

We will apply the new methodology to combine two studies on pediatric TBI to investigate the association between socioeconomic status and the longitudinal psychological and social outcomes of TBI across all ages of children. A single study has not been conducted to address this question, so we will combine data from a study that is focused on preschool age children (Chapman et al., 2010) and a study that is focused on school age children (Fay et al., 2009). Together these two studies will provide information about children who were between 3 and 12 years old at the time of injury. We will use this combined data set to estimate the effect of socioeconomic status on psychological and social outcomes related to pediatric TBI.

1.2 Pediatric TBI Studies to be Combined

As mentioned above, there are two studies that were conducted on pediatric TBI that will provide the datasets for our research. The school age study was conducted on children ages 6 to 12 years at the time of injury (Fay et al., 2009). Families were recruited from admissions to four different hospitals in Ohio. Children needed to have no evidence of child abuse and have English as the primary language at home to be eligible. The TBI group consisted of children who suffered documented moderate-to-severe TBI. The control group consisted of children who suffered orthopedic injuries and were hospitalized overnight with no findings suggestive of brain injuries. The study was able to recruit 52 children with severe TBI, 56 with moderate TBI, and 80 with orthopedic injuries. Baseline measurements were administered as soon as
possible after injury. Follow-ups were conducted at 6 months, 12 months, and at a
final time point that averaged 4.1 years after injury. Measurements were collected on
injury-related burden, parental psychological distress, and family functioning.

The preschool age study was conducted on children ages 36 to 83 months at the
time of injury (Chapman et al., 2010). This study was also conducted based on ad-
missions to four Ohio hospitals and had the same eligibility criteria as the school age study. This study also had a control group of children who suffered orthopedic injuries. The TBI group consisted of children with severe, moderate, and complicated mild TBI. Complicated mild TBI was defined as children whose neuroimaging results suggested a more significant brain injury than is typical for cases of mild TBI. The study was able to recruit 23 children with severe TBI, 64 with moderate and comp-
plicated mild TBI, and 119 with orthopedic injuries. Baseline measurements were
administered as soon as possible after injury and follow-up visits were conducted at
6 months, 12 months, and 18 months after injury. Measurements were collected on
injury-related burden and parental distress.

Table 1.1 displays a very brief summary of the variables that we consider through-
out this dissertation. We primarily focus our attention on one outcome variable, the
normed internalizing subscale of the Child Behavior Checklist (CBCL-I) (Achenbach,
1991), and are interested in estimating the effects of various injury and demographic
related factors on this outcome. Summaries of two other outcomes that are consid-
ered in Chapter 4, the normed externalizing subscale of the Child Behavior Checklist
(CBCL-E) (Achenbach, 1991) and the normed Brief Symptom Index General Severity
Index (BSI GSI) (Gioia et al., 2000), are also presented in Table 1.1.

4
The Family Hollingshead (Hollingshead, 1975) measure is systematically missing for the preschool age study. The income and number of additional children variables are systematically missing for the school age study. Thus, we cannot incorporate these covariates into a combined analysis without a procedure that can handle systematic missingness. We also drop the long term follow-up visits in both studies since they occur at different time intervals post-injury.

1.3 Preliminary Method for Meta-Analysis of TBI Studies

A model for the meta-analysis of longitudinal trials with individual patient data was proposed by Jones et al. (2009). The model for the length \( k \) vector of observations, \( Y_{ij} \), on individual \( j \) in study \( i \) is given by:

\[
Y_{ij} = X_{ij}\beta + \epsilon_{ij}
\]  

where \( X_{ij} \) is the matrix of covariates for individual \( j \) in study \( i \) with fixed effects \( \beta \) and \( \epsilon_{ij} \) is vector of random errors. The covariate matrix should include an indicator of study membership to allow the intercepts to differ by study. Also, \( \epsilon_{ij} \) is distributed as \( N(0, \Sigma_i) \). We model \( \Sigma_i \) for each study by using a fully unstructured covariance matrix. This model will be more fully discussed in Section 2.2.4.

Since this model does not handle missing data, we limit our preliminary analysis to variables that are observed in both studies. We also only include variables that are defined identically for both studies. Specifically, we regress CBCL-I on the injury and demographic characteristics present in Table 1.1 that were observed in both studies. The model was fit using PROC MIXED in SAS.
<table>
<thead>
<tr>
<th>Variable</th>
<th>School Age</th>
<th>Preschool Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Size</td>
<td>188</td>
<td>206</td>
</tr>
<tr>
<td>Age at Injury (years)</td>
<td>9.5 (2.0)</td>
<td>5.1 (1.1)</td>
</tr>
<tr>
<td>Number of Additional Children</td>
<td>NA</td>
<td>1.4 (1.1)</td>
</tr>
<tr>
<td>Hollingshead Index</td>
<td>33.6 (15.0)</td>
<td>NA</td>
</tr>
<tr>
<td>Income</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt; $25,000</td>
<td>NA</td>
<td>40%</td>
</tr>
<tr>
<td>$25,000 - $50,000</td>
<td>NA</td>
<td>33%</td>
</tr>
<tr>
<td>&gt; $50,000</td>
<td>NA</td>
<td>27%</td>
</tr>
<tr>
<td>Maternal Education</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Did not Graduate from High School</td>
<td>15%</td>
<td>12%</td>
</tr>
<tr>
<td>High School Graduate</td>
<td>72%</td>
<td>59%</td>
</tr>
<tr>
<td>College Graduate</td>
<td>13%</td>
<td>29%</td>
</tr>
<tr>
<td>Head Injury Severity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Severe TBI</td>
<td>28%</td>
<td>11%</td>
</tr>
<tr>
<td>Complicated Mild or Moderate TBI</td>
<td>30%</td>
<td>31%</td>
</tr>
<tr>
<td>Orthopedic Control</td>
<td>42%</td>
<td>58%</td>
</tr>
<tr>
<td>CBCL-I</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline</td>
<td>50.6 (10.7)</td>
<td>47.8 (10.8)</td>
</tr>
<tr>
<td>6 Months Post-Injury</td>
<td>51.1 (11.0)</td>
<td>47.8 (11.0)</td>
</tr>
<tr>
<td>12 Months Post-Injury</td>
<td>51.3 (11.0)</td>
<td>46.5 (10.7)</td>
</tr>
<tr>
<td>CBCL-E</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline</td>
<td>50.9 (10.2)</td>
<td>48.3 (12.0)</td>
</tr>
<tr>
<td>6 Months Post-Injury</td>
<td>51.2 (10.5)</td>
<td>48.3 (12.2)</td>
</tr>
<tr>
<td>12 Months Post-Injury</td>
<td>51.3 (12.1)</td>
<td>49.1 (11.9)</td>
</tr>
<tr>
<td>BSI GSI</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline</td>
<td>55.3 (11.2)</td>
<td>50.7 (10.9)</td>
</tr>
<tr>
<td>6 Months Post-Injury</td>
<td>52.5 (10.7)</td>
<td>51.3 (11.9)</td>
</tr>
<tr>
<td>12 Months Post-Injury</td>
<td>52.2 (11.6)</td>
<td>50.0 (11.2)</td>
</tr>
</tbody>
</table>

Table 1.1: Summary of variables from the two studies that are included in the example data analysis. Entries corresponding to the variables are means (standard deviations) for continuous variables, and percentages for categorical variables. Note that there is no estimate for the Hollingshead Index in the preschool age study, as this study did not calculate this variable. Likewise, there are no estimates for income and number of additional children in the school age study.
1.4 Results of Preliminary Meta-Analysis

The main covariates of interest for the model are time post-injury, injury severity group, age at injury, and maternal education. Time is modeled as a categorical effect, as there does not appear to be a linear effect of time on CBCL-I. Injury groups are severe TBI, moderate TBI, and orthopedic injury. Maternal education is the proxy for socioeconomic status (SES) that was common to both studies. We also include a fixed effect of study to account for across-study variability. The reference group is the orthopedic injury group with a mother who is a college graduate in the school age study. We select the covariates for inclusion in the model via backward selection and consider up to second order interactions. The main effect of each covariate is retained in the model regardless of significance. The resulting model estimates are shown in Table 1.2.

There are a few results to highlight from Table 1.2. First, we notice that the mean CBCL-I at each visit differs by injury group as shown by the significant injury group by time post-injury interaction ($p < 0.001$). We see worse outcomes (higher CBCL-I scores) over time associated with more severe TBI injuries. We also see a significant main effect of maternal education ($p < 0.001$), which shows worse outcomes for children with less educated mothers. Thus, in addition to injury severity, socioeconomic status appears to be an important factor that influences psychological conditions related to internalizing behaviors.

The error covariance matrix, $\Sigma$, is estimated separately for each study. The estimate of the residual covariance for the preschool study is

$$
\hat{\Sigma}_{\text{preschool}} = \begin{bmatrix}
113.4 & 82.9 & 72.1 \\
82.9 & 112.4 & 80.9 \\
72.1 & 80.9 & 105.7
\end{bmatrix},
$$

(1.2)
<table>
<thead>
<tr>
<th>Variable</th>
<th>Level</th>
<th>Estimate</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td></td>
<td>45.62</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Study</td>
<td>Preschool</td>
<td>-1.84</td>
<td>0.272</td>
</tr>
<tr>
<td></td>
<td>School Age</td>
<td>ref</td>
<td></td>
</tr>
<tr>
<td>Age at Injury</td>
<td></td>
<td>0.16</td>
<td>0.604</td>
</tr>
<tr>
<td>Group</td>
<td>Severe TBI</td>
<td>0.18</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td></td>
<td>Complicated Mild or Moderate TBI</td>
<td>1.98</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Orthopedic Injury</td>
<td>ref</td>
<td></td>
</tr>
<tr>
<td>Education</td>
<td>Did not Graduate from High School</td>
<td>7.20</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td></td>
<td>High School Graduate</td>
<td>2.82</td>
<td></td>
</tr>
<tr>
<td></td>
<td>College Graduate</td>
<td>ref</td>
<td></td>
</tr>
<tr>
<td>Time Post-Injury</td>
<td>Baseline</td>
<td>ref</td>
<td>0.334</td>
</tr>
<tr>
<td></td>
<td>6 months</td>
<td>-1.31</td>
<td></td>
</tr>
<tr>
<td></td>
<td>12 months</td>
<td>-2.51</td>
<td></td>
</tr>
<tr>
<td>Group*Time Post-Injury</td>
<td>Severe TBI*Baseline</td>
<td>ref</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td></td>
<td>Severe TBI*6 months</td>
<td>4.38</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Severe TBI*12 months</td>
<td>6.16</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Complicated Mild or Moderate TBI*Baseline</td>
<td>ref</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Complicated Mild or Moderate TBI*6 months</td>
<td>1.87</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Complicated Mild or Moderate TBI*12 months</td>
<td>3.08</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Orthopedic Injury*Baseline</td>
<td>ref</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Orthopedic Injury*6 months</td>
<td>ref</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Orthopedic Injury*12 months</td>
<td>ref</td>
<td></td>
</tr>
</tbody>
</table>

Table 1.2: Estimates of $\beta$ from fitting Model 1.1 for CBCL-I. Reference categories are denoted by “ref”.
and the estimate for the school age study is

\[
\hat{\Sigma}_{\text{school}} = \begin{bmatrix}
114.6 & 59.1 & 62.8 \\
59.1 & 107.5 & 73.6 \\
62.8 & 73.6 & 100.6
\end{bmatrix},
\]  

(1.3)

where the rows of each matrix correspond to the baseline, 6 month, and 12 month visits. We see some evidence of nonconstant variance across time and differences in the covariance structures of the two studies. To see the differences more clearly, we examine estimates of the correlation matrices:

\[
\hat{R}_{\text{preschool}} = \begin{bmatrix}
1 & 0.73 & 0.66 \\
0.73 & 1 & 0.74 \\
0.66 & 0.74 & 1
\end{bmatrix},
\]  

(1.4)

\[
\hat{R}_{\text{school}} = \begin{bmatrix}
1 & 0.53 & 0.58 \\
0.53 & 1 & 0.71 \\
0.58 & 0.71 & 1
\end{bmatrix}.
\]  

(1.5)

The estimated correlation structure for the preschool age study is what we tend to expect for longitudinal data, with stronger correlations between consecutive measurements compared to measurements taken farther apart. However, a different pattern is seen for the school age study. The estimated correlation between the 6 month visit and the 12 month visit is higher than any of the other correlations. In addition, the 12 month and baseline visits have a higher correlation than the 6 month and baseline time points. We hypothesize that this is because by the 12 month visit the immediate impacts and stresses of the injury have diminished, and the child’s behavior has started to move back toward baseline (pre-injury) levels. Thus, we will emphasize modeling techniques that are appropriate for modeling unstructured error covariance matrices throughout the rest of this dissertation.
1.5 Analysis of Each Study Separately

As mentioned in Section 1.3, we could only include variables that were common to both studies in our preliminary analysis. This is quite limiting, as we had to discard roughly half of the variables collected in each study. By doing so, we may have missed other important covariates that could be associated with CBCL-I. To examine what we may have missed, we analyze each study separately thus eliminating the problem of systematic missingness.

Each separate analysis reveals additional variables that were not included in the combined analysis that may be of interest. For the preschool study, we find a significant main effect of household income, a three-level categorical variable. We also see a main effect of the number of additional children in the home that we would like to investigate further. In our analysis of the school age study, the Hollingshead Family Scale (a measure of SES) has a significant effect on CBCL-I and in fact modifies the change in CBCL-I over time. Since both separate analyses reveal significant SES effects beyond maternal education, we would like to consider the inclusion of an additional indicator of SES, either Family Hollingshead or income, in our model for CBCL-I that uses the combined data from both studies.

1.6 Missing Data

Thus far, we have not included subjects with partially observed covariates in any of our analyses. As in any study, there are data values that are unable to be observed. This could be due to drop out or nonresponse (e.g., refusal to provide one’s answer). While sporadic missing data of this type is also important, it is not the focus of this dissertation. We describe some of the well developed literature on
methods for handling sporadic missing data in Chapter 2, but focus the remainder of this dissertation on systematic missingness like we encounter for income, the number of additional children in the home, and Family Hollingshead. Each of these variables was only collected in one study, but we would like to include them in our synthesized analysis of CBCL-I.

Table 1.3 shows the proportion of complete observations for each variable that we use in this dissertation. We see that maternal education is almost completely observed in both studies. The number of additional children in the home and Family Hollingshead variables are almost completely observed for the study in which they were collected. Income is only observed in two-thirds of the subjects in the preschool study which is likely related to the sensitivity of reporting one’s income.

Since our focus is on systematic missing data, we omit discussion of the detailed patterns of missingness for the repeated responses. Table 1.3 shows the proportion of subjects that have observed measurements at all three time points for the three outcomes, CBCL-I, CBCL-E, and BSI GSI. The table also shows the proportion of subjects for whom we have completely observed all three of the longitudinal outcomes at all three time points, as this will be required for our data example in Chapter 4. We see that roughly 75% of subjects in the combined data set have completely observed the three outcomes of interest.

To incorporate the variables that were not observed in both studies, we need to develop methodology to account for systematically missing data. Specifically, the methodology will need to be able to account for missing data at the subject level, i.e., non-time varying baseline values. In addition, we will need to be able to handle variables that are not normally distributed since the number of children living in the
<table>
<thead>
<tr>
<th></th>
<th>School Age</th>
<th>Preschool Age</th>
<th>Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Covariates</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maternal Education</td>
<td>0.98</td>
<td>0.99</td>
<td>0.98</td>
</tr>
<tr>
<td>Number of Additional Children</td>
<td>NA</td>
<td>0.98</td>
<td>0.51</td>
</tr>
<tr>
<td>Income</td>
<td>NA</td>
<td>0.66</td>
<td>0.34</td>
</tr>
<tr>
<td>Family Hollingshead</td>
<td>1</td>
<td>NA</td>
<td>0.48</td>
</tr>
<tr>
<td><strong>Outcomes</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CBCL-I</td>
<td>0.81</td>
<td>0.74</td>
<td>0.78</td>
</tr>
<tr>
<td>CBCL-E</td>
<td>0.81</td>
<td>0.74</td>
<td>0.77</td>
</tr>
<tr>
<td>BSI GSI</td>
<td>0.77</td>
<td>0.74</td>
<td>0.75</td>
</tr>
<tr>
<td>All 3 outcomes</td>
<td>0.77</td>
<td>0.71</td>
<td>0.74</td>
</tr>
</tbody>
</table>

Table 1.3: Summary of variables from the two studies that are included in the example data analysis. We present the proportion of subjects for each variable that are fully observed. For outcomes, the proportion shown is the proportion of subjects who are fully observed at all three time points. NA denotes that a variable is systematically missing.

home is a count and income is categorical. Because baseline variables tend to have sparse missingness in the single-study setting, as shown by Family Hollingshead and the number of additional children, few methods appropriate for longitudinal datasets with time-invariant missing variables have been developed or compared. This leaves open the question of how to proceed when synthesizing longitudinal datasets with design-induced systematic missingness in variables that do not vary with time.
Chapter 2: Literature Review

Our research will build on current methodology that exists for longitudinal data, research synthesis, and missing data. This chapter provides an introduction to parts of each area that are relevant to our problem of missing data in a longitudinal meta-analysis. Section 2.1 describes methods that have been designed for the general problem of dealing with longitudinal data. Section 2.2 examines approaches for combining information from multiple studies, including multiple longitudinal studies. Section 2.3 introduces general methodology for dealing with missing data. Finally, Section 2.4 explores work that has been done to specifically address clustered missing data, which includes missing data in meta-analysis.

2.1 Longitudinal Data

Longitudinal studies are designed to examine how individuals change over time. This is done by following an individual for a period of time and recording measurements at multiple time points. There are two primary types of variables that are observed in longitudinal studies: time varying and non-time varying. Time varying variables are variables whose values change over time within an individual and are also referred to as observation-level variables. An example of an observation-level variable is the result of a particular diagnostic test that is administered at multiple
visits. Non-time varying variables are variables whose values do not change over time within an individual and are called subject-level variables. Baseline and demographic characteristics, such as age at injury and sex, are examples of subject-level variables.

The subject- and observation-level terminology helps to place longitudinal data in the more general framework of hierarchical or clustered data. The observation-level is the level below the subject-level. Observations that fall within the same subject are related, or clustered. This relationship must be accounted for by using models that properly incorporate correlation between observations. Two approaches for handling clustered data are mixed-effects models and generalized estimating equations. When multiple clustered outcomes are of interest, joint modeling is commonly employed to model the correlation within and between outcomes.

2.1.1 Mixed Models

One method for analyzing hierarchical data involves using mixed models. Mixed models specify probability distributions to model individual trajectories. This is commonly done through specification of a hierarchical model. Separate probability distributions are specified for the random effects and the observations given the random effects, thus creating the hierarchy. Random effects are used to model latent differences between individuals, or clusters.

Mixed models are specified using both fixed and random effects. Random effects induce correlation between observations that are in the same cluster. For instance, two observations for the same individual would share the same random effect for that individual and, thus, would be correlated. Observations from different subjects are assumed to be independent. A typical generalized linear mixed model for an
observation-level outcome vector $\mathbf{Y}_i$ with an element for each time point, $j$, for subject $i$ is

$$
\mathbf{Y}_i | \mu_i \sim F(\mu_i) \tag{2.1}
$$

$$
g(\mu_i) = \mathbf{X}_i \beta + \mathbf{Z}_i \mathbf{b}_i
$$

where $F(\cdot)$ is a distribution function, $g(\cdot)$ is a link function, $\mu_i$ is the mean vector of $\mathbf{Y}_i$, $\mathbf{X}_i$ is the matrix of covariates, with the $j^{th}$ row corresponding to the $j^{th}$ time point, with fixed effects $\beta$, and $\mathbf{Z}_i$ is the matrix of covariates with random effects $\mathbf{b}_i$.

It is commonly assumed that $\mathbf{b}_i \overset{ind}{\sim} N(\mathbf{0}, \mathbf{\Psi})$ where the covariance matrix, $\mathbf{\Psi}$, can be specified to allow for correlation between elements of $\mathbf{b}_i$.

The fixed effect parameter, $\beta$, for a mixed model has a conditional interpretation (Rabe-Hesketh and Skrondal, 2009): the parameter is interpreted as the effect within a particular individual. For example, consider a model for family stress with a positive fixed effect for severe TBI. Then, holding random effects constant, we expect a family with a child who suffered severe TBI to have more stress than a family with a child who suffered an orthopedic injury.

### 2.1.2 Generalized Estimating Equations

A method for analyzing correlated data to estimate population average effects (as opposed to individual trajectories) is generalized estimating equations (GEE) (Liang and Zeger, 1986). GEE is an approach based on the theory of quasi-likelihoods, which means that a functional form of the likelihood is never specified. We will not be using GEE in the rest of our work for two main reasons. First, the missing data methods being explored are model-based approaches. Since GEE does not involve specification of a likelihood function, it does not fit well with the proposed missing
data strategies. Additionally, for our TBI application, we are not primarily interested in the population-based inferences provided by GEE. The goal of the TBI studies is to learn about individual trajectories to help shape individual treatment plans. Population average inferences are not well suited to this goal.

2.1.3 Joint Modeling

Often, multiple variables are collected that are of interest, rather than a single outcome as was considered in Sections 2.1.1 and 2.1.2. For example, we might be interested in jointly assessing a child’s internalizing and externalizing behavior problems, rather than considering them independently. In this section, we will discuss two approaches for jointly modeling longitudinal variables that extend mixed effects models using latent variables.

Joint models use correlated or shared random effects to capture the dependence between outcome variables. Random effects in a joint model are often called latent variables. Shared latent variables are commonly used in structural equation modeling (Little, 2013). Often, latent variables are interpreted as characteristics of a subject that are not directly observable. For example, a latent variable could correspond to an individual’s level of sickness. Song et al. (2011) and Dunson (2003) present approaches that use shared latent variables for jointly modeling multiple longitudinal observation-level variables within the generalized linear models framework.

Suppose we are interested in jointly modeling \( p^{(o)} \) observation-level variables that are observed for \( n \) subjects. Superscripts will be used to differentiate between the subject- and observation-levels. Each observation-level variable is measured at \( t \) time points. Let observation-level variable \( m \) for subject \( i \) at time \( j \) be denoted by \( Y_{ijm}^{(o)} \),
where \( i = 1, \ldots, n, j = 1, \ldots, t, \) and \( m = 1, \ldots, p^{(o)} \), and have a distribution in the exponential family with canonical parameter, \( \theta^{(o)}_{ijm} \).

Song et al. (2011) proposed a joint longitudinal structural equation model that is based on two vectors of latent variables: a length \( q^{(o)} \) vector of observation-level latent variables for each time \( j \), \( \xi^{(o)}_{ij} \), and a length \( q^{(s)} \) vector of subject-level latent variables, \( \xi^{(s)}_i \). The observation-level latent variables represent unobservable characteristics of a subject that vary over time, and the subject-level latent variables represent unobservable characteristics that are time-invariant. Each observation-level variable at time \( j \) is assumed to be a function of both the subject- and observation-level characteristics at time \( j \).

The vectors of latent variables are shared across observation-level variables, and the elements of each vector can be correlated within a subject. That is, we assume

\[
\begin{align*}
(\xi^{(o)}_{i1}, \ldots, \xi^{(o)}_{it})' & \overset{iid}{\sim} N(0, \Psi^{(o)}) \\
\xi^{(s)}_i & \overset{iid}{\sim} N(0, \Psi^{(s)})
\end{align*}
\]

(2.2)

(2.3)

where \( \Psi^{(o)} \) is dimension \( tq^{(o)} \times tq^{(o)} \) and \( \Psi^{(s)} \) is dimension \( q^{(s)} \times q^{(s)} \). One particularly intuitive approach for modeling \( \Psi^{(o)} \) for longitudinal data is through an autoregression of the latent variables at the current time point on the latent variables at the previous time points (Dunson, 2003; Song et al., 2011), which is outlined in general in Daniels and Pourahmadi (2002). The model for the vector of observation-level latent variables at time \( j \) for subject \( i \) is given by:

\[
\xi^{(o)}_{ij} = \alpha w_{ij} + \sum_{k=0}^{j-1} (\xi^{(o)}_{ik} - \alpha w_{ij}) \phi_{jk} + \delta_{ij},
\]

(2.4)

where \( w_{ij} \) is a length \( w \) vector of covariates with a dimension \( q^{(o)} \times w \) matrix of fixed effects \( \alpha \), \( \xi^{(o)}_{i0} = 0 \), \( \phi_{jk} \) is a length \( q^{(o)} \) vector of autoregressive parameters with
\( \phi_{j0} = 0 \), and errors \( \delta_{ij} \overset{iid}{\sim} N(0, \psi_j) \) where \( \psi_j \) is diagonal with dimension \( q^{(o)} \times q^{(o)} \).

Thus, observation-level latent variables at the current time point are related to the latent variables from previous time points. We can similarly decompose \( \Psi^{(s)} \), although without the autoregressive interpretation of the parameters since the subject-level latent variables are time-invariant. Alternatively, \( \Psi^{(s)} \) can be treated as in any standard multivariate normal analysis.

The model for outcome \( Y_{ijm}^{(o)} \) with distribution from the exponential family, link function \( h_m(\cdot) \) and canonical parameter \( \theta_{ijm}^{(o)} \) as defined by Song et al. (2011) is:

\[
\begin{align*}
    h_m^{(o)}(\theta_{ijm}^{(o)}) &= \mu_{jm}^{(o)} + \xi_{ijm}^{(o)} \lambda_{jm}^{(o)} + \xi_{i}^{(s)} \gamma_{m}^{(o)} + \nu_{im}^{(o)}, & m = 1, \ldots, p^{(o)}, \\
    \nu_{im}^{(o)} &\overset{ind}{\sim} N(0, \omega_{m}^2)
\end{align*}
\]

where \( h_m^{(o)}(\cdot) \) is a link function and \( \mu_{jm}^{(o)} \) is a fixed intercept at time \( j \) specific to observation-level variable \( m \). Observation-level characteristics are related to \( Y_{ijm}^{(o)} \) at time \( j \) through \( \xi_{ijm}^{(o)} \) and the associated length \( q^{(o)} \) vector of factor loadings, \( \lambda_{jm}^{(o)} \). Thus, \( \mu_{jm}^{(o)}, \xi_{ijm}^{(o)}, \) and \( \lambda_{jm}^{(o)} \) are the time-varying parts of the model. Subject-level characteristics are related to \( Y_{ijm}^{(o)} \) through \( \xi_i^{(s)} \) and the associated length \( q^{(s)} \) vector of factor loadings, \( \gamma_{m}^{(o)} \). There is also a latent subject and variable specific intercept, \( \nu_{im}^{(o)} \). Thus, \( \xi_i^{(s)}, \gamma_{m}^{(o)}, \) and \( \nu_{im}^{(o)} \) are the time invariant parts of the model.

Dunson (2003) proposed a model similar to (2.5). However, Dunson (2003) does not include a vector of subject-level latent variables. That is, \( \xi_i^{(s)} = 0 \) in (2.5). Instead, only \( \nu_{im}^{(o)} \) is considered, which Dunson (2003) refers to as measurement error. Observation-level latent variables are included by Dunson (2003) as in (2.5).
2.2 Research Synthesis

Meta-analysis is a beneficial research tool for several reasons. In general, the primary goal of meta-analysis is to efficiently synthesize available evidence to inform a conclusion on a particular research question. An analysis combining multiple studies will often have increased power and improved efficiency compared to a single study. Another advantage is the ability to check for consistency across a collection of previously conducted studies (Deeks et al., 2011). More recently, meta-analysis has allowed the investigation of questions that were not originally posed by each individual study (Lumley, 2002). For example, our TBI meta-analysis will explore the effect of age at injury on the psychological and social outcomes across all ages of children by combining a study that was designed to address only preschool aged children and a study that was designed to address only school-age children. We thus intend to use meta-analysis to get more precise estimates across a wider range of ages at injury than either single study could provide. The remainder of this section is devoted to briefly reviewing meta-analytic methods.

2.2.1 Types of Data

Data for a meta-analysis can come in different forms. The two primary forms are published summary statistics and individual patient data. The analytical techniques used to synthesize the collection of studies are constrained by which type of data is available.

Meta-analysis data are often in the form of published summary statistics, or summary statistics that are extracted from previously published studies (DerSimonian
and Laird, 1986), and much of the classical meta-analysis methodology was developed for this type of data. While published summary statistics can be acquired fairly easily through a thorough literature review, there are some issues surrounding their use that require consideration. One important issue to consider is the ecological fallacy. The ecological fallacy occurs when inferences based on data for groups of individuals, or ecological data, are assumed to also apply to individuals within the groups (Piantadosi, 1988). The fallacy occurs when this assumption fails, and a study on the individual level would have reached different conclusions. Care must be taken so that the inferences and their interpretations are at the correct level of the data or extended to other levels only if there is evidence to warrant it.

An alternative to relying on published summary statistics is the use of individual patient data (IPD). However, this can be a challenge in practice since the full data on all individual patients in all studies for the meta-analysis need to be available for use. A meta-analysis with IPD is considered to be the gold standard (Simmonds et al., 2005). IPD are becoming more readily available for researchers with the mandated release of publicly funded study data and clinical trial registries, such as those organized by the National Institutes of Health (NIH). There are also many collaborative research groups that work together and share their data internally, which facilitates use of IPD. The TBI studies were conducted by a research group of this nature.

A meta-analysis using IPD is the most flexible since it is not constrained by the summary statistics that appear in a publication. This allows the meta-analysis to address research questions that may not have been addressed in original publications of the studies. IPD present the opportunity to model effects on the individual level, thus possibly preventing the ecological fallacy. The analysis can also consider additional
confounding factors that may not have been adjusted for in all the original analyses. With this added flexibility, there are some important decisions to make about the statistical method to use for the meta-analysis.

2.2.2 Methods for Summary Statistics

A meta-analysis of summary statistics is a synthesis of results from a collection of published studies on a particular topic. Typically, an effect of interest is selected, and estimates are extracted from each study. Then the estimates are statistically combined to produce an estimate of the overall average effect and its uncertainty. The form of the combination depends on assumptions about the structure of the data (and approach to estimation). Although many estimators have been proposed, the two most prevalent are based on fixed and random effects models.

Specifically, the choice of estimator reflects whether the analyst believes that means or effect sizes are homogeneous across studies. Homogeneity of means implies that all of the studies come from the same population with a common fixed mean treatment effect. An example of a simple fixed effects model for statistic $\tilde{Y}_i$ in study $i$ is:

$$\tilde{Y}_i \sim N(\mu, \sigma_i^2), \quad i = 1, \ldots, N.$$  \hspace{1cm} (2.6)

This model implies that a good estimator of $\mu$ is a weighted average where the estimate from each study is weighted by the inverse of its variance, which is assumed to be known, putting more weight on the estimates from studies with less variability and less weight on studies with more variability. As a consequence, larger studies are often given more weight than smaller studies.
A random effects model incorporates heterogeneity between studies. This is reflected in a simple random effects model for statistic $\tilde{Y}_i$ in study $i$ as:

\[
\tilde{Y}_i | \mu_i \sim N(\mu_i, \sigma_i^2) \tag{2.7}
\]
\[
\mu_i \sim N(\theta, \tau^2), \ i = 1, \ldots, N. \tag{2.8}
\]

DerSimonian and Laird (1986) extended the inverse-variance weighted average via a method of moments estimation of $\tau^2$. The weights under their procedure are the inverse of the sum of within and between study variation. In contrast to a fixed effects approach, this method typically more evenly weights large and small studies.

Meta-regression is a technique that extends the basic meta-analysis model and accounts for known differences between studies. Meta-regression tries to explain some of the variability in the summary statistics by regressing the statistics from each study on study-level covariates (Sutton and Higgins, 2008). A typical meta-regression model for a statistic from study $i$, $\tilde{Y}_i$, is

\[
\tilde{Y}_i = X_i \beta_i + \epsilon_i, \ i = 1, \ldots, N \tag{2.9}
\]

where $X_i$ is a matrix of study level covariates with effects $\beta_i$ and $\epsilon_i$ is random error. For a fixed effects model $\beta_i = \beta$ for all $i$, and for a random effects model, $\beta_i$ is given a distribution. This type of model will be further discussed in the context of longitudinal data in Section 2.2.4.

### 2.2.3 Methods for Individual Patient Data

An IPD meta-analysis can be done in either one or two stages. A two-stage meta-analysis uses the IPD to first conduct a separate analysis for each study and then combines the resulting estimates using methods described in Section 2.2.2 (Simmonds
et al., 2005). A benefit of a two-stage approach is that it uses standard methodology that has been developed for published summary statistics, while still giving control of the study level analysis to the researcher performing the meta-analysis. Another benefit is that it can accommodate a combination of published summary statistics and IPD.

A one-stage meta-analysis puts all of the IPD together in a single data set (along with study identifiers) and uses a single model (Simmonds et al., 2005). The model can be chosen as if the data were from a single large study, and can easily incorporate both fixed and random effects to reflect the structure of the data. A major advantage to the one-stage approach is that it seamlessly incorporates all of the sources of variability. In contrast, typical popular two-stage or published summary statistics procedures underestimate variability by assuming that variation within each study is known. A one-stage IPD approach was used for the preliminary meta-analysis of the TBI studies in Section 1.3.

2.2.4 Longitudinal Methods for Meta-Analysis

When data from longitudinal studies are being combined, there is additional structure in the data that should be accounted for when modeling. Unfortunately, researchers typically synthesize information from multiple longitudinal studies by simplifying the data, only considering a single time point or considering each time point separately (Jones et al., 2009). This approach is generally a product of practical considerations. Often studies only report summary measures and their uncertainty at each time point and do not report correlations between time points. This hinders the ability of a meta-analysis relying on summary statistics to properly account for
the correlation structure, and the benefits of the longitudinal design of the original studies are lost. By looking at each time point separately, a time trend cannot be modeled, and there is no way to account for baseline or other previous measurements.

Recall, the goal of research synthesis is to gain more information by combining studies, and by not considering the longitudinal nature of studies, information is actually lost. Multiple authors have shown that efficiency is lost, and incorrect inferences are sometimes drawn as a result of not accounting for the correlations that are present in longitudinal data (Ishak et al., 2007; Jones et al., 2009; Peters and Mengersen, 2008). Modeling techniques for combining longitudinal data should be able to model the correlation structure that characterizes repeated measures data.

**Published Summary Statistics**

Ishak et al. (2007) developed an approach to enable the modeling of the longitudinal structure of the data while still only relying on published summary statistics as data. It extends the meta-regression model of Section 2.2.2 by including random effects to model heterogeneity between studies. The use of random effects is natural, as random effects models, like the DerSimonian and Laird method described in Section 2.2.2, are already widely used in cross-sectional meta-analysis (Ishak et al., 2007). Additionally, the relationship between outcomes at different time points within the same study can be modeled by allowing the error terms within a study to be correlated. The general structure of the model follows that of a mixed model as described in Section 2.1.1 and is given by:

\[
\tilde{Y}_i = X_i \beta + Z_i b_i + \epsilon_i, \ i = 1, \ldots, N
\]  

(2.10)
where $\tilde{Y}_i$ is the vector of summary statistics from each of the $k$ time points from study $i$, $X_i$ is the $k \times c$ matrix of study-level covariates, which includes time, for study $i$, with vector of fixed effects $\beta$, $Z_i$ is the $k \times q$ matrix of covariates for study $i$ with vector of random effects $b_i$, and $\epsilon_i$ is the vector of random errors. The selection of study-level covariates for inclusion in the model would proceed as in any other meta-regression based on summary statistics. The vector $\epsilon_i$ is assumed to be distributed as a multivariate normal, $MVN(0, R_i)$, where $R_i$ need not be constant for each study, as discussed below. Additionally, $b_i$ is independently distributed as $MVN(0, D)$. The model is driven by the specification of $D$, $Z_i$, and $R_i$.

The between-study variation is characterized by $D$ and $Z_i$. If $Z_i$ is specified as a $k \times 1$ vector of ones, then $D$ measures the amount of between-study heterogeneity and corresponds to a random intercept model. This has a similar interpretation to the DerSimonian and Laird method described in Section 2.2.2. Another assumption to consider is whether there may be heterogeneity in the effect of time or any other study-level covariate. If so, that covariate could be included in $Z_i$, in addition to the column of ones. For example, a random effect for time suggests that there are latent differences between the trajectory of the statistic’s change over time in a particular study and the population average expected trajectory. When multiple random effects are included in the model, $D$ can be specified to allow correlation between the random effects. The interpretation of the estimated elements of $D$ follow those of any other mixed-effects model.

The within-study variation for study $i$ is described by $R_i$. In a cross sectional analysis, this corresponds to the variance of the estimate from a particular study. Unfortunately, such information on the covariance is rarely reported in publications.
If one is willing to assume that the summary statistics within a study are unrelated, then $R_i$ can be specified as a diagonal matrix where each diagonal entry corresponds to a variance reported by the published study. However, it is often plausible that the summary statistics would be correlated within a study. Since correlations on individual observations are rarely reported, assumptions would need to be made to simplify the problem to be estimable. One possible simplifying assumption would be to assume that the correlation structure is the same for all studies, and then estimate it based on the data. This can still be a challenge since each study only contributes a single vector containing statistics for each time point. A way to overcome this issue is to use IPD to directly estimate the within-study relationships at the individual level.

**IPD**

As with a cross-sectional meta-analysis, it is preferable to use IPD, rather than summary statistics, in longitudinal settings so that all of the appropriate relationships can be properly modeled and explored. A regression model for a one-step IPD meta-analysis was proposed by Jones et al. (2009) as a way to model data from longitudinal studies.

Jones et al. (2009) proposed a fixed-effects model, under which they assumed that the true vector of coefficients, $\beta$, is identical across all studies. The model for the length $k$ vector of observations, $Y_{ij}$, on individual $j$ in study $i$ is given by:

$$Y_{ij} = X_{ij}\beta + \epsilon_{ij} \tag{2.11}$$

where $X_{ij}$ is the matrix of covariates for individual $j$ in study $i$ with fixed effects $\beta$ and $\epsilon_{ij}$ is the random error. The covariate matrix should include an indicator of study membership to allow the intercepts to differ by study. Also, $\epsilon_{ij}$ is distributed
as $N(0, \Sigma_i)$. Jones et al. (2009) suggest modeling $\Sigma_i$ for each study by using a fully unstructured covariance matrix of dimension $k \times k$. They also suggest allowing $\Sigma_i$ to differ by groups within each study which allows correlations between outcomes within an individual to differ over studies and groups (Jones et al., 2009). Thus, the correlation structure in the data is modeled directly through the covariance matrix of the errors. However, it is not straightforward to characterize heterogeneity in the data using this model.

2.3 Missing Data

Missing data occur in almost every practical application using observational data. Very few studies are able to collect every variable from every participant. Missing data are often a large problem in longitudinal studies, where loss to follow-up can occur. This creates a problem because many analysis techniques require that all variables used be completely observed. However, if only fully-observed cases are considered, estimates can be biased and less efficient than if all of the fully and partially observed cases were used (Ibrahim et al., 2005). To try to minimize the negative impacts, assumptions must be made regarding the mechanisms underlying the missingness so models can be constructed to properly incorporate both fully and partially observed data, as well as the uncertainty that accompanies any missing data.

2.3.1 Types of Missing Data

Missing data are usually categorized based on the missing data pattern and mechanism. The method used for analysis will depend on the assumptions about these characteristics.
Missing data patterns describe which values are missing in a particular data set. There are a variety of missing data patterns that can be used to characterize a data set (Little and Rubin, 2002). A few that are particularly relevant to examining missing data in meta-analysis and longitudinal studies include monotone missingness, sporadic missingness, systematic missingness, and file matching missingness.
A monotone missing data pattern is where the data set can be sorted so that if a variable is missing, every subsequent variable is missing. In other words, the data vector is fully observed to a certain point and then unobserved for the rest of the vector. A monotone missing data pattern is illustrated in Figure 2.1(a). Longitudinal studies often suffer from loss to follow-up, or attrition, which is a special case of monotone missing data (Little and Rubin, 2002). This means that a subject is observed up to a certain point in time and after that point in time all of the observations for that subject are missing.

The missing data pattern that exhibits the least structure is known as sporadic missing data (Burgess et al., 2013), or a general pattern of missingness (Little and Rubin, 2002). This pattern describes missingness that occurs sporadically throughout the data set, as seen in Figure 2.1(b). In this case, where values are observed and where they are missing cannot be easily described as in the cases above. For example, this could occur if a patient refuses a particular test or refuses to answer a particular question. Sporadic missing data are the most common type of missing data and are often present in addition to a more structured primary pattern of missingness.

When combining multiple studies, missing data can occur simply because different studies may have collected different sets of variables. If a study does not collect a variable that other studies did, that variable would be missing for that study. There are two different missing data patterns that can occur in this situation. The first is referred to as systematic missing data (Resche-Rigon et al., 2013). This occurs when some studies collect a subset of the variables collected by other studies. In this case, all variables are jointly observed in at least one study since the smaller set of variables is a subset of the larger set of variables. This is illustrated in Figure 2.1(c). The other
relevant missing data pattern is a file matching pattern (Little and Rubin, 2002). This means that there are variables that were never jointly observed in any study, as is seen in Figure 2.1(d). This could occur if one set of studies used a particular measurement scale and another set of studies used a different scale to measure the same underlying trait. For example, one study may use maternal education as a measure for socioeconomic status and another study may use income, but neither study has both. Both systematic and file matching missing data are common in analyses of collections of studies since multiple studies rarely collect exactly the same set of variables. In this dissertation, we focus on systematic missing data.

Mechanisms

Missing data are generally categorized into one of three different types based on mechanism. Using notation similar to Little and Rubin (2002), let \( Y \) be a vector of the variable of interest, and let \( Y = (Y_o, Y_m) \), where \( Y_o \) are the observed values and \( Y_m \) are the missing values. Let \( M \) be a vector of indicators of whether \( Y \) was missing. So, if \( y_i \) is observed, \( m_i = 0 \); otherwise, \( m_i = 1 \), where \( y_i \) and \( m_i \) are the \( i^{th} \) elements of \( Y \) and \( M \), respectively. Let \( X \) be a matrix of any other fully-observed variables.

The first type of missingness is missing completely at random (MCAR). Data are said to be MCAR when the probability that a variable is missing does not depend on any of the observed or missing values (Little and Rubin, 2002). This means that \( P(M|Y, X, \theta) = P(M|\theta) \) for all \( Y \) and \( \theta \), where \( P(\cdot) \) is the probability mass function that depends on an unknown parameter \( \theta \). When data are MCAR, the missing data are a random sample of the full data. MCAR is a rather strong assumption that is usually unrealistic.
A less strict assumption is that the data are missing at random (MAR). This means that the probability that a value is missing can depend on observed values but not on the underlying value that is missing (Little and Rubin, 2002). That is, 
\[ P(M | Y, X, \theta) = P(M | Y_o, X, \theta), \] for all \( Y_m \) and \( \theta \). When data are MAR, a complete case analysis can be biased.

The most general form of missingness is what is known as “not missing at random” (NMAR). Data are NMAR when the chance of not observing a value depends on the true unobserved value (Little and Rubin, 2002). So, 
\[ P(M | Y, X, \theta) = P(M | Y_o, Y_m, X, \theta) \] for all \( Y_m \) and \( \theta \) cannot be further simplified. Here the probability of missingness depends on the value of the unobserved variable. A complete case analysis will typically introduce bias for this situation. When data are NMAR, a model for the missing data mechanism must be specified to make valid inferences (Little and Rubin, 2002). That is, the data are modeled as 
\[ P(Y | X, \phi) \] with unknown parameter \( \phi \) and the missingness is modeled as 
\[ P(M | Y_o, Y_m, X, \theta). \] Since a model must be specified, correct inferences will depend on how well the model for the missing data mechanism is modeling the truth. Data that are NMAR are the most difficult to deal with because the required assumptions about the missing data mechanism, 
\[ P(M | Y_o, Y_m, X, \theta), \] are untestable.

**Ignorability**

An alternative way to classify missing data mechanisms is into ignorable and non-ignorable missingness. The consequence of ignorable missingness is that the missing data mechanism can be ignored for inference about the parameters of the data model, \( \phi \). A missing data mechanism is ignorable for maximum likelihood inference if it meets
two conditions (Little and Rubin, 2002). First, the mechanism must be MAR. Second, the parameters in the data model, \( \phi \), must be distinct from the parameters in the missing data model, \( \theta \). Parameters are distinct if their joint parameter space is the product of each of the marginal parameter spaces (Little and Rubin, 2002). In a Bayesian setting, a missing data mechanism is ignorable if the mechanism is MAR and the parameters, \( \phi \) and \( \theta \), are \textit{a priori} independent (Little and Rubin, 2002). If either condition is not met, the missingness is nonignorable.

If the missingness is ignorable, we can ignore the missing data mechanism because the full likelihood function, \( L(\phi, \theta|Y_o, M, X) \), is proportional to the ignorable likelihood, \( L(\phi|Y_o, X) \), as a function of \( \phi \). Thus, likelihood-based inferences will be the same using either the full or observed data likelihood function. That is,

\[
L(\phi, \theta|Y_o, M, X) = P(Y_o, M|X, \phi, \theta) \tag{2.12}
\]
\[
= P(M|Y_o, X, \theta) \int P(Y_o, Y_m|X, \phi) dY_m \tag{2.13}
\]
\[
= P(M|Y_o, X, \theta) P(Y_o|X, \phi) \tag{2.14}
\]
\[
\propto P(Y_o|X, \phi) \tag{2.15}
\]
\[
= L(\phi|Y_o, X). \tag{2.16}
\]

This implies that as long as the missing data mechanism is ignorable, there is no need to explicitly model the missing data mechanism, \( P(M|Y_o, X, \theta) \). However, if the mechanism is nonignorable, this will not be the case, and the missing data mechanism will need to be modeled explicitly.

### 2.3.2 General Methods for Analysis With Missing Data

General methods have been developed to handle each type of missing data mechanism. A complete case analysis discards any observation that is not fully observed.
For data that are MCAR, a complete case analysis will suffice to provide unbiased inferences. However, efficiency will be lost since only the subset of complete observations are considered, and so a complete case analysis is not recommended for most situations. It is usually more efficient to use all of the available data in an analysis.

Analyses incorporating all of the available data can require the specification of models to describe the missing data. The type of model that is specified depends upon assumptions about the missing data mechanism and practical considerations. A few of the most common models will be briefly discussed here. More details and methods can be found in Little and Rubin (2002).

One potential method for analysis is to use maximum likelihood. There are a variety of computational methods for maximizing the likelihood function in the presence of missing data. One popular method is the Expectation-Maximization (EM) algorithm (Little and Rubin, 2002).

Alternatively, the missing data problem can be framed in the Bayesian paradigm. Here, all unobserved quantities – including parameters and missing values – are treated as latent or unobserved random variables. As such, each is modeled with a probability distribution (called prior distributions, in the case of parameters). The objective is typically to estimate or otherwise summarize the posterior distribution of the parameters of interest given the observed data. One computational method for accomplishing this that is convenient for the missing data problem is to use a Gibbs Sampler that samples missing values and unknown parameters from their joint posterior distribution (Ibrahim et al., 2005). Usual Bayesian inference for the parameters of interest can then be carried out.
Another widely used method is an approximation to a fully Bayesian analysis known as multiple imputation. Multiple imputation is a technique that involves using stochastic imputation to obtain plausible values for the missing data multiple times (Little and Rubin, 2002). In other words, the imputation creates multiple “completed” data sets with the missing values having been filled in by the imputation. Since the imputations are done multiple times, one can quantify the uncertainty that is associated with the imputation for the missing values. Each of these completed data sets is then analyzed separately using regular complete-data methods, such as maximum likelihood estimation. Inferences are then combined, commonly with Rubin’s Rules (Little and Rubin, 2002), to present a summary that reflects the uncertainty associated with estimating unknown parameters and the uncertainty due to missing values. A more detailed explanation of multiple imputation will be presented in Section 2.3.3.

For data that are NMAR, the joint distribution of the missingness indicator and the variable of interest (that contains missingness) is modeled. In other words, $P(Y, M)$ is modeled, where $Y$ is the variable of interest. There are two common approaches for modeling this joint probability: selection models and pattern mixture models (Little and Rubin, 2002). A selection model factors the joint probability as $P(Y, M) = P(M|Y)P(Y)$, whereas a pattern mixture model factors the joint probability as $P(Y, M) = P(Y|M)P(M)$. The choice of model formulation is often determined by which parameterization is most convenient for a particular situation. The analysis can then proceed by using maximum likelihood, Bayesian, or multiple imputation methods to estimate the parameters of interest.
2.3.3 Model-based Multiple Imputation

Multiple imputation is a flexible procedure for handling missing data. It provides an approximation to a fully-Bayesian approach, but allows analysts to fit any frequentist models that could be fit to a complete data set. There are three main steps to multiple imputation: imputation, analysis, and combining. The imputation step involves building a model to draw the imputations. The analysis step generates point estimates and standard errors using the multiply-imputed data. Then these point estimates and standard errors are combined to make inferences. It is a flexible method because the analysis and combining steps can be repeated to conduct different analyses on the same set of multiply-imputed data.

The imputation step involves imputing missing values multiple times, thus creating multiple “completed” data sets. The imputations are drawn from a posterior predictive distribution for the missing data so that averaging over the multiple imputations effectively numerically integrates over the missing values. The posterior predictive distribution is the conditional distribution for the missing variables given the observed data (Gelman et al., 2004). The posterior predictive distribution is derived from a joint Bayesian model for the data, called the imputation model. A joint model for the data jointly models both the covariates and outcomes. The goal is to propagate the uncertainty due to the missing data while preserving the associations that are present in the data. The imputation model is said to be proper if uncertainty in model parameter estimates is incorporated in the imputation procedure, i.e., by use of prior distributions for all of the parameters in the model (Little and Rubin, 2002).
After the creation of multiply-imputed data sets, a standard complete-data analysis can be conducted on each data set. The summaries from each completed data set are then combined, commonly with Rubin’s Rules (Little and Rubin, 2002). Rubin’s Rules give asymptotic approximations of the mean and variance of the posterior distribution for the model parameters, based on a sufficiently large number of imputations. In practice, White et al. (2011) suggest that the number of imputations should be at least equal to the percentage of incomplete cases (e.g., use 10 imputations for 10% incomplete cases). The estimators from Rubin’s Rules, where $\theta$ is the parameter of interest, $D$ is the number of imputations, $Y_o$ is the vector of observed data, and $\hat{V}_d$ is the complete-data posterior variance of $\theta$ for the $d^{th}$ data set, are:

$$
\hat{E}[\theta|Y_o] = \frac{1}{D} \sum_{d=1}^{D} \hat{\theta}_d = \bar{\theta} \quad (2.17)
$$

$$
\hat{Var}(\theta|Y_o) = \frac{1}{D} \sum_{d=1}^{D} \hat{V}_d + \frac{1}{D-1} \sum_{d=1}^{D} (\hat{\theta}_d - \bar{\theta})^2. \quad (2.18)
$$

Inferences are then carried out based on a normal distribution with the above mean and variance. Additional results have been derived for calculating degrees of freedom to use a $t$-distribution for inference when degrees of freedom for the complete data are small (Little and Rubin, 2002).

When building an imputation model, it is important to keep in mind the form of the desired analysis model. The analysis model attempts to describe the behavior of particular outcome variables. Therefore, when building an imputation model, the joint model should reflect these same beliefs about the outcomes. This helps to ensure that the imputation model and analysis model are coherent, or congenial. A congenial analysis model is one that can be derived from the joint distribution that is used for imputation, and the analysis model must be congenial to the imputation model.
for valid inference (Meng, 1994). For example, if both the analysis and imputation models are linear regressions, then the imputation model must contain the same set, or a more general set, of predictors as the analysis model. This becomes more of a challenge as analysis models grow in complexity. For an analysis that involves hierarchical or clustered data, the imputation model must account for the clustering and hierarchical structure that will be modeled in the analysis model. This will be of particular concern for modeling missing data in a longitudinal meta-analysis, since the data have multiple levels of hierarchy.

**Joint Methods**

The most commonly implemented joint model for multiple imputation is the multivariate normal model. The multivariate normal model jointly models all incomplete variables with a multivariate normal distribution and unstructured covariance matrix (Schafer, 1997). Let $Y_i$ be the incomplete vector of observations for subject $i$. Then the multivariate normal joint model is

$$Y_i \sim N(\mu, \Sigma).$$ (2.19)

If available, fully observed covariates can be incorporated into the model for the mean, $\mu$. Imputations are drawn from the posterior distribution for the missing values based on the above joint model. Imputed data sets are then combined using Rubin’s Rules (Little and Rubin, 2002).

When there are non-normally distributed variables, it is much more difficult to specify a joint model. In practice, researchers sometimes assume multivariate normality, even when it does not hold, to take advantage of the computational simplicity and robustness of the multivariate normal model (Schafer, 1997). This approach can
result in nonsensical values, such as negative values for distances or non-integer values for counts.

One simple approach is to impute missing integers under the multivariate normal model, and then round the imputed values to the nearest observed value. However, simple rounding of imputed values can lead to biased estimates, particularly if the true distribution is not close to symmetric or if there are high rates of missingness (Yucel et al., 2011). Various approaches have been developed to try to improve the performance of rounding methods through the use of calibrated rounding rules, instead of simple rounding (Bernaards et al., 2007; Yucel et al., 2008, 2011). In the case of binary covariates, Horton et al. (2003) recommends not rounding continuous imputations to attain unbiased estimates of the mean. However in practice, it is often undesirable to impute values that could not have been observed because an end user’s analysis may be tailored to the specific observable values, like a binary outcome in a logistic regression model. Thus there is continued interest in developing calibrated rounding approaches to impute only plausible values.

**Sequential Conditional Methods**

A sequential conditional approach to multiple imputation has been implemented under several different names, such as fully conditional imputation, sequential regression imputation, and multiple imputation by chained equations (MICE) (Raghunathan et al., 2001; van Buuren and Groothuis-Oudshoorn, 2011). Since these implementations are similar, we only discuss the MICE approach here. The MICE procedure is appealing because it provides a practical way to specify distributions for missing variables, especially if some are not normally distributed (Azur et al., 2011).
Specification of joint models becomes much more difficult in the presence of non-normal data. It is much simpler to think about univariate conditional distributions for each variable.

MICE relies on specification of a sequence of conditional distributions. In other words, a univariate conditional distribution is proposed for each variable that has missing observations. The joint imputation model in MICE is implied by the specification of the conditional distributions, however, it is never explicitly specified and is not generally theoretically guaranteed to exist (van Buuren and Groothuis-Oudshoorn, 2011). In the special case when all of the conditional distributions specify linear regressions, the sequence of conditional distributions implies a joint multivariate normal model (Raghunathan et al., 2001). In simulation studies, the MICE algorithm has been shown to work reasonably well despite no theoretical guarantee that a valid joint model exists (van Buuren and Groothuis-Oudshoorn, 2011).

The procedure is an approximation to a Gibbs Sampler. The major difference is that each step in the algorithm is now sampling from one of the conditional distributions, rather than a full conditional distribution that is derived from a joint posterior distribution. So, the MICE algorithm samples each of the $p$ elements from the vector of outcomes, $\mathbf{Y}$, from the following sequence of conditional distributions:

$$P(Y_1|\mathbf{Y}_{-1}, \theta_1^*)$$

$$\vdots$$

$$P(Y_p|\mathbf{Y}_{-p}, \theta_p^*)$$

where each distribution is conditional on all other observed values (i.e., $\mathbf{Y}_{-i}$ for observation $i$), the most recent values drawn for missing values, and the most recently
drawn model parameters, $\theta^*_j$, for each conditional density. The model parameters are each drawn from the marginal posterior distribution that is specific to the respective conditional distribution (van Buuren and Groothuis-Oudshoorn, 2011). In other words, model parameter draws are not based on a joint posterior distribution for all the model parameters. The sequence is then iteratively repeated until the draws converge, much like for a Gibbs Sampler. After convergence, the process is iterated until $D$ imputations are drawn for missing values to create $D$ multiply imputed data sets. Then the analysis is carried out on each imputed data set and combined using Rubin’s Rules, as discussed previously.

2.4 Clustered Missing Data

As with any analysis, missing data can occur when data are clustered, and missing data methods need to incorporate the clustered structure. The methods, described in Section 2.3, must be extended to properly account for the correlation structure while modeling missing data.

2.4.1 Multiple Imputation for Clustered Data

The problem of missing data in meta-analysis can be thought of as the more general problem of missing clustered data. An IPD meta-analysis can be naturally considered to be hierarchical or multi-level. For example, in the general case of a cross-sectional meta-analysis, the hierarchy consists of observations within the different studies. In a longitudinal meta-analysis, the hierarchy is extended so that there are now observations within subjects within studies. The model can be extended to handle more complex situations by adding more levels to the hierarchy.
Multiple imputation is one approach for handling missing data that are clustered. However, complications arise when a hierarchical analysis will be used to model the correlation structure, or clustering, in the data. This means that the imputation model must also be specified to account for the correlation in the data so that it remains congenial to the analysis model (Meng, 1994). For example, if a mixed model is to be used to model correlation in the analysis, then one option would be to use a mixed model for the imputation model as well. The imputation model needs to preserve all of the relationships that are to be modeled in the analysis model.

It has been shown that if the clustering in the data is not properly accounted for in the imputation model, the results from the analysis can be biased (Andridge, 2011; Reiter et al., 2006). When there are strong effects of clustering and clustering is ignored in the imputation model, variance of parameter estimates can be underestimated (Andridge, 2011), which will impact confidence interval coverage and p-values. A strategy that is often used for compatibility with standard software is to fit an imputation model that includes fixed effects for cluster membership rather than random effects. However, this imputation model is still not congenial to a mixed model. Somewhat counter-intuitively, this can cause an overestimation of the variance of the parameters in situations with weaker clustering effects because the fixed effects specification forces differences that may not actually exist (Andridge, 2011; Reiter et al., 2006).

Two approaches have been used to account for clustered data in multiple imputation. One approach is to specify a joint hierarchical model for the data and sample from the appropriate posterior predictive distributions. This is an extension of a
joint model as described in Section 2.3.3. Another approach is to extend the sequential conditional methods, as described in Section 2.3.3, to include random effects to capture the clustering. We describe each in turn.

**Joint Models**

An approach for handling clustered missing data is to specify a joint model for the data and derive the necessary posterior predictive distributions to simulate draws for multiple imputation. Schafer and Yucel (2002) developed a multiple imputation procedure based on a joint multivariate linear mixed-effects model. This procedure is appropriate for handling data that are missing at the observation level. The imputation model uses the random effects in the mixed model to account for the clustered nature of the observations.

The following multivariate mixed model, which extends what is described in Section 2.1.1, specifies the joint imputation model (Schafer and Yucel, 2002). Let \( Y_i \) be a matrix of observations for cluster \( i \) where each row represents an individual and each column represents a measurement of that individual. Suppose there are \( n_i \) individuals in cluster \( i \) and \( r \) measurements on each individual. Also let \( X_i \) be a fully observed matrix of covariates and \( \beta \) be a matrix of fixed regression coefficients. Likewise, let \( Z_i \) be a design matrix for the random effect matrix \( b_i \). Assume the rows of the error matrix, \( \epsilon_i \), are independently normally distributed. Also assume that the elements of the matrix, \( b_i \), have a multivariate normal distribution with mean \( 0 \) and covariance matrix \( \Psi \). Then the mixed-effects model that Schafer and Yucel (2002) propose for an imputation model is

\[
Y_i = X_i \beta + Z_i b_i + \epsilon_i. \tag{2.21}
\]
Conditionally conjugate inverse Wishart priors are specified for all variance parameters and improper flat priors are specified for the regression coefficients. Imputations for the missing values can then be drawn from the posterior predictive distributions to create the desired number of multiply-imputed data sets.

Yucel (2008) later extended this idea to account for missing data at any level of a three-level hierarchy. The algorithm relies on a separate Gibbs sampler for each level of the hierarchy. Yucel (2008) recommends that variables on lower levels be aggregated by taking a mean for use in the imputation of higher-level missing data. The algorithm starts by executing standard methods for non-clustered missing data on the top (most clustered) level of the hierarchy (Schafer, 1997). After imputed values are drawn for the top level, those values are carried into the middle level. They are then treated as known values and used to impute missing values on the middle level by using the methods described above (Schafer and Yucel, 2002). Again, imputed values are drawn and carried through to the bottom (observation) level as if they were observed. Imputation for missing data at the bottom level is carried out similarly to the middle level (Schafer and Yucel, 2002) by using mixed-effects models. After draws are made at the bottom level, those values are imputed, and a complete data set is formed. The procedure is then repeated until the desired number of multiply imputed data sets are formed.

It should be noted that this extension involves three distinct Gibbs Samplers. Each level of the hierarchy has its own separate sampler. Thus, for each imputed data set three Gibbs Samplers must be executed. The sampler at the next level down in the hierarchy depends on the imputed values that are sampled from the previous sampler, so each level cannot be parallelized and run simultaneously. Additionally,
this scheme does not allow information to be shared from lower to higher levels of the hierarchy. Each level is self-contained and only passes on the observed and imputed values to lower levels. The whole process is then repeated to create additional imputed data sets.

A similar, alternative approach is to specify a full joint multivariate normal hierarchical model (Liu et al., 2000; Schafer and Yucel, 2002). As above, a multivariate mixed model is specified for the variables that contain missingness at the bottom (observation) level, conditional on higher cluster level variables. Additionally, a marginal model is assumed for the subject-level variables. Standard conditionally conjugate priors are used for variance parameters and flat priors are used for means. The appropriate posterior predictive distributions can then be calculated and sampled to generate multiple imputations for missing data at either the individual or cluster level (Liu et al., 2000). This method assumes that all observation-level variables are independent given subject-level variables and random effects. Thus, correlation between variables is characterized by the correlation between the random effects.

The approaches described above are relatively straightforward when all of the data are normally distributed. However, it becomes more challenging to specify joint models when there are non-normally distributed variables. A joint modeling approach was proposed to address this issue in the presence of categorical variables (Goldstein et al., 2009). In a similar fashion to the models described in Section 2.1.3, Goldstein et al. (2009) models categorical variables using a method of generalized probit regression called the maximum indicant model (Aitchison and Bennett, 1970). This assumes a set of normally distributed latent variables underly each categorical variable for each individual. These latent variables can then be jointly modeled in a multivariate
normal model, along with any normally distributed observed variables. To impute a missing categorical variable, a set of latent variables is drawn from the posterior predictive distribution, and the category that corresponds to the largest latent variable is imputed (Goldstein et al., 2009). Thus, the categorical variable is simply a transformation of the latent vector of normally distributed latent variables.

Since all variables modeled, whether latent or observed, are normally distributed, we can examine the Goldstein et al. (2009) model as a joint multivariate normal model. Consider modeling \( p^{(o)} \) observation-level variables and \( p^{(s)} \) subject-level variables. Let \( Y_{ijm}^{(o)} \) be the \( m^{th} \) observation-level variable at time \( j \) for subject \( i \) and \( X_{i\ell}^{(s)} \) be the \( \ell^{th} \) subject-level variable. Goldstein et al. (2009) incorporates covariates in their model, but we present a simplified version for the sake of comparison with Song et al. (2011). The simplified model is as follows:

\[
\begin{align*}
Y_{ijm}^{(o)} &= \mu_{ijm}^{(o)} + \xi_{i}^{(s)} \gamma_{m}^{(o)} + e_{ijm}^{(o)} \quad (2.22) \\
X_{i\ell}^{(s)} &= \mu_{i\ell}^{(s)} + \xi_{i}^{(s)} \gamma_{\ell}^{(s)} \\
e_{ij}^{(o)} &\sim iid \sim N(0, \Omega^{(o)}) \\
\xi_{i}^{(s)} &\sim iid \sim N(0, \Omega^{(s)})
\end{align*}
\]

where \( \mu_{ijm}^{(o)} \) is the observation-level mean for variable \( m \), \( \mu_{i\ell}^{(s)} \) is the subject-level mean for variable \( \ell \), \( \xi_{i} \) is the length \( p^{(o)} + p^{(s)} \) vector of subject-level random effects, and \( e_{ij}^{(o)} \) is the vector of observation-level residuals. The observation-level residuals for different variables are correlated within an individual at a particular time point but are not correlated over time. The length \( p^{(o)} + p^{(s)} \) vector \( \gamma_{m}^{(o)} \) has an \( m^{th} \) element equal to 1 and 0 elsewhere. The length \( p^{(o)} + p^{(s)} \) vector \( \gamma_{\ell}^{(s)} \) has \( p^{(o)} + \ell^{th} \) element
equal to 1 and 0 elsewhere. Notice that there is not a residual error term in the model for $x_{it}^{(s)}$, which implicitly assumes that the residual variance is 0.

There are aspects of the Goldstein et al. (2009) model that deserve extra attention. This model explicitly models and can impute subject-level data that are categorical or normally distributed. However, Goldstein et al. (2009) is limited to use of the probit link function and generalizations of the probit link. Specifically, under the maximum indicant model, the expression for each category probability is an integral of a product of integrals (Aitchison and Bennett, 1970). Thus, there is not a common link function, and the maximum indicant model does not fit into the usual generalized linear models framework.

The Goldstein et al. (2009) model was designed for clustered data (i.e., observations within cluster). Thus, the correlation between variables within a cluster is only modeled at the cluster-level through $\Omega^{(s)}$. It was not designed for longitudinal data (i.e., subject $i$ and time $j$) so it does not accommodate observation-level correlation between variables at different time points within the same subject. That is, $e_{ij.}^{(o)} \perp e_{ij'}^{(o)}$. Thus, this approach would need to be extended for use with multilevel missing data that arise from longitudinal studies.

We now compare the observation-level model of Goldstein et al. (2009) to the joint longitudinal model proposed by Song et al. (2011). While both are fundamentally based on latent variables, (2.22) relies on correlated latent effects; whereas, (2.5) utilizes shared latent effects. Assume all variables are normally distributed and consider the following covariance expressions derived from each model.
First, consider the covariance between two observations at different time points of variable \( m \) on subject \( i \). From (2.5) using identity link functions, we have

\[
\text{Cov}(Y_{ijm}, Y_{ij'm}) = \lambda_{jm}^{(o)} \text{Cov}(\xi_{ij}^{(o)}, \xi_{ij'}^{(o)}) \lambda_{jm'}^{(o)} + \gamma_{m}^{(o)} \text{Var}(\xi_{i}^{(s)}) \gamma_{m}^{(o)} + \text{Var}(\nu_{im})
\]

(2.23)

and from (2.22), we have

\[
\text{Cov}(Y_{ijm}, Y_{ij'm}) = \gamma_{m}^{(o)} \text{Var}(\xi_{i}^{(s)}) \gamma_{m}^{(o)}.
\]

(2.24)

Comparing (2.23) and (2.24), we see additional observation-level terms in (2.23) for modeling the covariance over time. As we mentioned, Goldstein et al. (2009) does not model observation-level correlation over time and so, in (2.24), the covariance between observations at time \( j \) and \( j' \) is only captured through the variance of the subject-level latent variables \( \xi_{i}^{(s)} \).

Now consider the covariance between two observations of different variables at time \( j \) on subject \( i \). From (2.5) using identity link functions, we have

\[
\text{Cov}(Y_{ijm}^{(o)}, Y_{ij'm}^{(o)}) = \lambda_{jm}^{(o)} \text{Var}(\xi_{ij}^{(o)}) \lambda_{jm'}^{(o)} + \gamma_{m}^{(o)} \text{Var}(\xi_{i}^{(s)}) \gamma_{m'}^{(o)}
\]

(2.25)

and from (2.22), we have

\[
\text{Cov}(Y_{ijm}^{(o)}, Y_{ij'm}^{(o)}) = \gamma_{m}^{(o)} \text{Var}(\xi_{i}^{(s)}) \gamma_{m'}^{(o)} + \text{Cov}(e_{ijm}^{(o)}, e_{ij'm}^{(o)}) = \text{Cov}(e_{im}, e_{im'})
\]

(2.26)

We see that (2.25) models covariance through the variance of the latent variables. In contrast, we see that in (2.26) covariance is modeled through the covariance between pairs of latent variables. Since covariances are estimated for each pair of latent variables, this puts less structure on the covariance than (2.25). However, it also requires the estimation of a larger number of parameters, which could be problematic for some settings.
Sequential Conditional Models

The sequential conditional methods that were described in Section 2.3.3 can be extended to handle some missing data that are clustered. In this case, each univariate conditional distribution is now specified as a mixed-effects model instead of a linear model (van Buuren and Groothuis-Oudshoorn, 2011). The model would be specified as in Section 2.1.1. The addition of random effect terms to the conditional distributions model the correlation that exists within clusters. Mixed-effects models would be specified for each observation-level variable that has missing values and is considered to be clustered.

MICE, as implemented in R (van Buuren and Groothuis-Oudshoorn, 2011), can also accommodate missing data at the cluster, or subject, level. This corresponds to a variable whose value applies to every observation within the cluster. MICE handles cluster-level missing data by following the suggestion of Yucel (2008) to aggregate variables on the observation-level for use as predictors in the imputation model for the cluster-level missing variable. The procedure for cluster-level missing data uses established imputation methods for independent observations to draw imputations (Schafer, 1997).

Whereas most methods discussed thus far have been intended for sporadic missing data, MICE has been proposed as a possible method for handling systematically missing data in an IPD meta-analysis (Resche-Rigon et al., 2013). As described in Section 2.2, random effects are often used to characterize between-study heterogeneity in meta-analysis. By applying a MICE strategy for imputation, heterogeneity between studies can be incorporated in the imputation model through random effects in each
conditional distribution. This will help to preserve the hierarchical structure in the data through the imputation process.

Specifically, for each variable with missing data, the model specified for the vector of \( n_i \) observations for the \( r^{th} \) variable in study \( i \), \( Y_{ir} \) is:

\[
Y_{ir} = X_{ir}\gamma_r + Z_{ir}b_{ir} + e_{ir}
\] (2.27)

where \( X_{ir} \) is a design matrix for fixed effects \( \gamma_r \), \( Z_{ir} \) is a design matrix for random effects \( b_{ir} \), and \( e_{ir} \) is the random error (Resche-Rigon et al., 2013). As is typical for mixed-effects models, it is assumed that \( b_{ir} \sim N(0, \Psi) \) and \( e_{ir} \sim N(0, \sigma_r^2 I_{n_i}) \). The imputation procedure then uses the maximum likelihood estimates from each of the models to form the large-sample approximations to the posterior distributions for the parameters (Resche-Rigon et al., 2013). This model can also be implemented using an MCMC algorithm to sample from the posterior predictive distributions (van Buuren and Groothuis-Oudshoorn, 2011). The approach described above assumes that the systematic missing data is ignorable.

Resche-Rigon et al. (2013) conducted a simulation study to evaluate the performance of MICE for systematically missing confounders at the observation level with clustering by study. They found that MICE performed adequately, although it showed slight bias toward zero in the estimation of regression coefficients (Resche-Rigon et al., 2013). They also showed that in the presence of stronger clustering effects, MICE tended to underestimate standard errors which resulted in a reduction in confidence interval coverage (Resche-Rigon et al., 2013).

Recently, Jolani et al. (2015) generalized the approach of Resche-Rigon et al. (2013) using a sequence of generalized linear mixed effects models to impute missing values with distributions in the exponential family. Thus, the Resche-Rigon et al.
(2013) model is a special case of the Jolani et al. (2015) approach when missing variables are normally distributed and the identity link function is used.

2.4.2 Missing Data in Research Synthesis

Missing data in meta-analysis is a special case of the general problem of clustered missing data, however some approaches have been developed that do not fit into the framework discussed in Section 2.4.1. We will specifically highlight two approaches here. The first approach defines alternative models for missing data that incorporate additional sources of data to help inform the imputation of the missing values (Molitor et al., 2009). The second approach details a specialty method developed for a two-stage meta-analysis with systematically missing data (Fibrinogen Studies Commission, 2009). Since a two-stage meta-analysis is not the focus of this dissertation, we omit discussion of several other methods developed specifically for two-stage meta-analyses (Burgess et al., 2013; White et al., 2008a,b).

Combining Multiple Data Sources

An alternative approach for combining multiple sources of information with systematic block missingness is to use a Bayesian graphical model (Molitor et al., 2009). Different sources of data are used to help reduce the uncertainty due to missing data. The idea is that information from each study can be used to impute the missing details in the other studies (Molitor et al., 2009).

Each data set contains differing measurements and levels of detail. In the example used by Molitor et al. (2009), a registry provided population-based data, a survey provided detailed individual-level data, and a national study was used to provide aggregate regional estimates. This is a particularly useful approach when there are
data sources that can complement and help in the modeling of missing values that are of primary interest.

**Systematic Missingness**

If the systematically missing variables are covariates, one can consider a method that compares fully and partially adjusted model estimates (Fibrinogen Studies Commission, 2009). The goal of the analysis is to estimate the model parameters adjusted for all of the desired covariates. In other words, the interest lies in the estimate of the fully-adjusted model parameters, considering information from both the fully and partially observed studies. That is, we are interested in inference about $\beta_1^f$ in the following regression

$$E[Y_i] = \beta_1^f X_1 + \beta_2 X_2,$$

where $X_1$ is fully observed and $X_2$ is systematically missing from some subset of the studies being combined. This procedure fits two models to studies that have all of the covariates observed and one model to the studies for which only a subset of the covariates are observed. The method relies on the assumption that the partially-adjusted estimates, from the model with the subset of the covariates, will be correlated with the fully-adjusted estimates, from the model with all of the covariates. If this relationship is constant across studies, this correlation can be exploited to allow the combined fully-adjusted estimate to “borrow strength” from the partially-adjusted estimates.

The procedure follows the framework of a two-stage meta-analysis with IPD and is applicable to any type of generalized linear model or Cox regression. We will consider a simple situation where $Y_i$ is a scalar outcome for subject $i$, $X_1$ is a covariate that
is fully observed in all of the studies, and $X_2$ is a covariate that is only observed in a subset of studies. In every study, a model will be fit such that

$$E[Y_i] = \beta_1 p X_1.$$  \hspace{1cm} (2.29)

For studies for which all of the covariates are observed, the regression shown by Equation 2.28 will also be fit. Here the notation, $\beta_1 f$, indicates that the coefficients are fully adjusted (and of interest for inference) for all of the covariates, compared to $\beta_1 p$, which is only adjusted for the partial set of covariates (Fibrinogen Studies Commission, 2009). For any given study, the asymptotic normality of regression estimates and the assumption of correlation gives

$$\begin{bmatrix} \hat{\beta}_1 f \\ \hat{\beta}_1 p \end{bmatrix} \sim N \left( \begin{bmatrix} \beta_1 f \\ \beta_1 p \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \rho \sigma_1 \sigma_2 \\ \rho \sigma_1 \sigma_2 & \sigma_2^2 \end{bmatrix} \right).$$  \hspace{1cm} (2.30)

The difficulty with using this specification lies in estimating $\rho$. Estimating $\rho$ is a critical part of the analysis because the “borrowing of strength” is completely determined by the strength of the relationship between the fully and partially adjusted estimates.

The inference proceeds by maximizing the log likelihood function for the parameters. Each study with fully observed covariates will contribute both a fully and partially adjusted estimate with joint density given by Equation 2.30. Call this $f_i$ for study $i$. Each study with partially observed covariates contributes a partially-adjusted estimate with marginal density derived from the joint distribution in Equation 2.30. Call this $g_j$ for study $j$. Thus, the log likelihood function is,

$$\ell(\hat{\beta}_1 f; \hat{\beta}_1 p) = \sum_{i \in I} \log(f_i(\hat{\beta}_1 f; \hat{\beta}_1 p)) + \sum_{j \in J} \log(g_j(\hat{\beta}_1 p))$$  \hspace{1cm} (2.31)

where $I$ is the set of studies that have all of the covariates fully observed and $J$ is the set of studies that have only the partial set of covariates observed (Fibrinogen
Studies Commission, 2009). Note that the parameter of interest for inference is the fully-adjusted coefficient, $\beta_1^f$.

Implementing this procedure was found to reduce the standard errors when compared to an analysis that only used the studies with completely-observed covariates, a complete case analysis (Fibrinogen Studies Commission, 2009). This serves as evidence that the estimates of the fully-adjusted coefficients are “borrowing strength” from the partially-adjusted estimates that come from the studies in which the covariates are not fully observed (Fibrinogen Studies Commission, 2009). In other words, the correlation, $\rho$, between the fully and partially adjusted estimates is strong enough to meaningfully inform the final estimates.

2.5 Summary

In this chapter, we have reviewed the three major topic areas that will be the focus of the remainder of this dissertation. In Section 2.1, we discussed current approaches for modeling longitudinal data. We then discussed methods for research synthesis in Section 2.2. Then in Section 2.3, we described current approaches for addressing the missing data problem. In the final part of the chapter, we examined current approaches for combining these three areas. In the remainder of this dissertation, we focus on the development of methodology for conducting a longitudinal research synthesis with systematically missing subject-level data.
Chapter 3: Comparing Multiple Imputation Methods for Systematically Missing Subject-Level Data

3.1 Introduction

Even if research synthesists could minimize barriers to accessing all relevant individual patient data (and thus eliminate publication bias and reliance on possibly incomplete published statistics), overcoming the cross-study differences in the variables that were measured and when and how each was measured would remain a central challenge to individual patient data (IPD) meta-regression. For example, one study of pediatric health may use the Hollingshead Index (Hollingshead, 1975) to measure socioeconomic status, while another may rely only on the mother’s highest level of education. This issue can be formulated as a missing data problem, where values are fully observed for each variable included in a participant’s study design, but are missing for every variable not collected in that participant’s study. This pattern of missing data is known as systematic missingness (Resche-Rigon et al., 2013). Recently, already developed single-study approaches to handling such missingness patterns have served as guides to handling systematic missingness in research synthesis of cross-sectional studies (e.g. Jolani et al., 2015; Resche-Rigon et al., 2013). This chapter explores the potential usefulness of established longitudinal missing-data
methods for similar extensions in meta-analyses of longitudinal studies with systematic missingness.

Multiple imputation is a commonly used method for analyzing incomplete data across a range of settings (Carpenter and Kenward, 2013; Little and Rubin, 2002; Van Buuren, 2012). In this chapter, we compare two multiple imputation methods that have been proposed for imputing missing subject-level data in single longitudinal studies: the joint method (Liu et al., 2000) and the sequential conditional, or chained equations, method (Raghunathan et al., 2001; van Buuren and Groothuis-Oudshoorn, 2011). These two approaches to imputation are often used in practice.

Since the joint and sequential conditional methods each take a different approach to constructing imputation distributions, we first compare the imputation distributions generated by the two methods for longitudinal data with missing subject-level variables. We then examine how estimates from an analysis using each imputation procedure differ using a simulation study. Based on these comparisons, we recommend use of the joint method for multiple imputation of missing subject-level variables to preserve the associations in the data and to take advantage of the increase in power and efficiency from combining multiple studies. We expect these methods to form a more sound foundation for longitudinal IPD meta-regressions.

This chapter is organized as follows. In Section 3.2 we present details on the joint and sequential conditional methods. We examine the differences between the imputation distributions for the two methods in Section 3.3. The effect of the differences in the conditional distributions on estimation is illustrated by a simulation study in Section 3.4. In Section 3.5, we demonstrate both methods using the data described
in Chapter 1. Finally, Section 3.6 contains a discussion of the differences between the joint and sequential conditional methods.

3.2 Multiple Imputation Models

As in the TBI data example, suppose we have two studies that we wish to combine in meta-regression, with a total of $n$ subjects across the two studies. Let $Y_i = (Y_{i1}, \ldots, Y_{it})$ denote an observation-level variable that is the vector of observations at $t$ time points for subject $i$, $i = 1, \ldots, n$. In our example, the observation-level variable is the normed internalizing subscale of the Child Behavior Checklist (CBCL-I) measured in both studies at $t = 3$ time points. Let $X_i = (X_{i1}, \ldots, X_{ip})$ be a vector of $p$ subject-level covariates, with the first variable, $X_{i1}$, a binary indicator of study. In our example, $p = 6$, where age at injury, ordinal maternal education, indicators for injury severity and the Hollingshead Index comprise variables 2-6. We will assume that the first $r$ subjects are from a study that fully observed all the observation- and subject-level variables (e.g., the school-age study), and the remaining $n - r$ subjects are from a different study that only observed the first $q$ subject-level variables, where $q < p$ (e.g., the preschool study). Note that in our example, $q = 5$ fully observed variables in the preschool study, and $p = 6$ fully observed variables in the school-age study. Thus, $Y_i$ and $(X_{i1}, \ldots, X_{iq})$ are fully observed for all $i = 1, \ldots, n$, but $(X_{i(q+1)}, \ldots, X_{ip})$ is systematically missing for the second study, $i = r + 1, \ldots, n$.

While one might argue that population characteristics that would be measured by $Y$ or $X$ could influence the design choices of which variables to measure, in many cases a reasonable assumption is that the systematic missingness of $X_i$ is completely at random (MCAR), i.e., the missingness does not depend on either $X_i$ or $Y_i$ (Little and...
Rubin, 2002). In our motivating example, the choice to not calculate the Hollingshead Index in the preschool study was due to scientific trends in measures of socioeconomic status rather than any characteristics of the population.

At its core, the goal of multiple imputation is to draw repeated imputations for missing values based on the joint distribution of the missing data. To construct an imputation model, we consider the following multivariate normal joint distribution for the observations in both studies consistent with fixed-effects meta-analytical ideas:

\[
(Y_i, X_i) | \alpha, B, \mu, \Psi \sim N((\alpha + B\mu', \mu'), \Psi)
\]  

(3.1)

\[\Psi = \begin{bmatrix}
\Psi_{11} & \Psi_{21} \\
\Psi_{21} & \Psi_{22}
\end{bmatrix}\]

where \(\mu = E(X_i)\) is a length \(p\) vector, \(\alpha' + B\mu' = E(Y_i)\) is a length \(t\) vector composed of the vector of time-specific intercepts \(\alpha\), and the product of a matrix of time-specific effects of the subject-level variables, denoted \(B\), and the mean of the subject-level variables, \(\mu\). The covariance matrix \(\Psi\) is decomposed into the variance of \(Y_i\), \(\Psi_{11} (t \times t)\), variance of \(X_i\), \(\Psi_{22} (p \times p)\), and their covariance, \(\Psi_{21} (p \times t)\). Note that study membership is in truth typically not considered to be random, as it was fixed by the study designs. However, because study membership, \(X_{i1}\), is always observed and therefore never imputed, it is convenient to treat study membership the same as the other subject-level covariates.

For simplicity in presentation, we suppose that for each subject, the association between the subject-level variables, \(X_1\), and each observation-level variable, \(Y_{ij}\), does not change with time, and that the observation-level variables have already been adjusted for time. Thus, \(\alpha\) reduces to a length \(t\) vector of identical values, \(\alpha 1_t\), and
$B$ reduces to a length $p$ vector $\beta$, reducing (3.1) to:

$$(Y_i \quad X_i)' | \alpha, \beta, \mu, \Psi \sim N \left( \left((\alpha + \beta \mu')1_t \quad \mu\right)', \Psi \right), \quad (3.2)$$

where $1_t$ is a length $t$ vector of ones. These assumptions are not necessary, but greatly simplify the theoretical comparisons presented in this and the following Section. When we return to modeling the TBI data in Section 3.5, we demonstrate the inclusion of time main effects.

This parametrization of the joint distribution is convenient due to its direct connection with the generalized least squares regression model of $Y$ on $X$. We assume this or a similar regression on a subset of $X$ is the substantive (analysis) model of interest. Under this joint model, $B$ is the matrix of (or $\beta$ is the vector of time-invariant) regression coefficients obtained from this generalized least squares regression model, i.e., the effect of $X_i$ on $Y_i$. It will also be convenient to use $\Sigma$ to denote the covariance matrix of $Y_i$ adjusted for $X_i$ (i.e., the residuals of the longitudinal regression of $Y_i$ on $X_i$), so that

$$\Psi_{11} = 1_t' \beta \Psi_{22} \beta' 1_t' + \Sigma \quad (3.3)$$

$$\Psi_{21} = \Psi_{22} \beta' 1_t'.$$

Similarly, one could decompose the subject-level portion of the model into one sub-model for the fully observed variables $(X_{i1}, \ldots, X_{iq})$ and a regression sub-model for the conditional distribution of the systematically missing variables $(X_{i(q+1)}, \ldots, X_{ip})$ given these fully observed variables. This second regression sub-model clearly demonstrates the unidentifiability in any model with a fixed effect of study. That is, the data cannot contain information about the effect of the study indicator $X_{i1}$ on the missing variables $(X_{i(q+1)}, \ldots, X_{ip})$, since these variables are entirely unobserved for one level.
of the study indicator. A number of solutions to this unidentifiability problem are possible. We take the simplest approach, which is to set this coefficient to zero. That is, we effectively make the strong assumption that the association between the missing variables \(X_i(q+1), \ldots, X_{ip}\) and the fully observed variables \((X_{i2}, \ldots, X_{iq})\) is identical for both studies. Although this assumption is consistent with our assumption of MCAR, it is likely to be strong in practice, and Section 3.6 describes extensions that could soften it in situations where more studies are available.

While this model is motivated by combing two studies in a meta-analysis, we note that the model described above would also apply to a single study with \(n\) subjects in which \(X_i\) is missing for \(r\) subjects. As mentioned above, in the context of meta-analysis, we assume that the systematic missingness of \(X_i\) is completely at random (MCAR). In the context of a single study, we might imagine more complex mechanisms such as missingness dependent on the fully observed \(Y_i\) (missing at random, MAR) or missingness dependent on the \(X_i\) values themselves (missing not at random, MNAR) (Little and Rubin, 2002). The two imputation models that we focus on, the joint method and the sequential conditional method, would be appropriate in the single-study setting either under MCAR or MAR, but not under MNAR.

The joint method directly models the multivariate normal distribution of all the subject- and observation-level variables shown in (3.2) by specifying a full Bayesian model for the data (Liu et al., 2000). For this Bayesian approach, the selection of prior distributions plays an important role in the model, particularly for variance parameters. Analysts often choose the inverse Wishart distribution to be the prior distribution for \(\Sigma\) because of the computational convenience of conjugacy. Following this approach, Schafer and Yucel (2002) recommend setting the mean of the inverse
Wishart distribution to a prior guess for $\Sigma$ or setting it equal to a diagonal matrix whose elements are empirical variance estimates. An alternative prior distribution for $\Sigma$ proposed by Daniels and Pourahmadi (2002) is based on a variation of the Cholesky decomposition of $\Sigma$ and is intuitively parameterized for longitudinal data. For scalar variance parameters, an inverse gamma prior is often used (Liu et al., 2000). For parameters associated with means, Liu et al. (2000) and Schafer and Yucel (2002) recommend using improper flat priors; alternatively, one could use multivariate normal priors as suggested by Gelman et al. (2004).

The most general version of the joint method based on Equation 3.1 and an inverse Wishart prior distribution for the covariance matrix is easy to implement for the special case of the normal model where the study indicator is not included as an explanatory variable. In this case, the joint imputation model is an available option in many software packages, such as SAS’s PROC MI (MCMC option) or Stata’s mi command (mvn option). WinBUGS or JAGS could be used to implement our simple identifiable version of the joint imputation model (where the coefficient of the regression effect of the study indicator on the missing variables is set to zero), with the more flexible Daniels and Pourahmadi (2002) prior distribution for the covariance matrix $\Sigma$. Because our prior distributions are conjugate, we found it straightforward to create our own Gibbs sampler algorithm in R. It is more difficult to construct joint distributions and MCMC algorithms when the data include non-normally distributed variables, and we are not aware of any software packages that implement joint models using alternative distributional assumptions.

The sequential conditional method is based on the specification of a sequence of univariate conditional distributions for each missing variable (Raghunathan et al.,
2001; van Buuren and Groothuis-Oudshoorn, 2011). Each distribution is sampled in sequence, acting as an approximation to a Gibbs sampler and thus an approximation to sampling from the joint distribution in (3.2). If the data have a multivariate normal distribution, then the sequential conditional method is equivalent to the joint method, when the conditional distribution for each variable conditions on all other variables (Carpenter and Kenward, 2013). Thus for (3.2), applying the sequential conditional method to impute the $j^{th}$ element of $X_i$ would require using the conditional distribution of $X_{ij}$ conditional on $X_{i(-j)}$ and $Y_i$. However, with multiple observation-level variables each measured at multiple times, the number of variables to condition on will quickly grow unmanageable. To combat this, Yucel (2008) recommends using the mean of each observation-level variable in the imputation model, and this is what is implemented in the multiple imputation by chained equations (MICE) R package (van Buuren and Groothuis-Oudshoorn, 2011). Using this approach, the conditional distribution based on (3.2) that would be used to impute the $j^{th}$ missing element of $X_i$ at the $k^{th}$ iteration is:

$$X_{ij}^{(k)*} = \gamma_0 j + \gamma_1 j (X_{i(-j)}^{(k-1)*} \ Y_i)' + \xi_{ij}$$

(3.4)

where $X_{i(-j)}^{(k-1)*}$ is the length $p - 1$ vector $X_i^{(k-1)*}$ of imputed values at the $(k-1)^{th}$ iteration of $X_i$ excluding the $j^{th}$ entry, $\overline{Y}_i$ is the scalar mean of the elements of vector $Y_i$, $\gamma_1 j$ is a length $t + p - 1$ vector of regression coefficients, and $\gamma_0 j$ is a scalar intercept term. The distributions defined in (3.4) are sampled in sequence from $j = 1, \ldots, p$ as necessary for partially observed subject-level variables, and repeated for the desired number of iterations. In addition to the MICE R package, the sequential conditional
method is available in many other software packages, including SAS’s PROC MI (FCS option) and Stata’s mi command (chained option).

Because, as noted above, one can construct the most general sequential conditional and joint methods to be equivalent under an assumption of multivariate normality, the default sequential conditional method with a fixed study effect suffers from the same lack of identifiability as the full joint model. Naïve use of the MICE R package executes seamlessly and does not produce any warnings regarding this identifiability issue. Intuitively, the sequential algorithm would result in a similar MCAR-based imputation with no study main effect in the conditional imputation model for the systematically missing variables (i.e., \( \gamma_{1j1} = 0 \) for \( j = q + 1, \ldots, p \)).

### 3.3 Comparisons Between Imputation Distributions

Regardless of which method is used, the imputation of missing subject-level variables (e.g., the Hollingsead Index) is based on the conditional distribution for those variables given the others (i.e., \( P(\mathbf{X}_i|\mathbf{Y}_i) \)). The sequential conditional method models this conditional distribution directly, whereas the conditional distribution is implied by the joint model for the data in the joint method. The differences between the two approaches are important to understand because they lead to differences in the estimation of the association between \( \mathbf{X}_i \) and the elements of \( \mathbf{Y}_i \).

#### 3.3.1 Imputation Distributions

The sequential conditional method that uses \( \mathbf{Y}_i \) rather than the vector \( \mathbf{Y}_i \) as a predictor for \( X_i \) will result in a different conditional distribution for \( X_i \) than that implied by the joint method. For the normal case that we are considering, closed form imputation distributions for each method can be derived and compared. In this and
the following section, we highlight differences between the two imputation methods by focusing on the special case of a single subject-level variable $X$, so that $p = 1$ and $q = 0$. This exploration may directly relate to the case of a single longitudinal study, or a meta-regression that does not incorporate study as an explanatory variable. That is, for the theoretical and simulation-based comparisons between the imputation methods, we treat the two studies as if they were one study. This is well-known to be poor practice, and so we discuss the generalization of our results to more sound meta-regression models in Section 3.6. In this simplified situation, it is relatively easy to compare the mean and variance for the subject-level variable’s imputation models under the two approaches.

**Joint Method**

The imputation distribution based on the joint method for the missing subject-level variable given the observation-level variables (e.g., for the Hollingshead Index given the CBCL-I measurements) is derived from (3.2) as:

$$X_i | Y_i, \beta, \mu, \Psi \sim N(\mu_{X_i|Y_i}, \eta_{X_i|Y_i}^2) \quad (3.5)$$

$$\mu_{X_i|Y_i} = \mu + \Psi_{21} \Psi_{11}^{-1} (Y_i - (\alpha + \beta \mu) 1_t)$$

$$\eta_{X_i|Y}^2 = \eta^2 - \Psi_{21} \Psi_{11}^{-1} \Psi'_{21}$$

where $\eta^2 = \Psi_{22}$ is a scalar, since we are considering only one subject-level variable. Imputations of missing $X_i$ are drawn from this conditional distribution for the joint method via a full Gibbs sampler.
Sequential Conditional Method

The sequential conditional method we consider directly uses the conditional distribution in (3.4) for imputation of missing $X_i$, which can be rewritten for a single subject-level variable:

$$X_i | \mathcal{Y}_i, \beta, \mu, \Sigma, \eta^2 \sim N(\mu_{X_i | \mathcal{Y}_i}, \eta^2_{X_i | \mathcal{Y}_i})$$  \hspace{1cm} (3.6)

$$\mu_{X_i | \mathcal{Y}_i} = \mu + \frac{1}{n} \beta \eta^2 \left( \beta^2 \eta^2 + \frac{1}{n^2} \mathbf{1}_t' \Sigma \mathbf{1}_t \right)^{-1} \mathbf{1}_t' (Y_i - (\alpha + \beta \mu) \mathbf{1}_t)$$

$$\eta^2_{X_i | \mathcal{Y}_i} = \eta^2 - \beta^2 \eta^4 \left( \beta^2 \eta^2 + \frac{1}{n^2} \mathbf{1}_t' \Sigma \mathbf{1}_t \right)^{-1}.$$  \hspace{1cm} (3.7)

3.3.2 Comparing Means

One way to compare the imputation distributions is to examine the difference between their means, $\mu_{X_i | \mathcal{Y}_i} - \mu_{X_i | \mathcal{Y}_i}$. We find

$$\mu_{X_i | \mathcal{Y}_i} - \mu_{X_i | \mathcal{Y}_i} = \left( \Psi_{21} \Psi_{11}^{-1} - \frac{1}{n} \beta \eta^2 \left( \beta^2 \eta^2 + \frac{1}{n^2} \mathbf{1}_t' \Sigma \mathbf{1}_t \right)^{-1} \right) \mathbf{1}_t' (Y_i - (\alpha + \beta \mu) \mathbf{1}_t)$$

$$= \mathbf{D} (Y_i - (\alpha + \beta \mu) \mathbf{1}_t).$$  \hspace{1cm} (3.7)

This difference is a product of two terms. The second term is the vector of centered values of $Y_i$, $(Y_i - (\alpha + \beta \mu) \mathbf{1}_t)$. The first term is the vector of differences, $\mathbf{D}$, that compares the weight on each element of $Y_i$ in the computation of the conditional mean for the two methods. Since we condition on $Y_i$, the vector of centered values is the same for both methods. This implies that the difference between the means for the two imputation distributions is driven by the differences in the weights put on each observation. That is, the difference between $\Psi_{21} \Psi_{11}^{-1}$ and $\frac{1}{n} \beta \eta^2 \left( \beta^2 \eta^2 + \frac{1}{n^2} \mathbf{1}_t' \Sigma \mathbf{1}_t \right)^{-1} \mathbf{1}_t'$.

We gain additional insight into the difference between the conditional means by assuming $t = 3$ timepoints (as in the TBI example) and thus reducing the number of
parameters of $\Sigma$:

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho_1\sigma_1\sigma_2 & \rho_3\sigma_1\sigma_3 \\ \rho_1\sigma_1\sigma_2 & \sigma_2^2 & \rho_2\sigma_2\sigma_3 \\ \rho_3\sigma_1\sigma_3 & \rho_2\sigma_2\sigma_3 & \sigma_3^2 \end{bmatrix}, \quad (3.8)$$

To further simplify the expression for comparison, we assume $\rho_1 \geq \rho_2 \geq \rho_3 \geq 0$ and homogenous variance $\sigma^2 = \sigma_1^2 = \sigma_2^2 = \sigma_3^2$. In addition, assuming that the elements of the vector of centered $Y_i$ are similar, we use the sum of the elements of $D$ to characterize the difference between the means as a function of the parameters of $\Sigma$.

We denote this sum $\Delta_{\mu}$:

$$\Delta_{\mu} = D1_3 = \frac{2\beta\eta^2\sigma^2 C}{AB} \quad (3.9)$$

$$A = \sigma^2(2\rho_1 + 2\rho_2 + 2\rho_3 + 3) + 9\beta^2\eta^2$$

$$B = \sigma^2(-\rho_1^2 + 2\rho_1\rho_2\rho_3 - \rho_2^2 - \rho_3^2 + 1) + \beta^2\eta^2(-\rho_1^2 + 2\rho_1\rho_2 + 2\rho_1\rho_3$$

$$- 2\rho_1 - \rho_2^2 + 2\rho_2\rho_3 - 2\rho_2 - \rho_3^2 - 2\rho_3 + 3)$$

$$C = \rho_1^2(-\rho_1 + \rho_2 + \rho_3 + 1) + \rho_2^2(\rho_1 - \rho_2 + \rho_3 + 1) + \rho_3^2(\rho_1 + \rho_2 - \rho_3 + 1)$$

$$- \rho_1\rho_2 - \rho_1\rho_3 - \rho_2\rho_3 - 3\rho_1\rho_2\rho_3.$$

By examining derivatives of $\Delta_{\mu}$, we find that the absolute value of the difference increases with the differences between the correlation parameters. (Relatedly, the distance also increases with the largest correlation, $\rho_1$, and decreases with the smallest correlation, $\rho_3$.) The absolute value of the difference is thus smallest when the correlations are all equivalent, as with an exchangeable correlation structure. In fact, the joint and sequential methods are identical if the correlation structure is exchangeable. This equivalence intuitively makes sense because by regressing on $Y_i$, the sequential conditional method is always implicitly giving each element of $Y_i$ the
same weight, \( \left( \frac{1}{n^2} \beta \eta^2 \left( \beta^2 \eta^2 + \frac{1}{n^2} \mathbf{I}_t \Sigma_i \right)^{-1} \right) \). We note that this finding is not new; Carpenter and Kenward (2013, §9.3) derive the conditional distribution of a subject-level variable given an observation-level variable under an exchangeability assumption and show that it depends only on the mean of the observation-level variable. They do not, however, examine situations other than exchangeability.

By examining \( \Delta \mu \), we see that the sign of \( \Delta \mu \) is the same as the sign for \( \beta \). Thus, \( \Delta \mu > 0 \) for \( \beta > 0 \) and \( \Delta \mu < 0 \) for \( \beta < 0 \). This has implications on the association between the elements of \( Y_i \) and \( X_i \) because the conditional mean from the sequential conditional method is closer to the unconditional mean than its joint counterpart. Thus, we would expect the sequential conditional method to underestimate the correlation between elements of \( Y_i \) and \( X_i \). This implies that estimates of \( \beta \) would be attenuated from an analysis using the sequential conditional method because the association between \( X_i \) and \( Y_i \) is underestimated in the imputation model.

### 3.3.3 Comparing Variances

In addition to the possible difference in the conditional means between the two methods, we also find that the conditional variances may differ. The difference in conditional variances works out to be

\[
\eta_{X|Y}^2 - \eta_{X|Y}^2 = -\beta \eta^2 \Delta \mu.
\] (3.10)

Since (3.10) is a linear function of (3.9), the difference between conditional variances will be large when the difference between conditional means is also large. Based on the differences shown in (3.7) and (3.10), we expect the performance of the sequential conditional and joint methods to differ for correlation structures that are not exchangeable. While an exchangeable structure is often appropriate for repeated
measures at a single time point, longitudinal observations often exhibit correlations related to the proximity of the measurements in time. We demonstrate the differential performance of the two imputation methods for more typical longitudinal correlation matrices via simulation.

3.4 Simulation to Assess Impact on Estimation

Since the primary goal of our analysis is to estimate the regression coefficient, $\beta$, we conduct a simulation study to examine how differences in the imputation distributions for $X_i$ impact estimation. We expect the estimates from the sequential conditional method that uses $Y_i$ as a predictor to be attenuated because the imputation is based on a conditional distribution for $X_i$ that underestimates the association between $X_i$ and $Y_i$. We confirm this expectation with our simulation study.

3.4.1 Data Generation

We generate 1000 independent data sets that each contain $n = 50$ subjects. The data sets are simulated according to the model described in (3.2), with $t = 3$ repeated observations of $Y_i$ and a scalar $X_i$ (i.e., $p = 1$). This resembles the structure of our example TBI data with three longitudinal measures of CBCL-I and Hollingshead Index calculated at baseline. We set $\alpha = 0$, $\beta = 2$, $\mu = 5$, and $\Psi_{22} = \eta^2 = 4$. To loosely mimic a two study research synthesis, systematically missing data are created by deleting the value of $X_i$ for half of the subjects that are selected by a missing completely at random (MCAR) mechanism. We do not vary $\alpha$, $\beta$, $\mu$, and $\eta^2$ but instead consider different longitudinal covariance structures for $\Sigma$.

We first consider covariance structures where the variance is constant across time points with the covariance matrix parameterized as in (3.8). We consider five different
Table 3.1: Values of the correlation and variance parameters of $\Sigma$, as parameterized in (3.8), for the 7 settings examined in the simulation study. “Extreme” indicates the parameter values that maximize $\Delta \mu$ and thus are expected to show the worst performance for the sequential conditional method.

correlation structures, with the parameter values for $\Sigma$ shown in Table 3.1. Setting 1 maximizes $\Delta \mu$ for a covariance matrix with a Toeplitz structure, i.e., a diagonal-constant matrix such that pairs of observations with the same temporal proximity have the same correlation (i.e., for $t = 3$, $\rho_1 = \rho_2$). Setting 2 maximizes $\Delta \mu$ for an unstructured covariance matrix. These two settings represent extreme departures from exchangeability and thus are “worst case” scenarios for the sequential conditional method. Setting 5 is an exchangeable (compound-symmetric) structure, i.e., $\rho_1 = \rho_2 = \rho_3$. Settings 3 and 4 are more moderate structures that fall between Settings 1 and 2 and Setting 5.

We also compare the model performance for situations with nonconstant variance, also shown in Table 3.1. We use preliminary estimates from the TBI study data to define the variance parameters. Setting 6 corresponds to an exchangeable correlation of 0.7 between all time points, and Setting 7 has an unstructured correlation matrix with correlations estimated from a previous analysis of the TBI data.
3.4.2 Imputation and Analysis Procedures

Each simulated data set is analyzed with four different procedures: the two imputation methods presented in Section 3.2, a complete case analysis, and an analysis based on the data prior to deletion of missing values (“before deletion”). We use the complete case and before deletion analyses as standards for comparison. The analysis model for all methods is the generalized least squares (GLS) regression model that follows from (3.2), fit using the `gls()` function in the `nlme` package of R. With the exception of the Toeplitz structures (Settings 1, 3), the variance-covariance structure in the GLS model is specified to correspond with the data generation model (e.g., the analysis model for Setting 5 assumes exchangeability and constant variance). A Toeplitz structure is not an option in the `gls` function, and so for these cases the analysis model assumes an unstructured correlation.

For imputation, we implement the joint method and the sequential conditional method as described in Section 3.2. For both of these methods we use 50 imputations, following the advice of White et al. (2011) who suggest that the number of imputations should be at least equal to the percentage of incomplete cases (50% in our simulation). Imputation results are combined using Rubin’s Rules with the small sample degrees of freedom approximation (Little and Rubin, 2002).

To implement the joint method we must specify prior distributions. The prior distributions for \( \sigma_1^2, \sigma_2^2, \sigma_3^2 \), and \( \eta^2 \) are independent inverse gamma distributions, as parameterized in Gelman et al. (2004), with both parameters set to 0.005 (so that the corresponding gamma distribution has mean 1). The prior distributions for \( \alpha, \beta \), and \( \mu \) are uniform on the real line. We use the prior distribution for \( \Sigma \) suggested by
Daniels and Pourahmadi (2002), which is based on the decomposition,

$$T \Sigma T' = D,$$

where $T$ is a lower triangular matrix with 1s on the diagonal and $i, j^{th}$ entry $-\phi_{ij}$ and $D$ is a diagonal matrix with $i^{th}$ entry $\sigma_i^2$. The prior distributions for $\phi_{21}$, $\phi_{31}$, and $\phi_{32}$ are uniform on the real line. In initial simulations we also considered the commonly-used Inverse Wishart distribution as the prior for $\Sigma$; its performance was similar to that of the Daniels and Pourahmadi prior except for the most extreme unstructured covariance structure (Setting 2) where associations between $X$ and $Y$ were slightly attenuated (results not shown).

The sequential conditional method is implemented through the MICE R package (van Buuren and Groothuis-Oudshoorn, 2011). We use the default prior distribution implemented in the package, which is the improper limiting form of the normal-inverse Wishart distribution (Schafer, 1997).

### 3.4.3 Performance Metrics

We compare the performance of each method for estimating $\beta$. We evaluate the performance of estimation based on the empirical mean of $\hat{\beta}$, the empirical mean of the model standard error of $\hat{\beta}$, the empirical standard error of $\hat{\beta}$, and the empirical root mean squared error for estimation of $\beta$. In addition, we calculate the mean 95% confidence interval width and the observed confidence interval coverage, based on symmetric intervals for $\beta$ constructed using estimates from applying Rubin’s Rules and the critical $t$-value with degrees of freedom estimated by the small sample approximation (Little and Rubin, 2002). Finally, we compare the relative estimated efficiencies of the methods for estimating $\beta$. 
Figure 3.1: 95% confidence interval average width and coverage for the analysis with data before deletion (○), complete case (□), sequential conditional method (♦), and joint method (△) with normally distributed systematically missing subject-level covariate data. Setting number is indicated by plotting symbol. Results are from 1000 independent replications. Points falling between the dashed lines are within Monte Carlo error of nominal coverage.

### 3.4.4 Results

Tables 3.2 and 3.3 show the performance metrics for estimation of $\beta$ with each method. Figure 3.1 illustrates the average confidence interval width and coverage for each method.

From Table 3.2, we see the sequential conditional method is practically identical to the joint method for Setting 5, which has constant variance and an exchangeable correlation structure. However, in general, we prefer the joint method over the sequential conditional method, and the strength of that preference depends on the underlying covariance structure of the data. As the covariance structure departs from homogenous variance and exchangeable correlation, the joint method becomes the
<table>
<thead>
<tr>
<th>Setting</th>
<th>N</th>
<th>Mean</th>
<th>Model SE</th>
<th>Empirical SE</th>
<th>RMSE</th>
<th>CI Width</th>
<th>CI Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 - Toeplitz, extreme</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Before Deletion</td>
<td>50</td>
<td>2.00</td>
<td>0.03</td>
<td>0.03</td>
<td>0.10</td>
<td>0.94</td>
<td></td>
</tr>
<tr>
<td>Complete Case</td>
<td>25</td>
<td>2.00</td>
<td>0.04</td>
<td>0.04</td>
<td>0.14</td>
<td>0.90</td>
<td></td>
</tr>
<tr>
<td>MICE</td>
<td>50</td>
<td>1.83</td>
<td>0.19</td>
<td>0.14</td>
<td>0.22</td>
<td>0.77</td>
<td>0.93</td>
</tr>
<tr>
<td>Joint</td>
<td>50</td>
<td>2.00</td>
<td>0.04</td>
<td>0.04</td>
<td>0.16</td>
<td>0.94</td>
<td></td>
</tr>
<tr>
<td>2 - Unstructured, extreme</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Before Deletion</td>
<td>50</td>
<td>2.00</td>
<td>0.07</td>
<td>0.07</td>
<td>0.27</td>
<td>0.92</td>
<td></td>
</tr>
<tr>
<td>Complete Case</td>
<td>25</td>
<td>2.00</td>
<td>0.10</td>
<td>0.10</td>
<td>0.38</td>
<td>0.92</td>
<td></td>
</tr>
<tr>
<td>MICE</td>
<td>50</td>
<td>1.87</td>
<td>0.19</td>
<td>0.15</td>
<td>0.20</td>
<td>0.78</td>
<td>0.96</td>
</tr>
<tr>
<td>Joint</td>
<td>50</td>
<td>2.01</td>
<td>0.10</td>
<td>0.10</td>
<td>0.41</td>
<td>0.94</td>
<td></td>
</tr>
<tr>
<td>3 - Toeplitz, moderate</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Before Deletion</td>
<td>50</td>
<td>2.01</td>
<td>0.16</td>
<td>0.16</td>
<td>0.61</td>
<td>0.95</td>
<td></td>
</tr>
<tr>
<td>Complete Case</td>
<td>25</td>
<td>2.00</td>
<td>0.22</td>
<td>0.24</td>
<td>0.88</td>
<td>0.92</td>
<td></td>
</tr>
<tr>
<td>MICE</td>
<td>50</td>
<td>1.95</td>
<td>0.23</td>
<td>0.22</td>
<td>0.91</td>
<td>0.95</td>
<td></td>
</tr>
<tr>
<td>Joint</td>
<td>50</td>
<td>2.01</td>
<td>0.21</td>
<td>0.21</td>
<td>0.83</td>
<td>0.94</td>
<td></td>
</tr>
<tr>
<td>4 - Unstructured, moderate</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Before Deletion</td>
<td>50</td>
<td>2.00</td>
<td>0.14</td>
<td>0.14</td>
<td>0.53</td>
<td>0.94</td>
<td></td>
</tr>
<tr>
<td>Complete Case</td>
<td>25</td>
<td>2.00</td>
<td>0.20</td>
<td>0.21</td>
<td>0.77</td>
<td>0.92</td>
<td></td>
</tr>
<tr>
<td>MICE</td>
<td>50</td>
<td>1.95</td>
<td>0.21</td>
<td>0.20</td>
<td>0.83</td>
<td>0.95</td>
<td></td>
</tr>
<tr>
<td>Joint</td>
<td>50</td>
<td>2.00</td>
<td>0.19</td>
<td>0.19</td>
<td>0.75</td>
<td>0.95</td>
<td></td>
</tr>
<tr>
<td>5 - Exchangeable</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Before Deletion</td>
<td>50</td>
<td>2.01</td>
<td>0.17</td>
<td>0.16</td>
<td>0.65</td>
<td>0.94</td>
<td></td>
</tr>
<tr>
<td>Complete Case</td>
<td>25</td>
<td>2.02</td>
<td>0.24</td>
<td>0.25</td>
<td>0.95</td>
<td>0.94</td>
<td></td>
</tr>
<tr>
<td>MICE</td>
<td>50</td>
<td>2.01</td>
<td>0.23</td>
<td>0.23</td>
<td>0.91</td>
<td>0.95</td>
<td></td>
</tr>
<tr>
<td>Joint</td>
<td>50</td>
<td>2.02</td>
<td>0.22</td>
<td>0.23</td>
<td>0.89</td>
<td>0.95</td>
<td></td>
</tr>
<tr>
<td>6 - Nonconstant Variance Exchangeable</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Before Deletion</td>
<td>50</td>
<td>1.98</td>
<td>0.52</td>
<td>0.55</td>
<td>2.05</td>
<td>0.94</td>
<td></td>
</tr>
<tr>
<td>Complete Case</td>
<td>25</td>
<td>1.95</td>
<td>0.76</td>
<td>0.84</td>
<td>2.97</td>
<td>0.91</td>
<td></td>
</tr>
<tr>
<td>MICE</td>
<td>50</td>
<td>1.81</td>
<td>0.74</td>
<td>0.75</td>
<td>2.97</td>
<td>0.94</td>
<td></td>
</tr>
<tr>
<td>Joint</td>
<td>50</td>
<td>1.95</td>
<td>0.73</td>
<td>0.78</td>
<td>2.92</td>
<td>0.92</td>
<td></td>
</tr>
<tr>
<td>7 - Nonconstant Variance Unstructured</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Before Deletion</td>
<td>50</td>
<td>1.98</td>
<td>0.46</td>
<td>0.48</td>
<td>1.79</td>
<td>0.94</td>
<td></td>
</tr>
<tr>
<td>Complete Case</td>
<td>25</td>
<td>1.98</td>
<td>0.69</td>
<td>0.78</td>
<td>2.69</td>
<td>0.91</td>
<td></td>
</tr>
<tr>
<td>MICE</td>
<td>50</td>
<td>1.84</td>
<td>0.69</td>
<td>0.71</td>
<td>2.75</td>
<td>0.95</td>
<td></td>
</tr>
<tr>
<td>Joint</td>
<td>50</td>
<td>1.99</td>
<td>0.67</td>
<td>0.72</td>
<td>2.68</td>
<td>0.93</td>
<td></td>
</tr>
</tbody>
</table>

Table 3.2: Simulation results comparing performance for estimating $\beta$ with normally distributed systematically missing subject-level covariate data. Results are averaged over 1000 independent replications. Correlation and variance parameter values for each setting are shown in Table 3.1.
more preferred approach. For example, the differences in the mean estimate of $\beta$ and the average model standard error from the two methods are smaller in Settings 3 and 4 than Settings 1 and 2 because Settings 3 and 4 have covariance matrices “closer” to constant variance and exchangeable correlation.

As we depart from homogenous variance and exchangeable correlation, the mean point estimate for $\beta$ is attenuated for the sequential conditional method (Table 3.2), which implies that the association between $X_i$ and the elements of $Y_i$ is underestimated. The confidence interval coverage remains reasonable because the standard errors tend to be overestimated (Figure 3.1), often even larger than those from the complete case analysis (Table 3.3). We observe this pattern especially in Settings 1 and 2.

In contrast, the mean estimate of $\beta$ from the joint method tends to be very slightly amplified (Table 3.2). The confidence interval coverage is reasonable despite a slight attenuation of the standard error estimates, particularly in Setting 7. In general, the joint method has smaller bias than the sequential conditional method and has smaller standard errors (Table 3.3) and thus narrower confidence intervals (with comparable coverage), as shown in Figure 3.1. Thus, we prefer to use the joint method over the sequential conditional method for the settings that we considered.

3.5 Comparison of Results for TBI Data Example

In this section, we demonstrate the use of the complete case, sequential conditional and joint methods in the real world setting of the TBI data described in Chapter 1. Recall that the normed internalizing subscale of the Child Behavior Checklist (CBCL-I) is the observation-level outcome of interest, which was measured at three
Sequential Conditional  Complete Case  Joint

| Setting 1 - Toeplitz, extreme | 5.27  | 0.21  |
| Setting 2 - Unstructured, extreme | 2.00  | 0.53  |
| Setting 3 - Toeplitz, moderate | 1.01  | 0.92  |
| Setting 4 - Unstructured, moderate | 1.05  | 0.90  |
| Setting 5 - Exchangeable | 0.93  | 0.98  |
| Setting 6 - Nonconstant Variance Exchangeable | 0.98  | 0.97  |
| Setting 7 - Nonconstant Variance Unstructured | 1.00  | 0.98  |

Table 3.3: Relative estimated efficiencies comparing average model SE($\hat{\beta}_1$) from an analysis with normally distributed systematically missing subject-level covariate data. Results are averaged over 1000 independent replications.

time points. Our goal in this example analysis is to estimate its association with household-level socioeconomic status (SES), as measured at baseline by the Hollingshead Index, controlling for other family characteristics. Because the Hollingshead Index was not measured in the preschool study, it is a systematically missing subject-level variable. As our focus is on the systematically missing variable, we follow the lead of Resche-Rigon et al. (2013) in deleting all observations with sporadic missingness so that all the variables are fully observed, except for the Hollingshead Index for the preschool subjects. After multiply imputing the Hollingshead Index using the two methods described in Section 3.2, we estimate the Hollingshead Index coefficient of the generalized least squares regression of CBCL-I on the Hollingshead Index, study indicator, age at injury, maternal education level, injury severity, and time post-injury, where the covariance matrix for the three CBCL-I measurements is fully unstructured. For comparison, we also fit this model using a complete case analysis.
<table>
<thead>
<tr>
<th>Method</th>
<th>$\sigma_1^2$</th>
<th>$\sigma_2^2$</th>
<th>$\sigma_3^2$</th>
<th>$\rho_1$</th>
<th>$\rho_2$</th>
<th>$\rho_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complete Case</td>
<td>126</td>
<td>107</td>
<td>101</td>
<td>0.57</td>
<td>0.71</td>
<td>0.60</td>
</tr>
<tr>
<td>MICE</td>
<td>109</td>
<td>104</td>
<td>91</td>
<td>0.60</td>
<td>0.67</td>
<td>0.56</td>
</tr>
<tr>
<td>Joint</td>
<td>109</td>
<td>104</td>
<td>91</td>
<td>0.60</td>
<td>0.67</td>
<td>0.56</td>
</tr>
</tbody>
</table>

Table 3.4: Estimates of the parameters of the error covariance matrix $\Sigma$, as parameterized in (3.8), for the CBCL-I regression model fit to the TBI data. Estimates that rely on the sequential conditional and joint imputation missing data methods are based on 50 imputations.

Estimates of the error covariance matrix, $\Sigma$, shown in Table 3.4 indicate that our example is closest to our simulation setting 7, with nonconstant variance and an unstructured correlation matrix. As such, we would expect the sequential conditional method to produce a slightly attenuated coefficient estimate with a standard error slightly larger than that estimated via the joint method. The estimates, standard errors, and 95% confidence intervals based on each missing data method, shown in Table 3.5, are consistent with this expectation. We note that Monte Carlo error in the estimates based on multiple imputation is of the same magnitude as the differences between the two methods. Thus, as predicted by the simulation study, the sequential conditional method performs at least as well as the joint method.

Both imputation methods result in smaller standard errors than the complete case analysis, which is based only on the school-age study. This is natural because the imputation methods nearly double the number of subjects used in the analysis. However, the standard error for the imputation methods are not a factor of $\sqrt{2}$ smaller than the complete case standard error, because the missing Hollingshead Index for one entire study means that these extra observations provide less information than the observations with complete data. However, both imputation methods rely on
<table>
<thead>
<tr>
<th>Missing Data Method</th>
<th>N</th>
<th>Estimate</th>
<th>Model SE</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complete Case</td>
<td>149</td>
<td>-0.124</td>
<td>0.074</td>
<td>(-0.268, 0.021)</td>
</tr>
<tr>
<td>Sequential Conditional</td>
<td>290</td>
<td>-0.108</td>
<td>0.065</td>
<td>(-0.236, 0.021)</td>
</tr>
<tr>
<td>Joint</td>
<td>290</td>
<td>-0.117</td>
<td>0.065</td>
<td>(-0.246, 0.011)</td>
</tr>
</tbody>
</table>

Table 3.5: Estimates, standard errors, and confidence intervals for the effect of Hollingshead Index of socioeconomic status on Child Behavior Checklist Internalizing score for the TBI data. Estimates are adjusted for study, age at injury, maternal education, injury severity, and time post-injury. Each model assumes a fully unstructured covariance matrix. Estimates that rely on the sequential conditional and joint imputation missing data methods are based on 50 imputations.

the strong assumption that the relationships between the Hollingshead Index and the other family characteristics do not differ across the studies. If this assumption is violated, both imputation standard errors would under-estimate the true variability in the estimated effect of SES on CBCL-I, and thus the difference between the imputation-based and complete case standard errors would be too large. With only two available studies, this assumption is difficult to assess. The possibly slightly attenuated imputation-based effect estimates may (if not a reflection of Monte Carlo error) imply that the distribution of the CBCL-I differs across the two studies, but this finding does not automatically also imply differences in the Hollingshead Index distribution.

In sum, we find in this one example analysis that the properties of the missing data methods demonstrated in the simulation study extend to more complex general linear models. In the example, we did not find a typical longitudinal correlation pattern that allowed us to choose a restricted covariance matrix. We demonstrated the use of the two imputation methods for a model for the time-specific CBCL-I means that allowed a time trend and controlled for other possibly important subject-level variables.
3.6 Discussion

In this chapter, we compared the sequential conditional and joint methods for multiple imputation of systematically missing subject-level covariates in the context of a generalized least squares regression analysis synthesizing two longitudinal studies via a fixed effects model. Based on examination of the expressions for the imputation distributions in a much simplified version of the meta-analysis problem, we found that the imputation distributions for the two methods differ unless there is constant variance across time points and the correlation structure is exchangeable. In particular, the association between missing subject-level variables $X_i$ and the elements of the observation-level variable $Y_i$ is attenuated in the imputation distribution derived under the sequential conditional method that uses the mean of the observation-level variable as a predictor.

Via simulation with similar simplified models, we confirmed that the regression coefficient estimates from the sequential conditional method were attenuated in all settings except for when there was constant variance and an exchangeable correlation structure. In contrast, the estimates from the joint method were nearly unbiased for each setting considered. In addition, the estimates from the sequential conditional method were less efficient than those from the joint method, which led to wider confidence intervals.

These findings in the simplified cases were confirmed in our analysis of the TBI example. Because the model suggested that the error covariance matrix was not extremely different from exchangeable, we expected the two imputation-based methods to be comparable. Our analyses confirmed this, and further demonstrated the potential efficiency advantages of either method over the complete case approach that does...
not use all available studies. While this example does not demonstrate a case where the joint method is superior, it does confirm that the joint method is practical to implement and the conservative choice without prior knowledge of the error covariance structure.

Some may argue that the sequential conditional method is adequate for all error covariances because the empirical confidence interval coverage is approximately at the nominal level. However, the average confidence interval is wider because of inflated estimates of the model standard error, which are often even larger than those from the complete case analysis. In the context of research synthesis, the complete case analysis represents an analysis where we do not synthesize multiple studies but instead only analyze the single study that is fully observed. These results suggest that we should not bother to use the sequential conditional method based on variable means to combine studies because it is, at best, no more efficient than the complete case analysis. However, because we know there are often benefits associated with combining studies, we recommend using the joint method when missing subject-level data are present to preserve both the associations in the data and the increase in power and efficiency from including additional studies.

Our results for the sequential conditional method are limited to the default implementation in the R MICE package that uses the sample mean of the elements of $Y_i$ as a predictor in the conditional distribution for $X_i$. In our particular example, we could have instead treated each observation-level measurement as a separate subject-level variable and included all $t$ elements of $Y_i$ as predictors in the conditional distribution of $X_i$. However, this approach suffers from significant drawbacks in the longitudinal setting. The conditional models for the subject-level variables quickly become
unmanageable as the model grows in complexity. The number of predictors in the conditional distribution of $X_i$ can outnumber the number of subjects as $t$ grows and as the number of observation-level variables grows. In addition, it is not clear how to incorporate observations that are not taken at consistent times under a sequential conditional approach. For these reasons, we did not consider this difficult-to-extend method in our comparisons.

This comparison highlights an important, well-known difference between the joint and sequential conditional methods. While the sequential conditional method is appealing because the analyst can consider only one variable at a time, it is difficult to impose consistent simplifying constraints that make model estimation feasible, such as simplified mean models for some variables, or a Toeplitz or AR1 covariance matrix for longitudinal measurements. In addition, the naïve use of sequential conditional methods may quietly impose identifiability constraints that are not obvious or desirable. In particular, for more complex patterns of systematic missingness that are likely to appear in larger collections of studies, identifiability concerns may be even more difficult to discern. In contrast, although the implementation of the joint method is not as simple, logical and consistent constraints are easily imposed and identifiability issues more simply identified and openly addressed.

Our results are also limited in their direct applicability to many cases of IPD meta-regression. In particular, we relied on fixed-effects models that are practical in our real setting with limited data, but are not typically considered to be best practice when more studies are available. In particular, study-based random intercepts and random effects are often employed to reflect the across-study variation that is not explained by the patient profile and modeled study characteristics. Additionally,
changing from fixed to random study intercepts and effects may have the technical
benefit of eliminating model unidentifiability associated with the fixed study effects.
In effect, by assuming the study effects are similar (according to some distribution), we
may be able to estimate this similarity from multiple studies where the systematically
missing variables are observed and under an MAR assumption. The implementation
of such a model is not possible with only two studies and therefore was not developed
for this chapter. Nonetheless, we believe that the underlying shortcomings of the
sequential conditional imputation method implemented in this chapter will persist
with the addition of random-effects features often used in IPD meta-regression. As
Equations 3.9 and 3.10 suggest, this method does not capture nonexchangeable cor-
relations between or differing variances across observation-level measurements on the
same individual in the conditional distribution of the systematically missing subject-
level variables. These are limitations to modeling each individual’s data that we
expect to persist regardless of the method chosen to model across-individual asso-
ciations, such as fixed- versus random-effects meta-regression. We expect that the
joint method of imputation would still be preferred over the implemented sequential
conditional method for longitudinal meta-regressions that include enough studies to
follow the usual recommendations for the inclusion of random effects in the analysis
model.

We chose to primarily compare the two imputation methods in the extremely
simple case of a single observation-level variable (measured t times), a single subject-
level variable, and a single pattern of systematic missingness. We were thus able
to highlight important differences without the distractions that such complications
would introduce. However, both examined imputation methods are easily extended
to these more complex cases. As partly demonstrated in the example that included multiple subject-level variables, the joint model described in Equation 3.2 simply grows to accommodate the extra variables. For the sequential conditional method, the number of full conditional distributions to be specified grows with the number of systematically missing variables. As is typical, the usefulness of both methods may be limited by a large number of variables relative to the number of individuals and/or longitudinal observations.

More complex missing data patterns (e.g., multiple systematic patterns across multiple studies) pose no insurmountable implementation issues to either imputation method, but may introduce more complicated unidentifiability for models that include fixed study effects. In the sequential conditional case, we must carefully determine that the identifiability assumptions imposed by the collection of conditional distributions are reasonable. For the joint imputation model, we must be careful to impose sufficiently strong prior distributions to construct a well-behaved sampling algorithm. This is especially true for patterns in which there is no overlap (e.g., $X_{i2}$ and $X_{i3}$ are never jointly observed in the same study). The development of approaches to determine reasonable assumptions or prior distributions for such patterns is an open question, but will likely be more transparent under the joint model paradigm.

Beyond issues of common assumptions in algorithm implementation, we see no theoretical reason that the larger bias or reduced efficiency of the sequential method relative to the joint method seen in the simple case would vanish or change direction simply by adding variables or other complexities to the model, unless these changes happened to result in conditional independence between the missing subject-level and the observation-level variables. This exception seems unlikely to occur in practice.
While naturally related, our study is unlikely to be applicable to extensions for more general multivariate response vectors. We presented our examination in the context of a meta-regression with a longitudinal outcome variable, and we took advantage of the longitudinal structure to explore common longitudinal correlation structures. In theory, any vector of variables $Y$ could take the place of the repeated measurements, as long as an appropriate covariance matrix is preserved. However, in this case, few analysts would knowingly implement the default sequential conditional procedure that uses the mean of this arbitrary vector $Y$ as an explanatory variable.

Via direct examination of imputation distributions and simulation, we were able to understand important differences between a joint method and the default implementation of the sequential conditional method for imputing missing subject-level variables in longitudinal meta-regression. While we limited our study to a case with few variables and no random effects, we expect the joint method to continue to outperform the sequential conditional method for more complex hierarchies and dependence structures. As more longitudinal and dependent data become available for synthesis, it will be important to use methodology that can preserve the structure of the data while accounting for systematically missing variables.

In this chapter, we only examined the case where all of the variables are normally distributed. In practice, we know that this is unlikely to be the situation. A shortcoming of the joint method is that it can be quite difficult to specify a joint model when variables are not all normally distributed. This is one reason that the sequential conditional approach is popular because non-normality is much easier to accommodate through the use of generalized linear models. Since we prefer the joint method for imputation, we would like to develop a joint model within the generalized
linear models framework that can be used to impute variables with distributions in the exponential family.
4.1 Introduction

In Chapter 3, we concluded that a joint imputation model should be used for multiple imputation of systematically missing normally distributed subject-level variables. However, the joint modeling approach described in Chapter 3 becomes difficult to specify when variables are not normally distributed. In this chapter, we develop a two-level joint modeling approach for multiple imputation based on the generalized linear models framework that incorporates an autoregressive approach for capturing serial dependence.

A two-level joint modeling approach has been previously proposed by Goldstein et al. (2009), as described in Section 2.4.1. They developed a multivariate normal model to jointly model subject- and observation-level data that are either normally distributed or able to be modeled with generalized probit models and established its use for imputation. Like Goldstein et al. (2009), our model jointly models variables on the subject- and observation-levels, and we build our model for use in multiple imputation. However, the method of Goldstein et al. (2009) does not account for serial dependence. We develop our model in the broader generalized linear models
framework so we are not limited to use of the probit link function and incorporate serial dependence. We do this by extending the work of Song et al. (2011) and Dunson (2003), which are both introduced in Section 2.1.3.

This chapter is organized as follows. Section 4.2 describes our proposed two-level joint model. We present a simulation study to compare the performance of our model to existing imputation approaches for systematically missing subject-level variables with different distributions in Section 4.3. Section 4.4 demonstrates our model using the TBI data described in Chapter 1. We suggest a possible approach for choosing the number of latent variables in the model in Section 4.5. Finally, Section 4.6 contains a discussion of our findings.

4.2 Joint Two-Level Model

Suppose we are interested in jointly modeling \( p = p^{(o)} + p^{(s)} \) variables that are observed for \( n \) subjects, where the superscripts refer to the observation- and subject-level, respectively. Observation-level variables are measured at \( t \) time points, and subject-level variables are measured once for each subject \( i \). We let \( Y_{ijm}^{(o)} \) represent observation-level variable \( m \) for subject \( i \) at time \( j \), where \( i = 1, \ldots, n, \ j = 1, \ldots, t, \) and \( m = 1, \ldots, p^{(o)} \). We denote subject-level variable \( \ell \) for subject \( i \) as \( X_{i\ell}^{(s)} \), where \( \ell = 1, \ldots, p^{(s)} \). We assume that \( Y_{ijm}^{(o)} \) has a distribution from the exponential family with canonical parameter \( \theta_{ijm}^{(o)} \) and scale parameter \( \sigma_{jm}^{(o)} \), if necessary. We also assume that \( X_{i\ell}^{(s)} \) has a distribution from the exponential family with canonical parameter \( \theta_{i\ell}^{(s)} \) and scale parameter \( \sigma_{\ell}^{(s)} \), if necessary. We consider several examples for implementation of our joint model with normally, Bernoulli, Poisson, and multinomial distributed variables in Section 4.3.3.
We model the associations among the observed variables via shared latent variables, $\xi$. Thus, our model assumes independence of all observed variables conditional on the latent variables. This includes conditional independence of repeated measurements of the same variable on a subject over time.

A simple case of our model is shown in Figure 4.1 as a path diagram, where observed variables are in boxes and latent variables are in circles. In this simple scenario, there are three observation-level variables ($Y_{ij1}^{(o)}$, $Y_{ij2}^{(o)}$, $Y_{ij3}^{(o)}$), observed at three time points, and one subject-level variable ($X_{i1}^{(s)}$). In addition, the model has one observation-level latent variable and two subject-level latent variables. Each observation-level observed variable also has a subject-level latent variable that is specific to observed variable $m$ ($\nu_{i1}^{(o)}$, $\nu_{i2}^{(o)}$, $\nu_{i3}^{(o)}$). Variance and scale parameters are represented by self loops. Arrows connecting model components represent associations.

Associations among observed variables are modeled by $\xi$, which is shown in Figure 4.1 by arrows connecting the latent variables to the observed variables. Conditional independence of the observed variables given the latent variables is represented in Figure 4.1 by the lack of arrows connecting the observed variables to one another.

### 4.2.1 Latent Variables

As in Song et al. (2011), our joint model is based on two vectors of latent variables: a length $q^{(o)}$ vector of observation-level latent variables for each time $j$, $\xi_{ij}^{(o)}$, and a length $q^{(s)}$ vector of subject-level latent variables, $\xi_{i}^{(s)}$. Note that $q^{(o)}$ need not be equal to the number of observation-level variables, $p^{(o)}$, and $q^{(s)}$ need not be equal to the number of subject-level variables, $p^{(s)}$. The observation-level latent variables represent
Figure 4.1: Diagram of the model defined in Section 4.2 where $p^{(o)} = 3$, $p^{(s)} = 1$, $q^{(o)} = 1$, $q^{(s)} = 2$. Boxes represent observed variables and circles represent latent variables.
unobservable characteristics of a subject that vary over time, and the subject-level latent variables represent unobservable characteristics that are time invariant. We assume \( \xi_i^{(s)} \) is independent of \( \xi_i^{(o)} \) for all \( j \).

For the observation-level latent variables, we assume the following model to capture serial dependence over time:

\[
(\xi_{i1}^{(o)}, \ldots, \xi_{it}^{(o)})' \overset{iid}{\sim} N(0, \Psi^{(o)}), \quad i = 1, \ldots, n,
\]

where \( \Psi^{(o)} \) has dimension \( tq^{(o)} \times tq^{(o)} \). In the style of Dunson (2003) and Daniels and Pourahmadi (2002), we adopt a dynamic approach for modeling the covariance matrix, \( \Psi^{(o)} \), of all the vectors of observation-level latent variables across time points for a particular individual. The model for the vector of latent variables for individual \( i \) at time \( j \) is given by:

\[
\xi_{ij}^{(o)} = \sum_{k=0}^{j-1} \xi_{ik}^{(o)} \phi_{jk}^{(o)} + \delta_{ij}^{(o)},
\]

where \( \xi_{i0}^{(o)} = 0, \phi_{jk}^{(o)} \) is a length \( q^{(o)} \) vector of autoregressive parameters with \( \phi_{j0}^{(o)} = 0 \), and errors \( \delta_{ij}^{(o)} \overset{iid}{\sim} N(0, \psi_j^{(o)}) \) where \( \psi_j^{(o)} \) is diagonal with dimension \( q^{(o)} \times q^{(o)} \). Thus, following Dunson (2003) and Song et al. (2011), we assume that all dependence among latent variables at the current time is captured by the values of the latent variables at the previous time points. This approach is particularly sensible for longitudinal data which naturally progresses over time.

The graphical representation of the model for the observation-level latent variables is shown in the bottom half of Figure 4.1. We see arrows pointing from the observation-level latent variables at earlier time points to those at later time points (i.e., arrow from \( \xi_{i1}^{(o)} \) to \( \xi_{i2}^{(o)} \)). This represents the dependence of observation-level latent variables at later time points on those at earlier time points.
For the subject-level latent variables, we assume

$$\xi_i^{(s)} \sim iid \, N(0, \Psi^{(s)}), \quad i = 1, \ldots, n,$$

(4.3)

where $\Psi^{(s)}$ is of dimension $q^{(s)} \times q^{(s)}$. We can also model $\Psi^{(s)}$ using the decomposition defined by Daniels and Pourahmadi (2002) or $\Psi^{(s)}$ can be estimated as in any typical multivariate normal analysis. We use the decomposition in our modeling as follows for the $r^{th}$ element of $\xi_i^{(s)}$:

$$\xi_{ir}^{(s)} = \sum_{k=0}^{r-1} \xi_{ik}^{(s)} \phi_{rk}^{(s)} + \delta_{ir}^{(s)}$$

(4.4)

where $\xi_{i0}^{(s)} = 0$, $\phi_{rk}^{(s)}$ is a regression coefficient with $\phi_{r0}^{(s)} = 0$, and errors $\delta_{ir}^{(s)} \sim N(0, \psi_{ir}^{2(s)})$. This is shown graphically in the top half of Figure 4.1 by the arrow connecting $\xi_{i1}^{(s)}$ to $\xi_{i2}^{(s)}$, where the second subscript represents the first and second element of $\xi_i^{(s)}$ respectively.

### 4.2.2 Observation-Level Sub-Model

First, we consider the sub-model for the observation-level variables, $Y_{ijm}^{(o)}$. We use the model from Song et al. (2011) for each canonical parameter of the distributions of the observation-level variables. That is,

$$h_m^{(o)}(\theta_{ijm}^{(o)}) = \mu_{jm}^{(o)} + \xi_{ij}^{(o)} \lambda_{jm}^{(o)} + \xi_i^{(s)} \gamma_m^{(o)} + \nu_{im}^{(o)}, \quad m = 1, \ldots, p^{(o)},$$

(4.5)

$$\nu_{im}^{(o)} \sim N(0, \omega_{m}^{2})$$

where $h_m^{(o)}(\cdot)$ is a link function and $\mu_{jm}^{(o)}$ is a fixed intercept at time $j$ specific to observation-level variable $m$. Subject-level characteristics are related to $Y_{ijm}^{(o)}$ through the non-time varying subject-level latent variables, $\xi_i^{(s)}$, and the associated length $q^{(s)}$ vector of factor loadings, $\gamma_m^{(o)}$. There is also a latent subject and variable specific
intercept, $\nu_{im}^{(o)}$. Thus, $\xi_{ij}^{(o)}$, $\gamma_{jm}^{(o)}$, and $\nu_{im}^{(o)}$ are the time invariant part of the observation-level model. Observation-level characteristics are related to $Y_{ijm}$ through the time-varying vector of latent variables, $\xi_{ij}^{(o)}$, and the associated length $q_{ijm}^{(o)}$ vector of factor loadings, $\lambda_{jm}^{(o)}$. Thus, $\mu_{jm}^{(o)}$, $\xi_{ij}^{(o)}$, and $\lambda_{jm}^{(o)}$ all vary over time.

Graphically, the observation-level sub-model is shown in Figure 4.1 by arrows connecting latent variables to the observed observation-level variables. We see both subject-level latent variables, $\xi_{i1}^{(s)}$ and $\xi_{i2}^{(s)}$, have arrows connecting them to each of the observation-level variables at each of the three time points. We also see the subject-level variable specific latent variable, $\nu_{im}^{(o)}$, is only connected to the three observations of observation-level variable $m$. The observation-level latent variable, $\xi_{ij}^{(o)}$, is connected to each observed observation-level variable at time $j$.

### 4.2.3 Subject-Level Sub-Model

Next, we define the sub-model for the subject-level variables, $X_{i\ell}^{(s)}$, and discuss its connection to (4.5). Our extension of Song et al. (2011) is to include subject-level variables by sharing the subject-level latent variables, $\xi_i^{(s)}$, across models for the observation- and subject-level variables. We define the model for the canonical parameter of the distribution of the subject-level variables as

$$h_\ell^{(s)}(\theta_i^{(s)}) = \mu_\ell^{(s)} + \xi_i^{(s)} \gamma_\ell^{(s)}, \quad \ell = 1, \ldots, p^{(s)},$$

(4.6)

where $h_\ell^{(s)}(\cdot)$ is a link function and $\mu_\ell^{(s)}$ is a fixed intercept specific to subject-level variable $\ell$. The vector of latent variables, $\xi_i^{(s)}$, is related to observed subject-level variables through the corresponding length $q_{ij}^{(s)}$ vectors of factor loadings, $\gamma_\ell^{(s)}$. This is shown graphically in Figure 4.1 by arrows connecting the subject-level latent variables $(\xi_{i1}^{(s)}, \xi_{i2}^{(s)})$ to the observed subject-level variable ($X_{i1}^{(s)}$).
The subject- and observation-level sub-models are connected by the vector of subject-level latent variables, \( \xi_{i}^{(s)} \). The subject-level latent variables appear in both the model for subject-level variables, (4.6), and the time-invariant part of the model for observation-level variables, (4.5). This is clearly illustrated in Figure 4.1 by arrows connecting the subject-level latent variables to every observed variable. The sharing of the latent variables, scaled by their associated subject- and observation-level factor loadings, model the correlation between the subject- and observation-level variables through the unobservable time-invariant characteristics of each individual.

### 4.2.4 Example

We consider a simple example to illustrate how to specify our joint model in practice. Suppose that we have \( p^{(o)} \) normally distributed variables on the observation-level measured at \( t \) time points and a normal and a binary variable on the subject-level \( (p^{(s)} = 2) \). Our observation-level sub-model using the identity link function for all variables is

\[
Y_{ijm}^{(o)} | \theta_{ijm}^{(o)}, \sigma_{jm}^{2(o)} \overset{\text{ind}}{\sim} N(\theta_{ijm}^{(o)}, \sigma_{jm}^{2(o)}), \ j = 1, \ldots, t, \ m = 1, \ldots, p^{(o)} \tag{4.7}
\]

\[
\theta_{ijm}^{(o)} = \mu_{jm}^{(o)} + \xi_{i}^{(o)} \lambda_{jm}^{(o)} + \xi_{i}^{(s)} \gamma_{m}^{(o)} + \nu_{im}^{(o)}.
\]

On the subject-level, we have a normally distributed variable (\( \ell = 1 \)) and a binary variable (\( \ell = 2 \)). We use the identity link function for the normally distributed variable and the logistic link function for the binary variable. Then the subject-level sub-model is

\[
X_{i1} | \theta_{i1}^{(s)}, \sigma_{1}^{2(s)} \overset{\text{ind}}{\sim} N(\theta_{i1}^{(s)}, \sigma_{1}^{2(s)}) \tag{4.8}
\]

\[
\theta_{i1}^{(s)} = \mu_{i}^{(s)} + \xi_{i}^{(s)} \gamma_{1}^{(s)}
\]
As this simple example illustrates, our joint method easily fits into the generalized linear models framework. Thus it is applicable to a wide variety of practical modeling situations since it fits into this flexible and well established framework. More detailed examples are considered in Section 4.3.3.

### 4.2.5 Congeniality

An important consideration when constructing an imputation model is that of congeniality. An analysis model is said to be congenial if it can be derived from the joint model used for imputation (Meng, 1994). This ensures that the imputation model and analysis model reflect the same beliefs about the structure and behavior of the outcome variable being modeled and lead to valid inference (Meng, 1994). Thus, it is generally desirable to build an imputation model that is congenial to the analysis model.

First, we consider the types of analysis models that might be of interest. Commonly, researchers examining longitudinal outcomes are interested in fitting regression models that incorporate the repeated measures structure of the data. Generally, this is done by using a mixed effects or generalized least squares regression model. The fixed effects of such models are estimated conditional on the random effects and error structure. This requires the ability to separate the covariance attributable to repeated measurements on the same subject (correlated residual errors) from covariance that is due to correlation with other variables. Correlation with other variables is captured

\[
X_{i2}\mid \theta^{(s)}_{i2} \overset{ind}{\sim} \text{Bern}(\theta^{(s)}_{i2})
\]

\[
\text{logit}(\theta^{(s)}_{i2}) = \mu^{(s)} + \xi^{(s)} \gamma^{(s)}.
\]
through the fixed effects portion of the regression model, while the correlation due to repeated measurements is captured in the error and random effect structure.

When all variables are normally distributed, the joint model in Section 4.2 specifies a multivariate normal distribution. However, in general, it may not be congenial to the regression models described above. It may not be congenial because, under the joint model, we are unable to separate the variability due to other variables and variability due to repeated measurements. In the joint model, all observation-level correlation, both due to other variables and repeated measurements, is modeled by $\xi^{(o)}_i$. This does not allow us to separate the sources of the observation-level correlation in the joint model. Thus, in general, we cannot derive the regression analysis model from the joint model since we cannot condition on the correlation structure of the residual error.

When the variables are not all normally distributed, the joint model no longer specifies a multivariate normal distribution. In general, we cannot easily compute the conditional distribution for the outcome of interest. Thus, the joint model may not be congenial with the regression models of interest when non-normally distributed variables are incorporated into the model.

Despite the importance and desirability of congeniality, as models incorporate non-normally distributed variables and grow in complexity, it becomes ever more difficult to specify a congenial imputation model. In many applications, a practical imputation model is needed to address complex missing data problems where it may not be possible to specify a congenial model. Despite theoretical shortcomings of uncongenial imputation models, they can often adequately address the problem at hand while remaining practical computationally (Carpenter and Kenward, 2013, §4.3).
We address the adequacy of our joint model in spite of the possible uncongeniality via simulation in Section 4.3.

4.2.6 Prior Distributions

We elect to take a Bayesian approach to fitting our joint model so we must specify prior distributions for all of the parameters in the model. The prior distributions for the regression parameters are:

\[
\begin{align*}
\mu_{jm}^{(o) \text{ ind}} &\sim N(0, \eta_{jm}^2), \\
\mu_{\ell}^{(s) \text{ ind}} &\sim N(0, \eta_{\ell}^2), \\
\lambda_{jm}^{(o) \text{ ind}} &\sim N(0, T_{jm}^{(o)}), \\
\gamma_{m}^{(o) \text{ ind}} &\sim N(0, \Upsilon_{m}^{(o)}), \\
\gamma_{\ell}^{(s) \text{ ind}} &\sim N(0, \Upsilon_{\ell}^{(s)}), \\
\phi_{jk}^{(o) \text{ ind}} &\sim N(0, \zeta_{jk}^{(o)}), k < j \quad k < r \\
\phi_{rk}^{(s) \text{ ind}} &\sim N(0, \zeta_{rk}^{2(s)}), k < r
\end{align*}
\]

where \( T_{jm}^{(o)}, \Upsilon_{m}^{(o)}, \Upsilon_{\ell}^{(s)}, \) and \( \zeta_{jk} \) are diagonal matrices. Following Gelman (2006), we use uniform priors on the standard deviation for scalar variance and scale parameters. That is,

\[
\begin{align*}
\sigma_{jm}^{(o) \text{ ind}} &\sim U(0, \infty), \\
\sigma_{\ell}^{(s) \text{ ind}} &\sim U(0, \infty), \\
\omega_{m}^{\text{ ind}} &\sim U(0, \infty), \\
\psi_{jk}^{(o) \text{ ind}} &\sim U(0, \infty), \\
\psi_{rk}^{(s) \text{ ind}} &\sim U(0, \infty),
\end{align*}
\]
where $\psi^{(o)}_{jk}$ is the square root of the $k^{th}$ element of $\psi^{(o)}_j$. Hyperparameters can be chosen to give weakly informative prior distributions or to reflect prior knowledge.

4.3 Simulation Study to Assess Use in Multiple Imputation

In Section 4.2, we described a joint model for data on the subject- and observation-levels that have distributions in the exponential family. However, since our joint model may not be congenial with commonly used regression analysis models, in this section we examine the properties associated with using our joint model for imputation via simulation when the joint model is not the data generating model. Thus, we generate data according to the structure of a commonly used hierarchical model that is compatible with a regression analysis model. We then create systematically missing data and impute the missing data with our joint model. We compare the performance of our joint model for estimating regression coefficients in the analysis model to two other common approaches for multiple imputation: a joint model that assumes multivariate normality and a sequential conditional approach.

4.3.1 Data Generation

We generate 1000 independent data sets that each contain 400 subjects and three repeated observations per subject, $j = 1, 2, 3$. We simulate three observation-level variables ($p^{(o)} = 3$), $m = 1, 2, 3$, and one subject-level variable ($p^{(s)} = 1$), $\ell = 1$, according to the following hierarchical model:

\begin{align*}
Y_{ij1} &= \beta_{01} + \beta_{11}X_{i1} + \beta_{21}Y_{ij2} + \beta_{31}Y_{ij3} + e_{ij1} \\
Y_{ij2} &= \beta_{02} + \beta_{12}X_{i1} + \beta_{22}Y_{ij3} + e_{ij2} \\
Y_{ij3} &= \beta_{03} + \beta_{13}X_{i1} + e_{ij3}
\end{align*}  

(4.9)
Table 4.1: Values of the correlation and variance parameters of $\Sigma$, parameterized as in (4.10), for the settings examined in the simulation study.

<table>
<thead>
<tr>
<th>Setting</th>
<th>Description</th>
<th>$\sigma_1^2$</th>
<th>$\sigma_2^2$</th>
<th>$\sigma_3^2$</th>
<th>$\rho_1$</th>
<th>$\rho_2$</th>
<th>$\rho_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>Toeplitz</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.85</td>
<td>0.85</td>
<td>0.55</td>
</tr>
<tr>
<td>4</td>
<td>Unstructured</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.80</td>
<td>0.60</td>
<td>0.20</td>
</tr>
<tr>
<td>5</td>
<td>Exchangeable</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.70</td>
<td>0.70</td>
<td>0.70</td>
</tr>
<tr>
<td>6</td>
<td>Nonconstant Exchangeable</td>
<td>1</td>
<td>1.25$^2$</td>
<td>1.5$^2$</td>
<td>0.70</td>
<td>0.70</td>
<td>0.70</td>
</tr>
</tbody>
</table>

We use a subset of the correlation structures considered in Chapter 3. However, we use different variance values for this simulation.

We consider distributions $F$ that are normal, Bernoulli, Poisson, and multinomial in our simulation study. For each case, we set the regression parameters of the
data generation model so that the associations between the variables are of similar magnitude to those observed in the TBI data. We aim to keep the strength of the associations nearly constant throughout the simulation study. Our primary goal is to evaluate the performance of our joint model for different distributions for subject-level variables and for different observation-level covariance structures. We generate data for each case as described below.

Normally Distributed Subject-Level Covariate

For a normally distributed subject-level covariate, we consider the following scenario. Let \( X_{i1} \sim N(0,1) \). We set the regression parameters as follows:

\[
\beta_1 = [10 -0.3 0.25 0.5]'
\]
\[
\beta_2 = [4 -0.15 0.3]'
\]
\[
\beta_3 = [2 -0.1]' .
\]

Thus, the correlation matrix for the generated data, \([Y_{i1} Y_{i2} Y_{i3} X_{i1}]' \), under the exchangeable covariance setting (Setting 5) is

\[
\begin{bmatrix}
1 & 0.73 & 0.73 & 0.37 & 0.28 & 0.28 & 0.49 & 0.35 & 0.35 & -0.32 \\
0.73 & 1 & 0.73 & 0.28 & 0.37 & 0.28 & 0.35 & 0.49 & 0.35 & -0.32 \\
0.73 & 0.73 & 1 & 0.28 & 0.28 & 0.37 & 0.35 & 0.35 & 0.49 & -0.32 \\
0.37 & 0.28 & 0.28 & 1 & 0.71 & 0.71 & 0.30 & 0.21 & 0.21 & -0.17 \\
0.28 & 0.37 & 0.28 & 0.71 & 1 & 0.71 & 0.21 & 0.30 & 0.21 & -0.17 \\
0.28 & 0.28 & 0.37 & 0.71 & 0.71 & 1 & 0.21 & 0.21 & 0.30 & -0.17 \\
0.49 & 0.35 & 0.35 & 0.30 & 0.21 & 0.21 & 1 & 0.70 & 0.70 & -0.10 \\
0.35 & 0.49 & 0.35 & 0.21 & 0.30 & 0.21 & 0.70 & 1 & 0.70 & -0.10 \\
0.35 & 0.35 & 0.49 & 0.21 & 0.21 & 0.30 & 0.70 & 0.70 & 1 & -0.10 \\
-0.32 & -0.32 & -0.32 & -0.17 & -0.17 & -0.17 & -0.10 & -0.10 & -0.10 & 1
\end{bmatrix} .
\]

Bernoulli Subject-Level Covariate

For a Bernoulli distributed subject-level covariate, we consider two different means that lead to a symmetric and a skewed distribution. It is important that we evaluate
our method under both scenarios because bias can sometimes result from imputing missing variables with distributions that are asymmetric (Yucel, 2008). First, we let $X_{i1} \overset{iid}{\sim} \text{Bern}(0.5)$, which is a symmetric distribution. We set the regression parameters as follows:

$$
\beta_1 = [10 \ -0.75 \ 0.5 \ 0.25]' \\
\beta_2 = [4 \ -0.5 \ 0.2]' \\
\beta_3 = [2 \ -0.3]' .
$$

Thus, the correlation matrix for the generated data, $[Y_{i1} \ Y_{i2} \ Y_{i3} \ X_{i1}]'$, under the exchangeable covariance setting (Setting 5) is

$$
\begin{bmatrix}
1 & 0.75 & 0.75 & 0.53 & 0.40 & 0.40 & 0.33 & 0.25 & 0.25 & -0.43 \\
0.75 & 1 & 0.75 & 0.40 & 0.53 & 0.40 & 0.25 & 0.33 & 0.25 & -0.43 \\
0.75 & 0.75 & 1 & 0.40 & 0.40 & 0.53 & 0.25 & 0.25 & 0.33 & -0.43 \\
0.53 & 0.40 & 0.40 & 1 & 0.72 & 0.53 & 0.23 & 0.17 & 0.17 & -0.26 \\
0.40 & 0.53 & 0.40 & 0.72 & 1 & 0.53 & 0.23 & 0.17 & 0.23 & -0.26 \\
0.40 & 0.40 & 0.53 & 0.72 & 0.72 & 1 & 0.17 & 0.17 & 0.23 & -0.26 \\
0.33 & 0.25 & 0.25 & 0.23 & 0.17 & 0.23 & 1 & 0.71 & 0.71 & -0.15 \\
0.25 & 0.33 & 0.25 & 0.17 & 0.17 & 0.71 & 1 & 0.71 & -0.15 \\
0.25 & 0.25 & 0.33 & 0.23 & 0.71 & 0.71 & 1 & -0.15 \\
-0.43 & -0.43 & -0.43 & -0.26 & -0.26 & -0.26 & -0.15 & -0.15 & -0.15 & 1 \\
\end{bmatrix}. \quad (4.12)
$$

Next, we let $X_{i1} \overset{iid}{\sim} \text{Bern}(0.1)$, which is a skewed distribution. We set the regression parameters as follows:

$$
\beta_1 = [10 \ -1.2 \ 0.5 \ 0.25]' \\
\beta_2 = [4 \ -0.75 \ 0.2]' \\
\beta_3 = [2 \ -0.5]' .
$$
Thus, the correlation matrix for the generated data, $[Y_{i,1} \ Y_{i,2} \ Y_{i,3} \ X_{i1}]'$, under the exchangeable covariance setting (Setting 5) is

$$\begin{bmatrix}
1 & 0.75 & 0.75 & 0.52 & 0.39 & 0.39 & 0.33 & 0.25 & 0.25 & -0.41 \\
0.75 & 1 & 0.75 & 0.39 & 0.52 & 0.39 & 0.25 & 0.33 & 0.25 & -0.41 \\
0.75 & 0.75 & 1 & 0.39 & 0.39 & 0.52 & 0.25 & 0.25 & 0.33 & -0.41 \\
0.52 & 0.39 & 0.39 & 1 & 0.72 & 0.72 & 0.22 & 0.17 & 0.17 & -0.24 \\
0.39 & 0.52 & 0.39 & 0.72 & 1 & 0.72 & 0.17 & 0.22 & 0.17 & -0.24 \\
0.39 & 0.39 & 0.52 & 0.72 & 0.72 & 1 & 0.17 & 0.17 & 0.22 & -0.24 \\
0.33 & 0.25 & 0.25 & 0.22 & 0.17 & 0.17 & 1 & 0.71 & 0.71 & -0.15 \\
0.25 & 0.33 & 0.25 & 0.17 & 0.22 & 0.17 & 0.71 & 1 & 0.71 & -0.15 \\
0.25 & 0.25 & 0.33 & 0.17 & 0.22 & 0.71 & 0.71 & 1 & -0.15 \\
-0.41 & -0.41 & -0.41 & -0.25 & -0.25 & -0.25 & -0.15 & -0.15 & 0 & -0.15 & 1
\end{bmatrix}$$ (4.13)

**Poisson Subject-Level Covariate**

We consider the following scenario for a Poisson subject-level variable. Let $X_{i1} \ iid \sim Pois(2.5)$, which is approximately the distribution of the number of children in the household in the TBI data. We set the regression parameters as follows:

$$\beta_1 = [10 \ -0.3 \ 0.25 \ 0.5]'$$

$$\beta_2 = [4 \ -0.15 \ 0.3]'$$

$$\beta_3 = [2 \ -0.1]'$$.

Thus, the correlation matrix for the generated data, $[Y_{i,1} \ Y_{i,2} \ Y_{i,3} \ X_{i1}]'$, under the exchangeable covariance setting (Setting 5) is

$$\begin{bmatrix}
1 & 0.77 & 0.77 & 0.42 & 0.33 & 0.33 & 0.50 & 0.37 & 0.37 & -0.47 \\
0.77 & 1 & 0.77 & 0.33 & 0.42 & 0.33 & 0.37 & 0.50 & 0.37 & -0.47 \\
0.77 & 0.77 & 1 & 0.33 & 0.33 & 0.42 & 0.37 & 0.37 & 0.50 & -0.47 \\
0.42 & 0.33 & 0.33 & 1 & 0.72 & 0.72 & 0.31 & 0.23 & 0.23 & -0.26 \\
0.33 & 0.42 & 0.33 & 0.72 & 1 & 0.72 & 0.23 & 0.31 & 0.23 & -0.26 \\
0.33 & 0.33 & 0.42 & 0.72 & 0.72 & 1 & 0.23 & 0.23 & 0.31 & -0.26 \\
0.50 & 0.37 & 0.37 & 0.31 & 0.23 & 0.23 & 1 & 0.71 & 0.71 & -0.16 \\
0.37 & 0.50 & 0.37 & 0.23 & 0.23 & 0.31 & 0.71 & 1 & 0.71 & -0.16 \\
0.37 & 0.37 & 0.50 & 0.23 & 0.23 & 0.31 & 0.71 & 0.71 & 1 & -0.16 \\
-0.47 & -0.47 & -0.47 & -0.26 & -0.26 & -0.26 & -0.16 & -0.16 & -0.16 & 1
\end{bmatrix}$$ (4.14)
Multinomial Subject-Level Covariate

We consider the following two scenarios for a multinomial subject-level variable. Let $X_{i1} \overset{iid}{\sim} \text{Mult}(0.23, 0.66, 0.11)$, which is approximately the distribution of maternal education in the TBI data. We let $X_{i1}$ take the values 0, 1, and 2. First, we assume $X_{i1}$ is ordinal and treat it as a continuous variable in the data generation and analysis of $Y$. We set the regression parameters as follows:

\[ \beta_1 = [10 -0.75 0.5 0.25]' \]
\[ \beta_2 = [4 -0.45 0.2]' \]
\[ \beta_3 = [2 -0.25]' \]

Thus, the correlation matrix for the generated data, $[Y_{i1}, Y_{i2}, Y_{i3}, X_{i1}]'$, under the exchangeable covariance setting (Setting 5) is

\[
\begin{bmatrix}
1 & 0.76 & 0.76 & 0.53 & 0.41 & 0.41 & 0.33 & 0.25 & 0.25 & -0.46 \\
0.76 & 1 & 0.76 & 0.41 & 0.53 & 0.41 & 0.25 & 0.33 & 0.25 & -0.46 \\
0.76 & 0.76 & 1 & 0.41 & 0.53 & 0.25 & 0.25 & 0.33 & -0.46 & -0.46 \\
0.53 & 0.41 & 0.41 & 1 & 0.72 & 0.72 & 0.23 & 0.71 & 0.17 & -0.27 \\
0.41 & 0.53 & 0.41 & 0.72 & 1 & 0.72 & 0.23 & 0.17 & 0.17 & -0.27 \\
0.41 & 0.41 & 0.53 & 0.72 & 0.72 & 1 & 0.17 & 0.17 & 0.23 & -0.26 \\
0.33 & 0.25 & 0.25 & 0.23 & 0.17 & 0.17 & 1 & 0.71 & 0.71 & -0.14 \\
0.25 & 0.33 & 0.25 & 0.17 & 0.23 & 0.71 & 1 & 0.71 & -0.14 & -0.14 \\
0.25 & 0.25 & 0.33 & 0.17 & 0.17 & 0.71 & 0.71 & 1 & -0.14 & -0.14 \\
-0.46 & -0.46 & -0.46 & -0.27 & -0.27 & -0.27 & -0.14 & -0.14 & -0.14 & 1
\end{bmatrix} \tag{4.15}
\]

In the second case, we treat $X_{i1}$ as a categorical variable in the data generation and analysis of $Y$. We let $X_{i1}$ have the same distribution as above in the ordinal case. We set the regression parameters as follows with $X_{i1} = 0$ as the reference group. We generate $Y$ according the following model:

\[
Y_{ij1} = \beta_{01} + \beta_{11}^{(1)} I(X_{i1} = 1) + \beta_{11}^{(2)} I(X_{i1} = 2) + \beta_{21} Y_{ij2} + \beta_{31} Y_{ij3} + e_{ij1} \tag{4.16}
\]
\[
Y_{ij2} = \beta_{02} + \beta_{12}^{(1)} I(X_{i1} = 1) + \beta_{12}^{(2)} I(X_{i1} = 2) + \beta_{22} Y_{ij3} + e_{ij2}
\]
\[ Y_{ij3} = \beta_{03} + \beta_{13}^{(1)} I(X_{i1} = 1) + \beta_{13}^{(2)} I(X_{i1} = 2) + e_{ij3} \] (4.17)

where

\[ \beta_1 = \begin{bmatrix} 10 & -0.4 & -1.3 & 0.5 & 0.25 \end{bmatrix} \]  
\[ \beta_2 = \begin{bmatrix} 4 & -0.2 & -1 & 0.2 \end{bmatrix} \]  
\[ \beta_3 = \begin{bmatrix} 2 & -0.1 & -0.5 \end{bmatrix} \]

and \( I(\cdot) \) is an indicator function. Since we treat \( X_{i1} \) as a categorical (i.e., nominal) variable, we omit presentation of the correlation matrix.

### 4.3.2 Analysis Methods

The analysis model is the generalized least squares regression model for \( Y_{ij1} \) that matches the data generation model (4.9):

\[ Y_{ij1} = \beta_{01} + \beta_{11} X_{i1} + \beta_{21} Y_{ij2} + \beta_{31} Y_{ij3} + e_{ij1}, \] (4.18)

where the covariance structure of the errors, \( \Sigma \), is correctly specified for each case, except for the Toeplitz case. The Toeplitz structure is not implemented in the nlme R package so instead we specify an unstructured correlation structure.

Each simulated data set is analyzed with five different procedures: an analysis based on the data prior to deletion of missing values (“before deletion”), a complete case analysis, multiple imputation using a multivariate normal distribution described in Section 2.3.3, multiple imputation using the sequential conditional method that is also described in Section 2.3.3, and multiple imputation using the joint latent variable model described in Section 4.2. We use the complete case and before deletion analyses.
as standards for comparison. For each imputation approach, we use 50 imputations, as suggested by White et al. (2011) to match the 50% missingness rate. The imputed data sets are combined using Rubin’s Rules with the small sample degrees of freedom approximation (Little and Rubin, 2002).

4.3.3 Implementation of Imputation Procedures

The details of the implementation of each imputation procedure for each distribution of $X_{i1}$ that we considered are as follows.

Normally Distributed Subject-Level Variable

We implement the multivariate normal imputation procedure using the norm R package (Schafer, 1997). We draw imputations for missing $X_{i1}$ from the following joint model:

$$\begin{bmatrix} Y_{i1} & Y_{i2} & Y_{i3} & X_{i1} \end{bmatrix}^{t \text{ ind}} \sim N(\mu, K).$$

(4.19)

We use the default prior distribution in the package (Schafer, 1997),

$$P(\mu, K) \propto |K|^{-\frac{n+1}{2}},$$

(4.20)

which is the improper limiting form of the normal-inverse Wishart distribution.

We implement the sequential conditional method as we did in Chapter 3 with the MICE R package (van Buuren and Groothuis-Oudshoorn, 2011) and the recommendation of Yucel (2008). That is, we impute from the following conditional distribution of $X_{i1}$:

$$X_{i1} = \alpha_0 + \alpha_1 \bar{Y}_{i1} + \alpha_2 \bar{Y}_{i2} + \alpha_3 \bar{Y}_{i3} + \epsilon_i$$

(4.21)

$$\epsilon_i \overset{\text{iid}}{\sim} N(0, \kappa^2).$$

(4.22)
We use the default prior distributions in the package,

\[ P(\kappa^2) \propto \frac{1}{\kappa^2}, \]  

and uniform distributions on the real line for the regression parameters (Schafer, 1997).

We implement our joint model described in Section 4.2 as follows. For the simulation, we fix the number of latent variables at \( q^{(o)} = 1 \) and \( q^{(s)} = 2 \) for all cases that we consider. We suggest a potential approach for choosing the number of latent variables in Section 4.5. Since each observation-level variable in our simulation study is normally distributed, we use the identity link function in (4.5). In addition, we must estimate variance parameters for each of the observation-level variables at each of the 3 time points, \( \sigma^{2(o)}_{jm} \); for these we use prior distributions \( \sigma^{(o)}_{jm} \sim U(0, \infty) \). That is,

\[ Y_{ijm}^{(o)} | \theta_{ijm}^{(o)}, \sigma_{jm}^{2(o)} \overset{\text{iid}}{\sim} N(\theta_{ijm}^{(o)}, \sigma_{jm}^{2(o)}), \ j = 1, 2, 3, \ m = 1, 2, 3 \]  

\[ \theta_{ijm}^{(o)} = \mu_{jm}^{(o)} + \xi_{ij}^{(o)} + \xi_{i}^{(s)} \gamma_{jm}^{(o)} + \nu_{im}^{(o)}. \]  

Likewise, since \( X_{i1} \) is normally distributed, we use the identity link function in (4.6) and must estimate the variance, \( \sigma^{2(s)}_{1} \), which has prior distribution \( \sigma^{(s)}_{1} \sim U(0, \infty) \). That is,

\[ X_{i1} | \theta_{i1}^{(s)}, \sigma_{1}^{2(s)} \overset{\text{iid}}{\sim} N(\theta_{i1}^{(s)}, \sigma_{1}^{2(s)}) \]  

\[ \theta_{i1}^{(s)} = \mu_{1}^{(s)} + \xi_{i}^{(s)} \gamma_{1}^{(s)}. \]  

The other prior distributions we use are as follows:

\[ \mu_{jm}^{(o)} \overset{\text{iid}}{\sim} N(0, 100) \]
\[ \mu_1^{(s)} \sim N(0, 100) \]
\[ \lambda_{jm}^{(o)} \text{iid} \sim N(0, 100) \]
\[ \gamma_m^{(o)} \text{iid} \sim N(0, 100I_2) \]
\[ \gamma_1^{(s)} \sim N(0, 100I_2) \]
\[ \phi_{jk}^{(o)} \text{iid} \sim N(0, 20), k < j \]
\[ \omega_m \text{iid} \sim U(0, \infty) \]
\[ \psi_{j1}^{(o)} \text{iid} \sim U(0, \infty) \]
\[ \phi_{rk}^{(s)} \text{iid} \sim N(0, 20), k < r \]
\[ \psi_r^{(s)} \text{iid} \sim U(0, \infty). \]

For each independent simulated data set, we use 20,000 iterations of a Gibbs sampling algorithm to compute the posterior distribution. Using the last 5000 iterations, we generate 50 imputed data sets by keeping every 100th draw for missing values of \( X_{i1} \).

**Bernoulli Subject-Level Variable**

For a binary subject-level variable, we implement the multivariate normal approach where we (incorrectly) assume that \( X_{i1} \) is normally distributed. The implementation is a two-stage procedure. The first stage is implemented the same as described for the normal case above where values for missing \( X_{i1} \) are drawn from their predictive distributions based on a joint multivariate normal distribution. In the second stage, the draws are rounded to 0 or 1 so the imputed values are binary. We use the simple rounding approach, as that is what is commonly implemented in software (Bernaards et al., 2007). Simple rounding rounds values less than 0.5 to 0 and values greater than 0.5 to 1.
We implement the sequential conditional method with the MICE R package (van Buuren and Groothuis-Oudshoorn, 2011) and the recommendation of Yucel (2008). Since $X_{i1}$ is a binary variable, we now use a logistic regression model for imputation. That is, we impute from the following conditional distribution of $X_{i1}$:

$$X_{i1}|p_i \overset{ind}{\sim} \text{Bern}(p_i)$$

(4.27)

$$\text{logit}(p_i) = \alpha_0 + \alpha_1 Y_{i1} + \alpha_2 Y_{i2} + \alpha_3 Y_{i3}.$$  

(4.28)

We use uniform distributions on the real line for the prior distributions of the regression parameters, as is the default in the package. The package also uses the large sample approximation to the posterior distribution to draw parameter values.

The implementation of our joint model is similar to the normal case, described above. The observation-level sub-model and the models for the latent variables on both levels are identical to those described for the normal case. Since we now have a binary subject-level variable, we adjust the subject-level sub-model accordingly. We now use the logistic link in (4.6). That is,

$$X_{i1}|\theta_{i1}^{(s)} \overset{ind}{\sim} \text{Bern}(\theta_{i1}^{(s)})$$

(4.29)

$$\text{logit}(\theta_{i1}^{(s)}) = \mu_1^{(s)} + \xi_i^{(s)} \gamma_1^{(s)}.$$  

(4.30)

We change the prior distribution for $\gamma_1^{(s)}$ to $N(0,20I_2)$ to reflect the smaller scale of Bernoulli random variables relative to variables with a larger support. Likewise, the prior distribution for $\mu_1^{(s)}$ is set to $N(0,20)$. Computationally, our Gibbs sampling algorithm now includes a Metropolis-Hastings step. We use the adaptive Metropolis-Hastings algorithm described in Browne and Draper (2006). For each independent simulated data set, we use 20,000 iterations of a Gibbs sampling algorithm to compute
the posterior distribution. Using the last 5000 iterations, we generate 50 imputed data sets by keeping every 100th draw for missing values of $X_{i1}$.

**Poisson Subject-Level Variable**

For a Poisson subject-level variable, we implement the multivariate normal approach using a square root transformation of $X_{i1}$, which is commonly used to stabilize the variance and improve the symmetry of count data. Thus, after transformation, we impute missing values of $\sqrt{X_{i1}}$ from a multivariate normal distribution. After imputation, we back transform by squaring the imputed values and rounding to the nearest integer.

We implement the sequential conditional method with the MICE R package (Kleinke and Reinecke, 2013; van Buuren and Groothuis-Oudshoorn, 2011) and the recommendation of Yucel (2008). For a Poisson variable, a Poisson regression model is fit using the log link function. Imputations are drawn from the following Poisson distribution:

$$X_{i1}|\lambda_i \sim \text{Pois}(\lambda_i)$$

(4.31)

$$\log(\lambda_i) = \alpha_0 + \alpha_1 Y_{i1} + \alpha_2 Y_{i2} + \alpha_3 Y_{i3}. \quad (4.32)$$

We use uniform distributions on the real line for the prior distributions of the regression parameters, as is the default in the package. The package also uses the large sample approximation to the posterior distribution to draw parameter values.

The implementation of our joint model is more complex for the case of a Poisson subject-level variable. To be compatible with a Poisson marginal distribution, the conditional distribution that we impute from is likely to be underdispersed. We
show this by contradiction. Let $E[X_{i1}] \geq Var(X_{i1})$, as in a Poisson or underdispersed Poisson distribution. Now suppose we condition on additional information $Y_i$ that is correlated with $X_{i1}$ such that $E[X_{i1}|Y_i]$ is not constant for all $i$. That is, $Var(E[X_{i1}|Y_i]) > 0$. Assume the conditional distribution of $X_{i1}|Y_i$ is not underdispersed so that $E[X_{i1}|Y_i] \leq Var(X_{i1}|Y_i)$ for all $Y_i$. The use of iterated expectations shows one relationship between the unconditional and conditional distributions:

$$Var(X_{i1}) = E[Var(X_{i1}|Y_i)] + Var(E[X_{i1}|Y_i]). \quad (4.33)$$

By assumed lack of underdispersion,

$$Var(X_{i1}) \geq E[E[X_{i1}|Y_i]] + Var(E[X_{i1}|Y_i]), \quad (4.34)$$

and thus since $E[E[X_{i1}|Y_i]] = E[X_{i1}]$ and $Var(E[X_{i1}|Y_i]) > 0$,

$$Var(X_{i1}) > E[X_{i1}] \quad (4.35)$$

which contradicts that $E[X_{i1}] \geq Var(X_{i1})$. Therefore, $Var(X_{i1}|Y_i) < E[X_{i1}|Y_i]$.

Thus the conditional distribution must be underdispersed when the marginal distribution that generates the data is assumed to be conditionally Poisson.

Thus, instead of using a standard Poisson generalized linear model, we use a generalized linear model based on the less familiar Conway-Maxwell-Poisson (COM-Poisson) distribution (Conway and Maxwell, 1962). A detailed description of the COM-Poisson distribution can be found in Shmueli et al. (2005) and Sellers et al. (2012). Briefly, the COM-Poisson distribution is a Poisson distribution that is generalized to accommodate overdispersion, underdispersion, and the standard Poisson distribution. Thus, a COM-Poisson model provides the flexibility of a quasi-Poisson model but with a true likelihood.
Before we define the subject-level sub-model, we describe the COM-Poisson distribution. The COM-Poisson distribution has probability mass function

\[ X_{i1} \sim CMP(\theta, \sigma) \quad (4.36) \]

\[ P(X_{i1} = x_{i1}) = \frac{\theta^{x_{i1}}}{(x_{i1})!^\sigma} \frac{1}{Z(\theta, \sigma)}, \quad x_{i1} = 0, 1, 2, \ldots \quad (4.37) \]

with intensity parameter \( \theta > 0 \), dispersion parameter \( \sigma \geq 0 \), and normalizing constant

\[ Z(\theta, \sigma) = \sum_{k=0}^{\infty} \frac{\theta^k}{(k!)^\sigma}. \quad (4.38) \]

Note that when \( \sigma = 1 \), the COM-Poisson distribution is simply the Poisson distribution. For overdispersed data \( \sigma < 1 \) and for underdispersed data \( \sigma > 1 \).

The COM-Poisson distribution is computationally difficult to implement because the normalizing constant is an infinite summation that is a function of the parameters (Wu et al., 2013). We use the following asymptotic approximation (Minka et al., 2003) to make computations feasible:

\[ Z(\theta, \sigma) = e^{\frac{\sigma \theta^{\frac{1}{\sigma}}}{\theta^{\frac{1}{\sigma} - \frac{1}{2}}} - \frac{\sigma}{2} \theta^{\frac{1}{\sigma}} \sqrt{\frac{\pi}{2 \sigma}}} \quad (4.39) \]

Using this approximation, we can implement feasible Markov chain Monte Carlo algorithms.

Conveniently, the COM-Poisson fits into the generalized linear models framework, and we use the general concepts from Wu et al. (2013) as a guide for specifying our model. As with a Poisson regression model, we use a log link function to model the intensity parameter. That is,

\[ X_{i1}|\theta^{(s)}_{i1}, \sigma^{(s)}_{i1} \sim CMP(\theta^{(s)}_{i1}, \sigma^{(s)}_{i1}) \quad (4.40) \]

\[ log(\theta^{(s)}_{i1}) = \mu^{(s)}_{i1} + \xi^{(s)}_{i1} \gamma^{(s)}_{i1}. \quad (4.41) \]
Prior distributions remain as defined in Section 4.2.6, except for the dispersion parameter. Hyperparameters are specified as in the normal case. Instead of a uniform prior on the positive real line for $\sigma_1^{(s)}$, we follow Wu et al. (2013) and assign the log of $\sigma_1^{(s)}$ a normal prior distribution where the variance is selected by using prior beliefs about the range of likely values for $\sigma_1^{(s)}$ divided by 6 as an estimate of standard deviation. That is,

$$\log(\sigma_1^{(s)}) \sim N(0, \kappa^2) \quad (4.42)$$

$$\kappa = \frac{\log(1.5) - \log(0.67)}{6}. \quad (4.43)$$

Our choice reflects a belief that the dispersion parameter falls between 0.67 and 1.5.

We discuss the need for more guidance for the choice of $\sigma_1^{(s)}$ in Section 4.6. The observation-level sub-model and the models for the latent variables on both levels are identical to those described for the normal case. As before, we use the adaptive Metropolis-Hastings algorithm of Browne and Draper (2006). For each independent simulated data set, we use 20,000 iterations of a Gibbs sampling algorithm to compute the posterior distribution. Using the last 5000 iterations, we generate 50 imputed data sets by keeping every 100\textsuperscript{th} draw for missing values of $X_{i1}$.

To illustrate the shortcomings of using a usual Poisson model, we also implement our joint model with a Poisson distribution. We use a Poisson regression model with the log link function. That is,

$$X_{i1} | \theta_{i1}^{(s)} \overset{ind}{\sim} Pois(\theta_{i1}^{(s)}) \quad (4.44)$$

$$\log(\theta_{i1}^{(s)}) = \mu_1^{(s)} + \xi_i^{(s)} \gamma_1^{(s)}. \quad (4.45)$$
We use the prior distributions as defined in normal case. As before, we use the adaptive Metropolis-Hastings algorithm of Browne and Draper (2006). For each independent simulated data set, we use 20,000 iterations of a Gibbs sampling algorithm to compute the posterior distribution. Using the last 5000 iterations, we generate 50 imputed data sets by keeping every 100th draw for missing values of $X_{i1}$.

**Multinomial Subject-Level Variable**

For a multinomial subject-level variable, we assume $X_{i1}$ is ordinal and implement the multivariate normal with rounding, as we did for the binary case. We again (incorrectly) assume that $X_{i1}$ is normally distributed and impute from a multivariate normal distribution. We then use simple rounding to round the imputed value to the nearest category, which in this case is 0, 1, or 2.

The sequential conditional method is implemented with the MICE R package (van Buuren and Groothuis-Oudshoorn, 2011) and follows the recommendation of Yucel (2008). We use a multinomial logistic regression model and impute from the following conditional distribution of $X_{i1}$:

$$X_{i1}|p_{i,0}, p_{i,1}, p_{i,2} \overset{ind}{\sim} Mult(p_{i,0}, p_{i,1}, p_{i,2})$$

$$\log \left( \frac{p_{i,k}}{p_{i,0}} \right) = \alpha_{0,k} + \alpha_{1,k}Y_{i,1} + \alpha_{2,k}Y_{i,2} + \alpha_{3,k}Y_{i,3}, \ k = 1, 2.$$  

(4.46)  

(4.47)

For each subject with a missing value, each $p_{i,k}$ is estimated and a value of $X_{i1}$ is drawn from the resulting multinomial distribution.

In addition, we fit a more restrictive version of the sequential conditional approach for ordinal data that imposes the proportional odds assumption. Thus, we fit the proportional odds model and impute from the following conditional distribution of $X_{i1}$:

$$X_{i1}|p_{i,0}, p_{i,1}, p_{i,2} \overset{ind}{\sim} Mult(p_{i,0}, p_{i,1}, p_{i,2})$$

$$\log \left( \frac{p_{i,k}}{p_{i,0}} \right) = \alpha_{0,k} + \alpha_{1,k}Y_{i,1} + \alpha_{2,k}Y_{i,2} + \alpha_{3,k}Y_{i,3}, \ k = 1, 2.$$  

For each subject with a missing value, each $p_{i,k}$ is estimated and a value of $X_{i1}$ is drawn from the resulting multinomial distribution.
$X_{i1}$:

$$X_{i1}|p_{i,0}, p_{i,1}, p_{i,2} \overset{ind}{\sim} \text{Mult}(p_{i,0}, p_{i,1}, p_{i,2})$$  \hfill (4.48)

$$\logit(P(X_{i1} \leq k)) = \alpha_{0,k} + \alpha_1 Y_{i1} + \alpha_2 Y_{i2} + \alpha_3 Y_{i3}, \quad k = 0, 1.$$  \hfill (4.49)

For each subject with a missing value, each $p_{i,k}$ is estimated and a value of $X_{i1}$ is drawn from the resulting multinomial distribution.

Our joint model is implemented similarly to the binary case, described above. Since we are now modeling a multinomial subject-level variable, we use multinomial logistic regression for the subject-level model. We use the more general multinomial logistic regression rather than imposing the proportional odds assumption. That is,

$$X_{i1}|\theta^{(s)}_{i1,0}, \theta^{(s)}_{i1,1}, \theta^{(s)}_{i1,2} \overset{ind}{\sim} \text{Mult}(\theta^{(s)}_{i1,0}, \theta^{(s)}_{i1,1}, \theta^{(s)}_{i1,2})$$  \hfill (4.50)

$$\log\left(\frac{\theta^{(s)}_{i1,k}}{\theta^{(s)}_{i1,0}}\right) = \mu^{(s)}_{1,k} + \xi^{(s)}_{1,k} \gamma^{(s)}_{1,k}, \quad k = 1, 2.$$  \hfill (4.51)

The prior distribution for $\gamma^{(s)}_{1,k}$ is $N(0, 20I_2)$ for $k = 1, 2$. The prior distribution for $\mu^{(s)}_{1,k}$ is $N(0, 20)$, also for $k = 1, 2$. As before, we use the adaptive Metropolis-Hastings algorithm of Browne and Draper (2006). For each independent simulated data set, we use 20,000 iterations of a Gibbs sampling algorithm to compute the posterior distribution. Using the last 5000 iterations, we generate 50 imputed data sets by keeping every 100th draw for missing values of $X_{i1}$.

We consider two cases of multinomial data: one where $X_{i1}$ is treated continuously in the generation of $Y_i$ and one where $X_{i1}$ is treated categorically in the generation of $Y_i$. For the situation when it is treated continuously, we fit each of the above approaches. When it is treated categorically, we omit the sequential conditional approach using the proportional odds model.
4.3.4 Performance Metrics

We use the same performance metrics as in Chapter 3, which are described in Section 3.4.3. We also present relative bias to facilitate comparison of performance across different parameter settings and distributions of subject-level.

4.3.5 Results

The results of our simulation study are presented by the type of subject-level variable being considered.

Normally Distributed Subject-Level Variable

Table 4.2 contains the performance metrics for estimation of the regression coefficients in (4.18) with a partially observed normally distributed subject-level covariate. We see that for all settings each imputation procedure has a smaller model standard error than the complete case analysis and thus gains efficiency. As expected, since it is a congenial model for all of the settings, the multivariate normal approach produces nearly unbiased estimates for each coefficient and confidence intervals with approximately nominal coverage.

Comparing the performance of our joint approach to MICE, we see very similar results for Settings 3-5. As shown in Chapter 3, MICE is congenial to the analysis model for Setting 5, and it is not surprising the joint approach has larger relative bias than MICE. Although, the joint method maintains appropriate confidence interval coverage. For the other settings, the estimate of $\beta_1$ from the joint method is slightly attenuated in each case with relative bias ranging from -0.104 to -0.048, but the relative bias is the same as MICE in Setting 4 and less than that for MICE in Setting 3 and 6. In fact, the relative bias for MICE in Setting 6 (nonconstant variance) is over...
Table 4.2: Simulation results comparing performance for estimating $\beta$ with normally distributed systematically missing subject-level covariate data. True values of the parameters are: $\beta_1 = -0.3$, $\beta_2 = 0.25$, and $\beta_3 = 0.5$. Results are averaged over 1000 independent replications. Correlation and variance parameter values for each setting are shown in Table 4.1.
twice that of the joint approach. The joint method provides appropriate confidence interval coverage for Settings 5 and 6. In Settings 3 and 4, we see slight undercoverage with rates of 0.935 and 0.902, but they are comparable to the coverage rates of MICE. For all settings, both the joint and MICE approaches give nearly unbiased estimates of \( \beta_2 \) and \( \beta_3 \).

**Binary Subject-Level Variable**

Table 4.3 contains the performance metrics for estimation of the regression coefficients in (4.18) with a partially observed Bernoulli distributed \((p = 0.5)\) subject-level covariate. Again, in all settings all of the imputation methods are more efficient than the complete case analysis.

In Settings 3, 4, and 6, the estimates of \( \beta_1 \) from all of the imputation approaches are attenuated with relative bias ranging from -0.117 to -0.015. We see the least bias in the estimates from the joint approach with the largest relative bias of only -0.037. In contrast, the smallest relative bias from the multivariate normal or MICE methods was -0.066. Thus, in terms of bias the joint method was superior in the settings that were not exchangeable with constant variance. For these three settings, the coverage rates of the joint approach were around 0.93 and better than the other methods except for MICE in Setting 3 which is essentially equivalent. For \( \beta_2 \) and \( \beta_3 \), the joint approach also is the least biased of the three approaches.

The multivariate normal approach with simple rounding shows the most attenuation, which is unsurprising given the high rate of missingness (Yucel et al., 2011). The estimates of \( \beta_2 \) and \( \beta_3 \) from the multivariate normal and MICE approaches are also slightly amplified.
Table 4.3: Simulation results comparing performance for estimating $\beta$ with binary systematically missing subject-level covariate data where $p = 0.5$. True values of the parameters are: $\beta_1 = -0.75$, $\beta_2 = 0.5$, and $\beta_3 = 0.25$. Results are averaged over 1000 independent replications. Correlation and variance parameter values for each setting are shown in Table 4.1.
In Setting 5 (exchangeable), the joint and MICE approaches perform similarly. We see nearly unbiased estimation of each of the parameters and confidence interval coverage that approaches the nominal value. The multivariate normal approach still suffers from attenuation of the estimate of $\beta_1$ and slight amplification of the estimates of $\beta_2$ and $\beta_3$, as in the other settings.

Table 4.4 contains the performance metrics for estimation of the regression coefficients in (4.18) with a partially observed Bernoulli distributed ($p = 0.1$) subject-level covariate. We see a very similar pattern of results in Table 4.4 as we saw in Table 4.3. We see a gain in efficiency for each of the imputation procedures compared to the complete case analysis.

In this case, the joint method produces the least biased results with relative bias ranging from -0.032 to -0.057. The next smallest relative bias is for MICE in Setting 5 at -0.058, and after that the next smallest relative bias is -0.116. For Settings 3, 4, and 6, the joint approach has relative bias that is less than half that of the other approaches. Even in the exchangeable case, the relative bias from MICE is still almost twice that of the joint method. Coverage of the joint method is at the nominal value for each parameter and estimation of $\beta_2$ and $\beta_3$ are nearly unbiased as well.

We see a nearly doubling in the relative bias of the multivariate normal approach from the previous scenario. This is not unexpected because for the previous case, with $p = 0.5$, the distribution of $X_{i1}$ is symmetric, but in this case, $p = 0.1$, we no longer have symmetry. It is known that the multivariate normal approach can lead to bias as distributions depart from symmetry (Yucel et al., 2011). MICE also performs worse compared to the previous setting with larger values of relative bias and lower coverage rates. The best setting for MICE is Setting 5 where it has appropriate
Table 4.4: Simulation results comparing performance for estimating $\beta$ with binary systematically missing subject-level covariate data where $p = 0.1$. True values of the parameters are: $\beta_1 = -1.2$, $\beta_2 = 0.5$, and $\beta_3 = 0.25$. Results are averaged over 1000 independent replications. Correlation and variance parameter values for each setting are shown in Table 4.1.
coverage and the smallest relative bias of the four settings but is still outperformed by the joint method.

In summary, the joint approach performed the best across the two Bernoulli cases that we considered. It consistently had the smallest relative bias and coverage rates. The joint approach also did not suffer the decline in performance in the second case where the distribution of $X_{i1}$ was more skewed. Thus, the joint approach exhibited good performance across the range of potential Bernoulli distributions.

**Poisson Subject-Level Variable**

Table 4.5 contains the performance metrics for estimation of the regression coefficients in (4.18) with a partially observed Poisson distributed subject-level covariate. We consider two implementations of the joint approach for this case: one that imputes from the Poisson distribution and one that imputes from the COM-Poisson (CMP) distribution. We see for each setting that the relative bias of the joint method using the CMP is roughly half that of the joint method using the Poisson. In addition, the coverage of the joint method with the Poisson is very low with coverage rates from 0.171 to 0.801. This is not surprising since we established that we should impute from an underdispersed distribution.

In comparison to MICE and the multivariate normal methods, the joint (CMP) method also fares well. The MICE approach imputes from a Poisson distribution and so performs similarly to the joint Poisson approach. The multivariate normal approach is more competitive for Settings 3 and 4, although neither the joint (CMP) nor multivariate normal approaches perform particularly well. In Setting 3, the relative bias from the joint approach is -0.111 compared to -0.153 from the multivariate normal, but both methods suffer from undercoverage with coverage of 0.872 for the
<table>
<thead>
<tr>
<th>Setting 3 - Toeplitz</th>
<th>N</th>
<th>Mean</th>
<th>Rel. Bias</th>
<th>Model SE</th>
<th>Empirical SE</th>
<th>RMSE</th>
<th>CI Width</th>
<th>CI Coverage</th>
<th>( \hat{\beta}_1 )</th>
<th>( \hat{\beta}_2 )</th>
<th>( \hat{\beta}_3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before Deletion</td>
<td>400</td>
<td>-0.300</td>
<td>0.001</td>
<td>0.027</td>
<td>0.028</td>
<td>0.027</td>
<td>0.106</td>
<td>0.940</td>
<td>0.248</td>
<td>0.029</td>
<td>0.957</td>
</tr>
<tr>
<td>Complete Case</td>
<td>200</td>
<td>-0.301</td>
<td>0.003</td>
<td>0.038</td>
<td>0.039</td>
<td>0.030</td>
<td>0.150</td>
<td>0.954</td>
<td>0.247</td>
<td>0.041</td>
<td>0.949</td>
</tr>
<tr>
<td>Normal(Rounding)</td>
<td>400</td>
<td>-0.254</td>
<td>-0.153</td>
<td>0.034</td>
<td>0.031</td>
<td>0.055</td>
<td>0.136</td>
<td>0.761</td>
<td>0.254</td>
<td>0.030</td>
<td>0.952</td>
</tr>
<tr>
<td>MICE</td>
<td>400</td>
<td>-0.241</td>
<td>-0.198</td>
<td>0.034</td>
<td>0.025</td>
<td>0.065</td>
<td>0.133</td>
<td>0.613</td>
<td>0.260</td>
<td>0.030</td>
<td>0.944</td>
</tr>
<tr>
<td>Joint(Poisson)</td>
<td>400</td>
<td>-0.233</td>
<td>-0.224</td>
<td>0.034</td>
<td>0.026</td>
<td>0.072</td>
<td>0.133</td>
<td>0.477</td>
<td>0.265</td>
<td>0.030</td>
<td>0.929</td>
</tr>
<tr>
<td>Joint(CMP)</td>
<td>400</td>
<td>-0.267</td>
<td>-0.111</td>
<td>0.034</td>
<td>0.031</td>
<td>0.046</td>
<td>0.136</td>
<td>0.872</td>
<td>0.259</td>
<td>0.030</td>
<td>0.949</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Setting 4 - Unstructured</th>
<th>N</th>
<th>Mean</th>
<th>Rel. Bias</th>
<th>Model SE</th>
<th>Empirical SE</th>
<th>RMSE</th>
<th>CI Width</th>
<th>CI Coverage</th>
<th>( \hat{\beta}_1 )</th>
<th>( \hat{\beta}_2 )</th>
<th>( \hat{\beta}_3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before Deletion</td>
<td>400</td>
<td>-0.300</td>
<td>-0.001</td>
<td>0.024</td>
<td>0.024</td>
<td>0.024</td>
<td>0.094</td>
<td>0.947</td>
<td>0.250</td>
<td>0.029</td>
<td>0.936</td>
</tr>
<tr>
<td>Complete Case</td>
<td>200</td>
<td>-0.299</td>
<td>-0.002</td>
<td>0.034</td>
<td>0.035</td>
<td>0.035</td>
<td>0.133</td>
<td>0.938</td>
<td>0.251</td>
<td>0.041</td>
<td>0.960</td>
</tr>
<tr>
<td>Normal(Rounding)</td>
<td>400</td>
<td>-0.254</td>
<td>-0.154</td>
<td>0.031</td>
<td>0.028</td>
<td>0.054</td>
<td>0.121</td>
<td>0.697</td>
<td>0.258</td>
<td>0.030</td>
<td>0.938</td>
</tr>
<tr>
<td>MICE</td>
<td>400</td>
<td>-0.232</td>
<td>-0.228</td>
<td>0.030</td>
<td>0.022</td>
<td>0.072</td>
<td>0.118</td>
<td>0.349</td>
<td>0.267</td>
<td>0.030</td>
<td>0.896</td>
</tr>
<tr>
<td>Joint(Poisson)</td>
<td>400</td>
<td>-0.219</td>
<td>-0.270</td>
<td>0.031</td>
<td>0.023</td>
<td>0.084</td>
<td>0.121</td>
<td>0.171</td>
<td>0.275</td>
<td>0.030</td>
<td>0.867</td>
</tr>
<tr>
<td>Joint(CMP)</td>
<td>400</td>
<td>-0.253</td>
<td>-0.157</td>
<td>0.031</td>
<td>0.028</td>
<td>0.055</td>
<td>0.124</td>
<td>0.723</td>
<td>0.267</td>
<td>0.030</td>
<td>0.904</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Setting 5 - Exchangeable</th>
<th>N</th>
<th>Mean</th>
<th>Rel. Bias</th>
<th>Model SE</th>
<th>Empirical SE</th>
<th>RMSE</th>
<th>CI Width</th>
<th>CI Coverage</th>
<th>( \hat{\beta}_1 )</th>
<th>( \hat{\beta}_2 )</th>
<th>( \hat{\beta}_3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before Deletion</td>
<td>400</td>
<td>-0.300</td>
<td>-0.001</td>
<td>0.029</td>
<td>0.030</td>
<td>0.030</td>
<td>0.113</td>
<td>0.944</td>
<td>0.252</td>
<td>0.029</td>
<td>0.950</td>
</tr>
<tr>
<td>Complete Case</td>
<td>200</td>
<td>-0.299</td>
<td>-0.004</td>
<td>0.041</td>
<td>0.042</td>
<td>0.042</td>
<td>0.161</td>
<td>0.943</td>
<td>0.251</td>
<td>0.041</td>
<td>0.944</td>
</tr>
<tr>
<td>Normal(Rounding)</td>
<td>400</td>
<td>-0.252</td>
<td>-0.161</td>
<td>0.037</td>
<td>0.034</td>
<td>0.059</td>
<td>0.145</td>
<td>0.767</td>
<td>0.257</td>
<td>0.030</td>
<td>0.952</td>
</tr>
<tr>
<td>MICE</td>
<td>400</td>
<td>-0.254</td>
<td>-0.153</td>
<td>0.035</td>
<td>0.029</td>
<td>0.054</td>
<td>0.140</td>
<td>0.810</td>
<td>0.258</td>
<td>0.030</td>
<td>0.933</td>
</tr>
<tr>
<td>Joint(Poisson)</td>
<td>400</td>
<td>-0.244</td>
<td>-0.185</td>
<td>0.036</td>
<td>0.030</td>
<td>0.063</td>
<td>0.141</td>
<td>0.716</td>
<td>0.265</td>
<td>0.030</td>
<td>0.924</td>
</tr>
<tr>
<td>Joint(CMP)</td>
<td>400</td>
<td>-0.279</td>
<td>-0.060</td>
<td>0.037</td>
<td>0.036</td>
<td>0.041</td>
<td>0.145</td>
<td>0.919</td>
<td>0.259</td>
<td>0.030</td>
<td>0.944</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Setting 6 - Nonconstant Exchangeable</th>
<th>N</th>
<th>Mean</th>
<th>Rel. Bias</th>
<th>Model SE</th>
<th>Empirical SE</th>
<th>RMSE</th>
<th>CI Width</th>
<th>CI Coverage</th>
<th>( \hat{\beta}_1 )</th>
<th>( \hat{\beta}_2 )</th>
<th>( \hat{\beta}_3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before Deletion</td>
<td>400</td>
<td>-0.302</td>
<td>0.006</td>
<td>0.032</td>
<td>0.031</td>
<td>0.031</td>
<td>0.124</td>
<td>0.960</td>
<td>0.251</td>
<td>0.029</td>
<td>0.948</td>
</tr>
<tr>
<td>Complete Case</td>
<td>200</td>
<td>-0.301</td>
<td>0.005</td>
<td>0.045</td>
<td>0.044</td>
<td>0.044</td>
<td>0.176</td>
<td>0.948</td>
<td>0.251</td>
<td>0.041</td>
<td>0.946</td>
</tr>
<tr>
<td>Normal(Rounding)</td>
<td>400</td>
<td>-0.254</td>
<td>-0.152</td>
<td>0.040</td>
<td>0.036</td>
<td>0.058</td>
<td>0.160</td>
<td>0.828</td>
<td>0.256</td>
<td>0.030</td>
<td>0.946</td>
</tr>
<tr>
<td>MICE</td>
<td>400</td>
<td>-0.241</td>
<td>-0.195</td>
<td>0.040</td>
<td>0.032</td>
<td>0.067</td>
<td>0.199</td>
<td>0.765</td>
<td>0.261</td>
<td>0.030</td>
<td>0.937</td>
</tr>
<tr>
<td>Joint(Poisson)</td>
<td>400</td>
<td>-0.247</td>
<td>-0.177</td>
<td>0.040</td>
<td>0.032</td>
<td>0.062</td>
<td>0.156</td>
<td>0.801</td>
<td>0.262</td>
<td>0.030</td>
<td>0.941</td>
</tr>
<tr>
<td>Joint(CMP)</td>
<td>400</td>
<td>-0.280</td>
<td>-0.068</td>
<td>0.041</td>
<td>0.038</td>
<td>0.043</td>
<td>0.161</td>
<td>0.942</td>
<td>0.257</td>
<td>0.030</td>
<td>0.948</td>
</tr>
</tbody>
</table>

Table 4.5: Simulation results comparing performance for estimating \( \beta \) with Poisson systematically missing subject-level covariate data. True values of the parameters are: \( \beta_1 = -0.3 \), \( \beta_2 = 0.25 \), and \( \beta_3 = 0.5 \). Results are averaged over 1000 independent replications. Correlation and variance parameter values for each setting are shown in Table 4.1.
joint (CMP) and 0.761 for the multivariate normal. The performance for both methods in worse for Setting 4. The joint (CMP) approach and the multivariate normal approach have nearly identical relative bias at -0.157 and -0.154. The coverage for the joint (CMP) is slightly better than the multivariate normal at 0.723 compared to 0.697, but the both coverages are not close to the nominal value.

In Settings 5 and 6, the estimates from the joint (CMP) approach are less biased than those from the multivariate normal approach. The relative bias of the joint approach for Settings 5 and 7 are -0.069 and -0.068 compared to -0.161 and -0.152 from the multivariate normal approach. We also see coverage rates that are much closer to the nominal value for the joint (CMP) method. Coverage is particularly good for Setting 6 (nonconstant variance) at 0.942. The coverage for Setting 5 is 0.919 which is the highest coverage rate for any of the imputation methods for the exchangeable setting. Setting 5 is also the only setting where MICE slightly outperforms the multivariate normal approach with smaller relative bias (-0.153 to -0.161) and higher coverage (0.810 to 0.767).

Thus, while the performance of the joint (CMP) approach is generally better than the competing approaches, the overall results are mixed. It does adequately for Setting 5 and 6 where there is an exchangeable correlation structure of the observation-level variables. However, the performance for Settings 3 and 4, still has room for improvement. However, it is clear that the COM-Poisson model is an improvement over the Poisson model in setting we considered. We address the need for improvement and further investigation of the joint approach with the COM-Poisson distribution in Chapter 5.


Multinomial Subject-Level Variable

Table 4.6 contains the performance metrics for estimation of the regression coefficients in (4.18) with a partially observed multinomial distributed subject-level covariate that is treated as a continuous variable in the analysis. Recall that for this scenario, we implemented MICE with an ordinal logistic regression model and a multinomial logistic regression model. We notice that the two implementations of MICE perform nearly identically.

The joint approach again outperforms the other imputation procedures for Settings 3, 4, and 6. Relative bias from the joint method ranges from -0.033 to -0.042 in these three settings compared to -0.072 to -0.113 for the other methods. Once again, we see relative bias that is roughly half that of the next smallest value. The joint approach is also the only method with coverage above 0.9 with no coverage below 0.926. In Setting 5 (exchangeable), both MICE approaches perform well. They have smaller bias than the joint method (-0.005 to -0.022), although they have lower coverage (0.928 to 0.945) due to underestimation of the standard error.

Tables 4.7 and 4.8 contain the performance metrics for estimation of the regression coefficients in (4.18) with a partially observed multinomial distributed subject-level covariate that is treated as a categorical variable in the analysis. The results for $\beta_{1}^{(1)}$ and $\beta_{1}^{(2)}$ are presented in Table 4.7, and the results for $\beta_{2}$ and $\beta_{3}$ are presented in Table 4.8.

First, we examine Table 4.7. We see that for both parameters the joint approach has smaller relative bias and closer to nominal confidence interval coverage than either the multivariate normal or MICE approaches with the exception the relative bias of $\beta_{1}^{(2)}$ of Setting 5. For $\beta_{1}^{(2)}$ of Setting 5, MICE has a smaller relative bias than the
Table 4.6: Simulation results comparing performance for estimating \( \beta \) with multinomial systematically missing subject-level covariate data. We assume that \( X_{i1} \) is ordinal and treat it as a continuous variable in the analysis model. True values of the parameters are: \( \beta_1 = -0.75 \), \( \beta_2 = 0.5 \), and \( \beta_3 = 0.25 \). Results are averaged over 1000 independent replications. Correlation and variance parameter values for each setting are shown in Table 4.1.
Table 4.7: Simulation results comparing performance for estimating $\beta$ with multinomial systematically missing subject-level covariate data. We assume that $X_{i1}$ is a categorical variable in the analysis model. True values of the parameters are: $\beta_{1}^{(1)} = -0.4$ and $\beta_{1}^{(2)} = -1.3$. Results are averaged over 1000 independent replications. Correlation and variance parameter values for each setting are shown in Table 4.1.
Table 4.8: Simulation results comparing performance for estimating $\beta$ with multinomial systematically missing subject-level covariate data. We assume that $X_{i1}$ is a categorical variable in the analysis model. True values of the parameters are: $\beta_2 = 0.5$, and $\beta_3 = 0.25$. Results are averaged over 1000 independent replications. Correlation and variance parameter values for each setting are shown in Table 4.1.
joint approach (-0.010 to -0.025), but the joint approach maintains a superior coverage rate (0.958 to 0.931) because the MICE approach tends to underestimate the standard errors which leads to confidence intervals that are too narrow. Now looking at Table 4.8, we see that each approach gives nearly unbiased estimates with approximately nominal confidence interval coverage. Thus, regardless of whether we treat $X_{i1}$ as a continuous or categorical variable, the joint method performs better than the competing methods in each setting we considered, except for the exchangeable setting where the performance of MICE is comparable.

**Summary**

Through this simulation, we have evaluated the performance of our joint model for imputing systematically missing subject-level variables with normal, Bernoulli, Poisson, and multinomial distributions. The joint method performed as well and often better than the competing approaches across all of the scenarios that we considered. The largest gains in performance compared to competing methods were in the Bernoulli and multinomial cases and the nonexchangeable correlation settings. The joint method also did well using the COM-Poisson model for both the constant and nonconstant exchangeable settings with the Poisson subject-level variable. However, the performance for the nonexchangeable correlation structures in the Poisson scenario needs improvement. With the exception of the nonexchangeable Poisson settings, the joint method generally led to smaller values of relative bias and closer to nominal coverage rates than either of the competing imputation approaches.
4.4 TBI Data Analysis

In this section, we demonstrate the use of the complete case, multivariate normal, MICE, and joint methods in a series of sample analyses of the TBI data described in Section 1.2. Our outcome of interest is the normed internalizing subscale of the Child Behavior Checklist (CBCL-I), which is measured at three time points in each study. We will conduct four sample analyses to illustrate imputation of systematically missing variables that are normally distributed, Bernoulli, Poisson, and multinomial. We will consider each of the three systematically missing subject-level variables that are shown in Table 1.3. In addition, we create a binary variable that is an indicator of income below $25,000, which we will refer to as low income. That is, an indicator of belonging to the first category of the income variable from the preschool study. For the purpose of this sample analysis, we will assume that income and low income are MAR, even though this may be a questionable assumption for the subjects in the preschool study for whom income is sporadically missing.

Each analysis will focus on estimating the effect of a single systematically missing subject-level variable. Following Resche-Rigon et al. (2013), we delete all observations with sporadic missing data that occur in any variable other than the subject-level variable of interest. For each subject-level variable we consider, we will impute all missing data, as shown in Table 1.3, regardless of which study the partially observed subject is in. After each imputation, we estimate the relevant coefficient of a generalized least squares regression of CBCL-I on the subject-level variable of interest adjusted for fixed effects of study indicator, age at injury, maternal education level, injury severity, and time post-injury. For each regression model, we use a fully unstructured
covariance matrix. For comparison, we fit a model using a complete case analysis, which omits study indicator since all of the complete cases belong to the same study.

The general implementation of each imputation method is the same as described in Section 4.3.3. For the Poisson case, we only fit the joint model using the COM-Poisson model. For the multinomial case, we only use the MICE implementation of multinomial logistic regression. While we try to stick closely to the scenarios considered in our simulation, there are a few differences when extending the methods in this chapter to a real data analysis.

The first difference relates to variables that are included in the imputation models. Recall that our original motivation for this dissertation is to develop methods that can be used to construct a data bank. As such, additional information will be available in the data sets that can aid in imputation but may not be of interest in a particular analysis. For this analysis, we are primarily interested in a regression of CBCL-I on injury and demographic characteristics. However, we will incorporate the normed externalizing subscale of the Child Behavior Checklist (CBCL-E) and the normed Brief Symptom Index General Severity Index (BSI GSI) as observation-level variables in the imputation model. By including additional variables in the imputation model, we can take advantage of their associations with the missing variables to increase the efficiency of our imputation model. Thus, as in the simulation, there will be three observation-level variables and one subject-level variable in the imputation model. However, we depart from the simulation study by not including CBCL-E and BSI GSI as predictors in the analysis model.

The second difference is that we model the fixed mean of each imputation model by regressing on fully observed subject-level variables, which are treated as covariates.
Each imputation procedure includes age at injury, maternal education, and injury severity. Study indicator is also not included in the multivariate normal and MICE models because a fixed effect of study cannot be identified for either approach when all observed values come from the same study. Thus, as in Chapter 3, we make the strong assumption that the relationships between the subject-level variable we wish to impute and the other variables in the model do not differ across studies.

We can extend the joint model to include fully observed covariates by including them as fixed effects in the mean at each level of the joint model. Instead of simply a fixed intercept as in (4.5), we now model the mean with covariates. For the observation-level sub-model, we now have

$$h_m(o(\theta_{ijm})) = Z_i^{(o)} \Gamma_j^{(o)} + \xi_{ijm} + \xi_i^{(o)} \gamma_m^{(o)} + \nu_{im}$$

(4.52)

where $Z_i^{(o)}$ is a length $c^{(o)}$ vector of fully observed covariates for subject $i$ with fixed effects vector $\Gamma_j^{(o)}$. Note that the covariates are time invariant. Additional fully observed observation-level variables would be modeled using the observation-level sub-model, as shown above. Also, we allow $\Gamma_j^{(o)}$ to vary with time so that the effect of $Z_i^{(o)}$ need not be constant over time. Often $Z_i^{(o)}$ would contain demographic-type characteristics that we would want to control for in our model. In our example, we include fixed effects of time post-injury, age at injury, maternal education, and injury severity. We also include a fixed effect of study in our observation-level model because we have observed each observation-level variable in both studies.

The incorporation of covariates in the subject-level sub-model follows similarly. However, we cannot include a fixed effect of study in the subject-level sub-model for the same reason it was excluded from the multivariate normal and MICE models. Thus, we again make the strong assumption that there are no study effects
on the subject-level variable. Although we are unable to account for study on the subject-level, our joint model is able to partially account for study by incorporating it into the observation-level sub-model. The impact of doing so is an area for further investigation.

For each of the four models that we fit, the coefficient estimate, model standard error, and 95% confidence interval are presented in Table 4.9. In general, we see coefficient estimates and standard errors that vary across the four approaches. We see slightly larger absolute effects with the joint method compared to the other two imputation methods for the normal and Poisson cases, although direct comparison is difficult because we do not know the truth. As was shown in the simulation study, MICE has the smallest standard errors for the multinomial case; however, that is likely due to underestimation.

The results for the Poisson case are concerning because the estimated standard error from the joint method is larger than the standard error of the complete case analysis which indicates a loss of efficiency from the inclusion of the second study. However, as indicated in Chapters 1 and 3, the TBI data do not appear to be exchangeable and thus the simulation results raise doubts about the performance of any of the three imputation approaches in this situation. In this case, we suspect that the imputation distribution should be severely underdispersed because, unlike the simulation study, the number of children in the home was slightly marginally underdispersed. Since a Poisson variable raised issues with underdispersion in the simulation study, it is reasonable to think that a marginally underdispersed variable would only exacerbate the underdispersion in the imputation distribution.
<table>
<thead>
<tr>
<th>Effect</th>
<th>Missing Data Method</th>
<th>N</th>
<th>Estimate</th>
<th>Model SE</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Normally Distributed Covariate</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Family Hollingshead</td>
<td>Complete Case</td>
<td>139</td>
<td>-0.147</td>
<td>0.077</td>
<td>(-0.298, 0.004)</td>
</tr>
<tr>
<td></td>
<td>Normal</td>
<td>281</td>
<td>-0.132</td>
<td>0.077</td>
<td>(-0.285, 0.020)</td>
</tr>
<tr>
<td></td>
<td>MICE</td>
<td>281</td>
<td>-0.135</td>
<td>0.073</td>
<td>(-0.279, 0.009)</td>
</tr>
<tr>
<td></td>
<td>Joint</td>
<td>281</td>
<td>-0.142</td>
<td>0.070</td>
<td>(-0.280, -0.004)</td>
</tr>
<tr>
<td>Bernoulli Covariate</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low Income</td>
<td>Complete Case</td>
<td>90</td>
<td>6.395</td>
<td>2.126</td>
<td>(2.228, 10.562)</td>
</tr>
<tr>
<td></td>
<td>Normal (Rounding)</td>
<td>281</td>
<td>5.882</td>
<td>1.625</td>
<td>(2.669, 9.095)</td>
</tr>
<tr>
<td></td>
<td>MICE</td>
<td>281</td>
<td>6.332</td>
<td>1.965</td>
<td>(2.429, 10.235)</td>
</tr>
<tr>
<td></td>
<td>Joint</td>
<td>281</td>
<td>6.004</td>
<td>2.045</td>
<td>(1.991, 10.018)</td>
</tr>
<tr>
<td>Poisson Covariate</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of Children</td>
<td>Complete Case</td>
<td>141</td>
<td>-1.406</td>
<td>0.700</td>
<td>(-2.778, -0.034)</td>
</tr>
<tr>
<td></td>
<td>Normal (Rounding)</td>
<td>281</td>
<td>-1.347</td>
<td>0.652</td>
<td>(-2.635, -0.058)</td>
</tr>
<tr>
<td></td>
<td>MICE</td>
<td>281</td>
<td>-1.193</td>
<td>0.724</td>
<td>(-2.627, 0.241)</td>
</tr>
<tr>
<td></td>
<td>Joint</td>
<td>281</td>
<td>-1.370</td>
<td>0.709</td>
<td>(-2.760, 0.021)</td>
</tr>
<tr>
<td>Multinomial (Categorical) Covariate</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Income: Level 2</td>
<td>Complete Case</td>
<td>90</td>
<td>-5.777</td>
<td>2.444</td>
<td>(-10.567, -0.987)</td>
</tr>
<tr>
<td></td>
<td>Normal (Rounding)</td>
<td>281</td>
<td>-5.324</td>
<td>1.669</td>
<td>(-8.617, -2.032)</td>
</tr>
<tr>
<td></td>
<td>MICE</td>
<td>281</td>
<td>-6.288</td>
<td>1.549</td>
<td>(-9.318, -3.258)</td>
</tr>
<tr>
<td></td>
<td>Joint</td>
<td>281</td>
<td>-6.753</td>
<td>1.987</td>
<td>(-10.654, -2.853)</td>
</tr>
<tr>
<td>Income: Level 3</td>
<td>Complete Case</td>
<td>90</td>
<td>-7.159</td>
<td>2.559</td>
<td>(-12.175, -2.143)</td>
</tr>
<tr>
<td></td>
<td>Normal (Rounding)</td>
<td>281</td>
<td>-7.795</td>
<td>1.975</td>
<td>(-11.694, -3.897)</td>
</tr>
<tr>
<td></td>
<td>MICE</td>
<td>281</td>
<td>-7.309</td>
<td>1.825</td>
<td>(-10.902, -3.717)</td>
</tr>
<tr>
<td></td>
<td>Joint</td>
<td>281</td>
<td>-6.615</td>
<td>2.382</td>
<td>(-11.290, -1.939)</td>
</tr>
</tbody>
</table>

Table 4.9: Estimates, standard errors, and confidences intervals for the effect of each subject-level variable considered on Child Behavior Checklist Internalizing score for the TBI data. Estimates that result from missing data procedures are adjusted for study, age at injury, maternal education, injury severity, and time post-injury. Complete case estimates do not adjust for study, as they are all in one study. Each model assumes a fully unstructured covariance matrix. Estimates that rely on the multivariate normal, sequential conditional and joint imputation missing data methods are based on 50 imputations.
Through the four sample analyses of the TBI data, we were able to illustrate how to apply our joint model in a real data situation to impute partially observed subject-level variables with four different distributions. We also demonstrated how to extend the model by including fully observed covariates and their fixed effects. This includes the ability to include fixed effects of study in the observation-level sub-model, which is something that neither of competing approaches could incorporate. Thus, we have shown that our joint model is practical to implement and apply to real data with systematically missing subject-level variables.

4.5 Choosing the Number of Latent Variables

In our simulation and data analysis, we have always used two subject-level latent variables and one observation-level latent variable in our joint model. Fundamentally, the number of latent variables needed to adequately model a data set is directly related to the underlying correlation structure of the data. Latent variables impose structure on the correlation that can be quite rigid for few latent variables but becomes more relaxed as more latent variables are added to the model. For simply structured data, few latent variables may be required to build a reasonable model. For complex data, we may need many latent variables to be able to adequately capture the intricacies of the correlation structure. Thus, it is important to consider the complexity of the data set of interest to determine if dimension reduction via latent variables is useful or even possible.

Because of their important role in modeling the correlation structure of the data, determining the number of latent variables to include in the model is an important and challenging problem. A variety of procedures have been suggested for selecting the
number of latent variables in exploratory factor analysis. Commonly used methods are Kaiser’s Criteria (Kaiser, 1960), the Scree Test (Cattell, 1966), and the examination of the cumulative proportion of variance explained (Williams et al., 2010). However, these approaches are not designed for use with models that have latent variables on both the subject- and observation-levels. Thus, these methods are not immediately appropriate for our joint model.

We suggest an alternative approach based on comparing the fit of the imputation model for different numbers of latent variables. He and Zaslavsky (2012) devise a procedure for assessing imputation model fit based on posterior predictive replicates. Posterior predictive replicates are defined as draws from the posterior predictive distribution for variables that are observed in the data set (Gelman et al., 2004). Thus, observed values are replicated by predicted values from the posterior distribution of the imputation model. The idea behind this procedure is to compare estimates from an analysis model fit on replicated data to the estimates from an analysis model fit on the partially observed data that is completed with imputed values for missing observations (i.e., the usual completed data that results from imputation). When the imputation model is adequate, we expect the estimates from the two data sets to be similar and not exhibit systematic differences (He and Zaslavsky, 2012).

It is relatively straightforward to implement the posterior predictive check for the imputation model. Since we are focused on systematically missing subject-level variables (X), we will explain the procedure in the context of partially observed X values. We begin by choosing the analysis model that we are interested in fitting following the imputation of the missing data. Our goal is to then assess the difference between the estimated regression coefficients in the selected analysis model for the
completed data and the replicate data. Recall the observation-level variables $Y$ are always fully observed for the scenario that we consider and thus will not be replicated since no imputations of $Y$ are necessary.

The completed data can be written as $(Y, X_{obs}, X_{mis}^*)$ and consists of the fully observed $Y$, the observed $X_{obs}$, and the imputed values for the missing portion of $X$, $X_{mis}^*$. We let an asterisk denote values that are drawn from the posterior predictive distribution. We keep $L$ imputations from the posterior distribution and fit the analysis model to each of the $L$ completed data sets. He and Zaslavsky (2012) suggest using $L = 5000$.

We define the replicate data as $(Y, X_{rep}^*)$, where $X_{rep}^* = (X_{obs}^*, X_{mis}^*)$. To generate the replicate data, we draw values for each $X$ from its posterior predictive distribution, regardless of whether it is originally missing or observed. Thus in addition to imputing $X_{mis}^*$, we also replace observed values $X_{obs}$ with draws from its predictive distribution under the imputation model, $X_{obs}^*$. Note that the imputed values for $X_{mis}^*$ in the completed and replicate data are the result of different draws from the posterior predictive distribution. Again, we keep $L$ simulations from the posterior distribution and fit the analysis model to the $L$ replicate data sets. Note that only variables that are partially observed in the original data set are replicated, i.e., $Y$ is not replicated.

Assume our analysis model of interest is the following regression model which includes $b$ covariates:

$$Y_i = \left( \beta_0 + \sum_{k=1}^{b} \beta_k X_{ik} \right) 1_t + \epsilon_i$$  \hspace{1cm} (4.53)

$$\epsilon_i \sim iid \sim N(0, \Sigma)$$  \hspace{1cm} (4.54)
where $\mathbf{Y}_i$ is a length $t$ column vector for subject $i$, $X_{ik}$ is the $k^{th}$ element of $\mathbf{X}$, $\mathbf{1}_t$ a length $t$ column vector of 1’s, and $\Sigma$ is dimension $t \times t$. If our imputation model fits the data well, we expect the estimates from the analysis model to be similar for the completed and replicate data sets. Denote the $k^{th}$ coefficient estimate from the completed data by $\hat{\beta}_k^{\text{comp}}$ and the $k^{th}$ coefficient estimate from the replicate data by $\hat{\beta}_k^{\text{rep}}$. To compare the estimates, we look at the average difference between each of the coefficient estimates over the $L$ iterations. That is, we compute

$$Q_k = \bar{\hat{\beta}}_k^{\text{rep}} - \bar{\hat{\beta}}_k^{\text{comp}}, \quad k = 0, \ldots, b,$$

where $b$ is the number of terms in the analysis model. Thus, $Q_k$ is an indicator of the bias introduced to coefficient $k$ by the imputation model (He and Zaslavsky, 2012).

We can also calculate a posterior predictive p-value to aid in our assessment of the imputation model. This is simply

$$p_k = P(\hat{\beta}_k^{\text{rep}} \geq \hat{\beta}_k^{\text{comp}}).$$

That is, the proportion of the $L$ simulations where $\hat{\beta}_k^{\text{rep}}$ is greater than or equal to $\hat{\beta}_k^{\text{comp}}$. The posterior predictive p-value is used by some for formal hypothesis testing (Gelman et al., 2004), but He and Zaslavsky (2012) use it as a guide to highlight inadequacies in the model, since it has been shown to be conservative in formal testing. We agree and rather than use a formal testing procedure, suggest considering p-values close to 0 or 1 to be indicators of poor model fit (He and Zaslavsky, 2012).

To use this procedure to determine the number of latent effects, we suggest first evaluating the fit of the joint model with $q^{(s)} = 1$ and $q^{(o)} = 1$ or some other small number of factors appropriate for a particular application. After applying the posterior predictive check, add another latent variable and evaluate the fit of the new
model. We, like He and Zaslavsky (2012), suggest evaluating the fit based on a combination of the evidence from $Q_k$ and $p_k$. Ideally, for a model that fits well, $Q_k$ would be close to 0, and $p_k$ would be close to 0.5. Based on simulation results from He and Zaslavsky (2012), $p_k$ values between 0.3 and 0.7 indicate adequate fit; how close to zero $Q_k$ must be to be “close enough” will depend on the application being considered. The process of setting the number of latent variables, imputing, and evaluating the fit is repeated until the desired level of fit is reached or until $q^{(s)} = p^{(s)} + p^{(o)}$ or $q^{(o)} = p^{(o)}$. If $q^{(s)} = p^{(s)} + p^{(o)}$ or $q^{(o)} = p^{(o)}$, we should consider choosing a different type of imputation model since the latent variables no longer reduce the dimension of the problem.

As an example, consider Setting 3 for the normal subject-level variable case, as specified in Section 4.3.1. Recall we chose parameter values to give moderate correlations between the variables. Thus, we could expect to benefit from the use of latent variables in our joint model. For the joint model that we considered in our simulations with $q^{(s)} = 2$ and $q^{(o)} = 1$, the posterior predictive p-values for coefficients $\beta_1$, $\beta_2$, $\beta_3$ from (4.18) are 0.512, 0.534, and 0.504 with $Q$ values of 0.0012, 0.0005, and 0.0001, respectively. Based on our above discussion, these p-values and values of $Q$ do not give us reason to doubt the adequacy of our model. However if we change the parameters in the data generation model to

$$\beta_1 = \begin{bmatrix} 10 & -3 & 0.75 & 0.5 \end{bmatrix}'.$$
$$\beta_2 = \begin{bmatrix} 4 & -2 & 1 \end{bmatrix}'.$$
$$\beta_3 = \begin{bmatrix} 2 & -4 \end{bmatrix}'$$,
we now have data with nearly perfect correlation between all of the variables. Now using the same model with $q^{(s)} = 2$ and $q^{(o)} = 1$, the p-values from the posterior predictive check are 0.802, 0.614, and 0.767 with $Q$ values of 0.0690, 0.0033, 0.0106, respectively. These p-values are far enough from 0.5 that we would suspect that our current model is inadequate for this case of high correlations. In this case, we would consider adding additional latent variables and performing the posterior predictive check again. By doing so, we find that a model with $q^{(s)} = 2$ and $q^{(o)} = 2$ performs adequately with p-values of 0.645, 0.536, and 0.639 with $Q$ values of 0.0154, 0.0008, and 0.0023, respectively.

The above discussion highlights a potential approach for selecting the number of latent variables to include in the model. As the number of latent variables grows and approaches the number of observed variables, it becomes counter productive to continue adding latent variables because of the additional complexity introduced into the joint model. We believe that our joint model will be useful in many applied situations but caution the user to diagnose the model adequacy before using the model to impute missing data.

We emphasize that the approach outlined in this section is a suggestion for how to choose the number of latent variables in our model. We have not attempted to validate the use of this procedure. Instead, we leave further development of this imputation model selection procedure to future work.

4.6 Discussion

In this chapter, we developed a two-level joint model based on latent variables to impute systematically missing subject-level covariates. Our work extends the
observation-level model of Song et al. (2011) by adding a subject-level sub-model to enable imputation of missing variables on the subject-level. As in Dunson (2003), we construct our model within the generalized linear models framework so that we have the flexibility to appropriately model variables with different distributions within the exponential family. All of our modeling is done within a joint framework that uses normally distributed latent variables on the subject- and observation-level to capture dependence within and between variables.

Via simulation, we examined the performance of our joint model when used for multiple imputation of systematically missing subject-level covariates in the context of a generalized least squares regression analysis. As in Chapter 3, we generated data to mimic the synthesis of two studies where only one study observed the subject-level variable of interest. We investigated the performance of our model for normally distributed, Bernoulli, Poisson, and multinomial subject-level variables. For each type of variable, we examined four different observation-level covariance structures. From our simulation study, we observed that our joint method had relative bias and confidence interval coverage that was similar to or better than MICE and the multivariate normal approach (with rounding as appropriate) when estimating regression coefficients from a generalized least squares regression model.

We then demonstrated our joint model in the context of a real analysis for the TBI data. In doing so, we suggested how to extend our joint model to incorporate fully observed subject-level covariates. We showed that our joint approach is reasonable to implement in practice and provides similar estimates to those from the multivariate normal and MICE approaches.
Our results are limited in several ways. First, we chose to limit our comparison to the multivariate normal and MICE approaches to the default implementations in R through the MICE and norm packages. For example, MICE would likely have performed better in the Poisson case if we had chosen a distribution that could accommodate underdispersion; however, such a distribution is not currently available in the software package. Our critique of MICE is also limited to the implementation using the suggestion of Yucel (2008) to include the mean of observation-level variables as predictors in the subject-level imputation model.

Our implementation of the multivariate normal approach used simple rounding, rather than a more complicated calibrated rounding procedure. The use of calibrated procedures would likely improve the performance of the multivariate normal method for the Bernoulli, Poisson, and multinomial cases. However, default implementations in software tend to implement simple rounding so our comparison was to a competing approach that a practitioner would likely consider.

Another limitation is that our simulation only considered a very specific pattern of missingness. We only examined the case where we had a single systematically missing subject-level covariate. We have not evaluated the performance of our joint model for imputing more general patterns of missingness that are likely to occur in practice. For example, we have not examined cases with multiple variables with missing data or any missing data on the observation-level variables. We envision that our joint model will be able to be extended to apply to more general patterns of missingness and leave this more general evaluation of our model to future work.

A further limitation is that we did not evaluate the adequacy of our choice of the number of latent variables in our simulation or data analysis. Based on the
relatively small magnitude of the bias in estimates of the regression coefficients in
our simulation study, it appears our choice was reasonable; however, there was no
formal evaluation of this choice. It may be the case that the inclusion of additional
latent variables could have improved our joint model performance. We suggested a
possible procedure based on posterior predictive checking to aid in the selection of
the number of latent variables, but we have not evaluated the appropriateness of this
suggestion. We think this approach is promising as it is closely related to posterior
predictive measures of fit commonly used in other Bayesian analyses (Gelman et al.,
2004). Further development of the model selection procedure is left for future work.

Our results in this chapter are limited in their wider applicability to IPD meta-
analysis because we have not yet fully considered the effect of study. We used a fixed
effect of study in the analysis model of the TBI data example, but current best practice
recommends the inclusion of random study effects to reflect across-study variation
(Higgins et al., 2001; Thompson and Higgins, 2002). Likewise, we should incorporate
random study effects into our joint model to reflect across-study heterogeneity in the
imputations. We were able to include a fixed effect of study in the observation-level
sub-model of our joint model, which was not possible in the multivariate normal and
MICE procedures. The implicit assumption that accompanies the exclusion of study
effects is that the relationships between variables do not differ across studies, which
is a strong and often unrealistic assumption. However, we have built our joint model
with the intention of adding a study level to the hierarchy in the future, a natural
extension.

In our TBI application, traditional random study effects would not be applica-
table since estimation of across-study variability is not possible with only two studies.
However, we believe that a potential future path would be to include a study-level latent variable, much like we have already used on the subject- and observation-levels. This would enable us to borrow strength across observed variables and provide a potential solution to the problem of having a small number of studies. A latent variable approach that aims to model across variable study-level correlation would be difficult without an underlying joint model and thus, would be particularly challenging to try to implement within the MICE framework. Any extension involving study effects would also require a thorough exposition of the assumptions being made regarding the relationships between studies.

We acknowledge the need for more investigation for imputation of count data and the use of the COM-Poisson model in the joint model. We were not able to find any literature that referenced using a COM-Poisson model to impute missing count data. Thus, investigation into the general properties of using the COM-Poisson distribution for multiple imputation is warranted. Our empirical experience points to two major areas that deserve further exploration. First, post-imputation inferences appear to be sensitive to the choice of the prior distribution for the scale parameter of the COM-Poisson. Thus, evaluation of this sensitivity and guidance on how to choose the hyperparameters of the prior distribution is needed. The second issue relates to modeling conditional distributions that are severely underdispersed. One example of this is when we assume a marginal Poisson distribution and the correlation between variables is extremely high. In this case, conditioning on the latent variables greatly reduces the variability in the count variable. In our simulation, the joint model with a regular Poisson conditional distribution did not perform well due to this underdispersion, but such a model may be adequate if we started with a count variable
with a marginal overdispersed Poisson distribution. We recommend additional study
to more fully understand the proposed model for the count case.

Via simulation we were able to establish the utility of the joint model for imputing
systematically missing subject-level variables. Our joint model is a flexible approach
that is constructed within the generalized linear models framework, which we illus-
trated through use to impute systematically missing normally, Bernoulli, Poisson, and
multinomial distributed subject-level variables. While this chapter is limited to one
specific missing data pattern, we expect that the joint model will perform similarly for
more complex missing data patterns. We also believe that the joint model provides
flexibility to incorporate additional levels of hierarchy, like study-level effects, that
are often present in practice. We believe our joint model represents an important
step in the development of methodology to address the issue of systematically miss-
ing subject-level data in longitudinal research synthesis. Methodological advances for
accounting for missing data in longitudinal research synthesis will only become more
important as more longitudinal data continue to become available for synthesis.
Chapter 5: Summary and Future Work

In this dissertation, we discussed multiple imputation approaches for handling systematically missing subject-level variables in the context of longitudinal individual patient data research synthesis. In Chapter 3, we analytically and empirically examined the differences between the use of a sequential conditional and a joint modeling approach to multiple imputation when all variables of interest were normally distributed. We specifically focused on the performance of each method for non-exchangeable covariance structures of the observation-level variable. In Chapter 4, we developed a new a joint model for multiple imputation in the generalized linear models framework that uses latent variables to model variables on the subject- and observation-levels.

In this chapter, we briefly recap the major conclusions from Chapters 3 and 4. We then discuss potential areas of future development. We specifically focus on further investigation of our joint modeling approach and the extension of our joint model to incorporate across-study variation, as is best practice for research synthesis.

5.1 Comparing Methods for Multiple Imputation

In Chapter 3, we compared two approaches for multiple imputation of a systematically missing subject-level covariate in a two study longitudinal research synthesis.
We examined the specific case where we had a fully observed normally distributed observation-level outcome and a systematically missing normally distributed subject-level variable. We compared the sequential conditional approach implemented in the MICE R package using the suggestion of Yucel (2008), which is to include the means of observation-level variables as predictors in the imputation model, to a multivariate normal joint model.

We analytically compared the imputation distributions for each of these two methods using some simplifying assumptions and found that the imputation distribution for the sequential conditional approach attenuates the association between the subject-level variables and the observation-level variables when the observation-level variable does not have an exchangeable correlation structure. We confirmed this finding via simulation where we found that regression coefficient estimates from the sequential conditional method were attenuated for all settings we considered, except for the setting with constant variance and an exchangeable correlation structure. In addition to attenuation, the estimates from the sequential conditional method were less efficient than those from the joint method and, at times, those from the complete case analysis.

As an additional comparison, we implemented both imputation approaches to analyze the TBI data. In this data analysis, the estimated error covariance structure was reasonably close to exchangeable, and therefore the two methods produced similar results. Both methods showed gains in efficiency from the complete case analysis. However, we believe that the joint method represents a practical and conservative choice for an imputation model, particularly if little is known about the error covariance structure of the observation-level variables.
5.2 A Novel Two-Level Joint Imputation Model

In Chapter 4 we developed a two-level joint model in the generalized linear models framework that uses latent variables to jointly model variables on the subject- and observation-levels. We then used our joint model to impute systematically missing subject-level variables with distributions in the exponential family in a simulation study and for a “real” data setting. Our model has the flexibility to jointly model a wide variety of variable types that fall within the generalized linear models framework.

We evaluated the performance of our joint model as an imputation model for a systematically missing subject-level variable via simulation. We used the estimated coefficients from a generalized least squares analysis after imputation as the basis for comparing our joint model, MICE, and a multivariate normal approach. In our simulation we considered systematically missing subject-level covariates that had normal, Bernoulli, Poisson, and multinomial distributions. We also considered several different covariance structures for the observation-level variables. We found that, in terms of relative bias and confidence interval coverage, our joint model performed at least as well and often better than the comparison approaches.

We then applied our joint model as well as the two competing methods to impute systematically missing variables in the TBI data. As in the simulation, we considered cases where the systematically missing subject-level variable had a normal, Bernoulli, Poisson, and multinomial distribution. In addition, we extended our model to incorporate fully observed subject-level variables as covariates. This included a study indicator in the observation-level sub-model. We were not able to include a study effect in the subject-level sub-model and thus made the strong assumption that the distribution of the subject-level variable did not differ across studies. Neither MICE
nor the multivariate normal approach were able to adjust for study in any part of their imputation models. We saw that our joint model could be implemented in a real data setting and has the advantage of including a study effect in the observation-level sub-model. Thus, we believe that our joint model is a reasonable and flexible joint approach for imputing systematically missing subject-level data that have distributions in the exponential family.

5.3 Future Work

In this section, we consider areas for future development in the synthesis of longitudinal studies with systematically missing data. We first focus on further development and investigation of the performance of our joint model described in Chapter 4. We then shift our attention to extending our joint model to address issues specific to research synthesis. Of particular interest is how to extend our joint model to incorporate latent study effects to model between study heterogeneity.

5.3.1 Further Development of the Joint Model

In Chapter 4, we established the initial concept of a joint model for multiple imputation that models variables on both the subject- and observation-level. In this dissertation we only considered the specific scenario when a single subject-level variable was systematically missing. In practice we can expect to see a variety of other patterns of missing data, including missingness in multiple variables on either the subject- or observation-level. In addition, we would expect to see sporadic missing data (for variables at both levels) within each of the studies that we are synthesizing. While our model is flexible enough to multiply impute data in any of these situations, we have not rigorously examined its properties. However, we have no reason to believe
that our model could not seamlessly handle such situations. We would like to further evaluate the performance of our joint model for imputing missing data in these other practical settings.

Another area that we would like to further investigate is the performance of our joint model when observation-level variables are of mixed type. Our model was developed within the generalized linear models framework so that we have the ability to jointly model variables of different distributions within the exponential family. However, in Chapter 4, we only evaluated the performance for subject-level variables of different distributions. As we saw with different distributions of subject-level variables, we would expect to see some performance improvements over existing methodology when we consider different distributions of observation-level variables. We expect to see improved performance because the flexibility of our joint approach lets us treat different types of variables appropriately within the context of a joint model. Unlike approaches such as the multivariate normal, we are not forced to make assumptions about our data (i.e., that all variables are normally distributed) that we know are false. As with examining additional patterns of missing data, further examining the properties of our joint model with different combinations of variables on each level would enhance the case for the use of our model in practice.

In our simulation we did not address how to choose the number of latent variables in our joint model. We suggested the use of posterior predictive checks to aid in the selection of the number of subject- and observation-level latent variables to include in the model. We have not yet evaluated the performance of this approach to model selection. However, we believe that this is a reasonable approach to determining
model accuracy and is closely related to existing posterior predictive checks of model fit (Gelman et al., 2004).

An innovative part of our joint model is the use of the Conway-Maxwell Poisson distribution for modeling count data. To our knowledge, this distribution has not been previously used for multiple imputation. Thus, a full investigation of the properties for imputing from a COM-Poisson model would be worthwhile. Since it is not a commonly used model, establishing guidance on the selection of prior hyperparameters would also be useful. In particular, we found through our empirical experience that the imputation results appear to be quite sensitive to the choice of hyperparameters for the prior distribution on the scale parameter. Following Wu et al. (2013), we suggested that the prior variance of the scale parameter be estimated by considering a reasonable range of values. However, more investigation is needed to fully understand the impact of prior selection in the Conway-Maxwell Poisson model and to provide more guidance for practitioners.

All discussion of multiple imputation approaches thus far has been in the context of missing data that are missing at random (MAR). We have not yet considered any cases involving missing data that are not missing at random (NMAR). Future extensions of our joint model may include modeling data that are NMAR.

5.3.2 Extension to Model Heterogeneity

In this dissertation, we relied solely on fixed effects models for meta-analysis and made the strong assumption that there were no study effects for systematically missing subject-level variables. This was a necessary assumption for fitting a fixed effects model because a study effect was not identifiable for studies that had systematic
missingness. While the models we implemented in this dissertation are practical for our TBI data since we only have two studies, they do not align with best practices for conducting a meta-analysis.

The current standard for conducting a meta-analysis recommends modeling heterogeneity across studies through the use of random study effects and intercepts (Higgins et al., 2001; Thompson and Higgins, 2002). Thus, to improve our joint model, we need to consider how to model study to study variability in a reasonable way. Also by including random study effects, we will eliminate the need for the strong assumption that there are no study effects for systematically missing variables. However, we may still require an informative prior distribution if the number of studies is small.

There are two possible approaches for incorporating study effects into our joint model. The first has been used in models for sequential conditional imputation of systematically missing variables by Resche-Rigon et al. (2013) and Jolani et al. (2015). This approach specifies conditionally independent vectors of random study effects for each variable being modeled. Thus, this approach hinges on the ability to estimate the between study variability for each variable. To do this, each variable must be observed in a sufficiently large number of studies to reliably estimate the between study variability or one must specify informative prior distributions. This approach would not be viable for a two study synthesis like we have considered in this dissertation without a very strong prior distribution.

Rather than specifying independent study effects for each variable, we could think about specifying study-level latent variables that are shared across variables. Then, we could borrow strength from observed variables to provide some information on study effects for systematically missing variables. This type of approach would fit very
nicely into the framework that we established in Chapter 4. Our joint model would now have latent variables at three levels: study-level, subject-level, and observation-level. An approach of this style could also be viable for use with a small number of studies because the study-level latent variables would borrow strength from all of the observed variables within a study.

Along with a model that incorporates study to study heterogeneity, we also need to consider the assumptions that will be made as we extend the model in Chapter 4 to more complex research syntheses. We will need to explicitly state the assumptions that are being made so our analysis will be transparent and the assumptions can be critically evaluated. Without explicit statements of the assumptions of our model, it will be difficult to evaluate its appropriateness for use in applied research syntheses.

In this dissertation, we have created a general and flexible two-level joint imputation model. Since our model fits in the hierarchical and generalized linear models frameworks, there are many potential avenues for future extension from incorporating addition levels of hierarchy to incorporating additional types of variables. Our work shows the promise of this modeling approach for handling systematically missing subject-level data in longitudinal research synthesis. We believe that this work is a solid initial step toward a comprehensive joint modeling approach for dealing with missing data in longitudinal research synthesis.
Bibliography


154


156


