Exploratory multivariate longitudinal data analysis and models for multivariate longitudinal binary data

Ozlem Ilk
Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/rtd
Part of the Statistics and Probability Commons

Recommended Citation
https://lib.dr.iastate.edu/rtd/832

This Dissertation is brought to you for free and open access by the Iowa State University Capstones, Theses and Dissertations at Iowa State University Digital Repository. It has been accepted for inclusion in Retrospective Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
Exploratory multivariate longitudinal data analysis and models for multivariate longitudinal binary data

by

Ozlem Ilk

A dissertation submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of
DOCTOR OF PHILOSOPHY

Major: Statistics

Program of Study Committee:
Dianne Cook, Major Professor
Michael Daniels
Philip Dixon
Heike Hofmann
Frederick Lorenz
Sigurdur Olafsson

Iowa State University
Ames, Iowa
2004
Copyright © Ozlem Ilk, 2004. All rights reserved.
UMI Number: 3136365

INFORMATION TO USERS

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

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

UMI®

UMI Microform 3136365
Copyright 2004 by ProQuest Information and Learning Company.
All rights reserved. This microform edition is protected against unauthorized copying under Title 17, United States Code.

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346
This is to certify that the doctoral dissertation of
Ozlem Ilk
has met the dissertation requirements of Iowa State University

Signature was redacted for privacy.
Major Professor
Signature was redacted for privacy.
For the Major Program
# TABLE OF CONTENTS

## NOTATION AND PARAMETERIZATION

<table>
<thead>
<tr>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>vii</td>
</tr>
</tbody>
</table>

## 1 GENERAL INTRODUCTION

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>1</td>
</tr>
<tr>
<td>1.1.1</td>
<td>2</td>
</tr>
<tr>
<td>1.1.2</td>
<td>4</td>
</tr>
<tr>
<td>1.1.3</td>
<td>4</td>
</tr>
<tr>
<td>1.1.4</td>
<td>4</td>
</tr>
<tr>
<td>1.2</td>
<td>5</td>
</tr>
<tr>
<td>1.2.1</td>
<td>5</td>
</tr>
<tr>
<td>1.2.2</td>
<td>6</td>
</tr>
<tr>
<td>1.2.3</td>
<td>9</td>
</tr>
<tr>
<td>1.2.4</td>
<td>9</td>
</tr>
<tr>
<td>1.2.5</td>
<td>12</td>
</tr>
<tr>
<td>1.3</td>
<td>16</td>
</tr>
</tbody>
</table>

## 2 GENERAL PRINCIPLES OF EXPLORATORY DATA ANALYSIS FOR MULTIVARIATE LONGITUDINAL DATA

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>20</td>
</tr>
<tr>
<td>2.2</td>
<td>21</td>
</tr>
<tr>
<td>2.2.1</td>
<td>21</td>
</tr>
<tr>
<td>2.2.2</td>
<td>25</td>
</tr>
<tr>
<td>2.2.3</td>
<td>28</td>
</tr>
<tr>
<td>2.3</td>
<td>35</td>
</tr>
<tr>
<td>2.4</td>
<td>69</td>
</tr>
<tr>
<td>2.5</td>
<td>71</td>
</tr>
<tr>
<td>2.6</td>
<td>74</td>
</tr>
</tbody>
</table>
3 EXPLORING THE MEAN STRUCTURE IN LONGITUDINAL DATA WITH GRAPHICS: LINKED BRUSHING APPROACH 

3.1 Abstract 

3.2 Introduction 

3.3 Data notation 

3.4 Exploratory data analysis of univariate longitudinal data 

3.4.1 Faraway's approach 

3.4.2 What did Faraway miss? 

3.5 Exploratory data analysis of multivariate longitudinal data 

3.5.1 About the data 

3.5.2 Exploring mean trend conditionally by covariate gender 

3.6 Conclusion and discussion 

3.7 Acknowledgements 

3.8 References 

3.9 Appendix 

4 MARGINALIZED TRANSITION RANDOM EFFECTS MODELS (MTREM(p)) FOR MULTIVARIATE LONGITUDINAL BINARY DATA 

4.1 Abstract 

4.2 Introduction 

4.3 Motivating example: Iowa youth and families project (IYFP) data 

4.4 Model: Marginalized transition random effects models (MTREM) 

4.4.1 Complete data model 

4.4.2 Missing responses and/or covariates 

4.5 Analysis of IYFP data 

4.5.1 Results 

4.5.2 Convergence diagnostics 

4.5.3 Diagnostics for model fit 

4.5.4 An attempt to improve model fit 

4.6 Conclusion and discussion 

4.7 Acknowledgements
4.8 References .................................................. 129
4.9 Appendix A: Details on sampling from posterior distribution of
parameters in MTREM(1) ........................................ 131
4.10 Appendix B: Details on sampling from posterior distribution of
parameters in MTREM(2) ........................................ 137
4.11 Appendix C: Details on missing data in MTREM(1) ............... 145

5 GENERAL CONCLUSIONS ............................................ 149
  5.1 Exploratory data analysis ........................................ 149
  5.2 Models ......................................................... 150
  5.3 Relationship between exploratory data analysis and models .......... 151
  5.4 Future work and possible extensions ................................ 153
  5.5 References ..................................................... 155

6 APPENDIX A. MARKOV CHAIN MONTE CARLO METHODS ............ 156
  6.1 Hybrid MC ..................................................... 156
  6.2 Gibbs sampling ................................................ 158

7 APPENDIX B. DATA, SOFTWARE AND LANGUAGES ..................... 159

8 ACKNOWLEDGEMENTS ............................................. 160
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 1.1</td>
<td>Comparison of common datasets: Longitudinal, time series, and cross-sectional data</td>
<td>3</td>
</tr>
<tr>
<td>Table 2.1</td>
<td>Unequally spaced uncommon time points</td>
<td>22</td>
</tr>
<tr>
<td>Table 2.2</td>
<td>A portion of NLSY data</td>
<td>24</td>
</tr>
<tr>
<td>Table 2.3</td>
<td>Unequally spaced common time points</td>
<td>25</td>
</tr>
<tr>
<td>Table 2.4</td>
<td>A portion of IYFP data</td>
<td>27</td>
</tr>
<tr>
<td>Table 2.5</td>
<td>Equally spaced common time points</td>
<td>28</td>
</tr>
<tr>
<td>Table 2.6</td>
<td>A portion of PSID data</td>
<td>30</td>
</tr>
<tr>
<td>Table 2.7</td>
<td>A portion of acoustic data</td>
<td>31</td>
</tr>
<tr>
<td>Table 2.8</td>
<td>A portion of LSOA data</td>
<td>32</td>
</tr>
<tr>
<td>Table 3.1</td>
<td>Unequally spaced uncommon time points</td>
<td>80</td>
</tr>
<tr>
<td>Table 3.2</td>
<td>Median of raw incomes in PSID data over selected years.</td>
<td>83</td>
</tr>
<tr>
<td>Table 4.1</td>
<td>Posterior summaries for parameters in MTREM(1) and GEE estimates and standard errors for baseline and later time points</td>
<td>115</td>
</tr>
<tr>
<td>Table 4.2</td>
<td>p-values for posterior predictive checks in MTREM(1)</td>
<td>123</td>
</tr>
<tr>
<td>Table 4.3</td>
<td>Posterior summaries for parameters in updated MTREM(1) for baseline and later time points</td>
<td>125</td>
</tr>
<tr>
<td>Table 4.4</td>
<td>p-values for posterior predictive checks in MTREM(1) (repeated here for easy comparison) and in updated MTREM(1)</td>
<td>127</td>
</tr>
</tbody>
</table>
NOTATION AND PARAMETERIZATION

Notation for Multivariate Longitudinal Data

$N$: number of cases
$r$: number of responses
$p_1$: number of time-dependent covariates
$p_2$: number of time-independent covariates
$n_i$: total number of followups for case $i$, $i = 1, \ldots, N$
$t_{ik}$: time of $k$th followup for case $i$, $i = 1, \ldots, N; k = 1, \ldots, n_i$
$Y_{ikj}$: response for item $j$ in $k$th follow-up at time $t_{ik}$ for case $i$, $i = 1, \ldots, N; k = 1, \ldots, n_i; j = 1, \ldots, r$
$Y_{itj}$: response for item $j$ in $t$th follow-up at time $t$ for case $i$, $i = 1, \ldots, N; t = 1, \ldots, n; j = 1, \ldots, r$ (for datasets with equally spaced common time points)
$Y_{ik}$: response vector for case $i$ at the $k$th follow-up (at time $t_{ik}$), $i = 1, \ldots, N; k = 1, \ldots, n_i$
$x^{(1)}_{ikj}$: time dependent covariate for item $j$ at time $t_{ik}$ for case $i$, $i = 1, \ldots, N; k = 1, \ldots, n_i; j = 1, \ldots, p_1$
$x^{(2)}_{ij}$: time independent covariate for item $j$ for case $i$, $i = 1, \ldots, N; j = 1, \ldots, p_2$

Assumption: Cases $i$ and $i'$ are independent for all $i \neq i' = 1, 2, \ldots, N$. 
Notation specific for Marginalized Transition Random Effects Models

\[ Y_{ijt} \]: response at \( j \) th item for the \( i \) th case at time \( t \), \( i = 1, ..., N; t = 1, ..., n; j = 1, ..., r \)

\[ X_{ijt} \]: covariates at \( j \) th response item for the \( i \) th case at time \( t \)

\[ Z_{ijt} \]: covariates in the transition model (typically, a subset of \( X_{ijt} \))

\( \beta \): marginal logistic regression coefficients for \textit{general} model

\( \beta^* \): marginal logistic regression coefficients for \textit{initial state} model

\( \alpha_{l,m} \): serial dependence parameters within each response at time \( t(\geq 2) \) for \( l \) th item in lag \( m \)

\( \gamma_{l,m} \): serial dependence parameters within each response as a function of covariates at time \( t(\geq 2) \) in lag \( m \)

\( b_{it} \): random effects to model the correlation across responses at a given time \( t \) for case \( i \)

\( \sigma_t \): standard deviation of random effects \( b_{it} \) at time \( t \)

\( \lambda \): parameters that allow correlation across responses to differ by response type for \textit{general} model

\( \lambda^* \): parameters that allow correlation across responses to differ by response type for \textit{initial state} model

\( \Delta_{ijt} \): intercept in the transition model for the \( i \) th case at time \( t(\geq 2) \) and response item \( j \)

\( \Delta_{i,j} \): intercept in the random effects model for the \( i \) th case at time \( t \) and response item \( j \)

\( \psi \): intercept and slope parameters in covariate model
1 GENERAL INTRODUCTION

1.1 Scope of thesis

The Greek philosopher Heraclitus said “Change alone is unchanging.” Change is part of our everyday life; our physical appearances change, our health and moods change, we change schools and jobs. Change is everywhere: measuring and analyzing it has been a scientific interest for a long time.

Willett (1989) emphasized that measuring change over time is not possible with cross-sectional data. It requires longitudinal data. Longitudinal data is the result of repeated measurements from the same subject observed over time. This kind of data occurs frequently in the medical and social sciences. The common terminology for longitudinal data in the social sciences and economic literature is panel data. It can often be complex. The repeated measures induce a correlation structure that includes a within-subject component. Measurements may be taken at infrequent and unequal time points, and covariates may be time dependent. There are commonly missing values which may occur in a non-random manner. The research in this thesis develops new graphical methods and models to help analyze multivariate longitudinal data, outlines what is different in the analysis from other types of data, and builds a bridge between exploratory analysis and models.

Some questions of interest in this kind of study include how a person’s health improves on a program of diet and exercise; how an elderly person’s home health care changes as the number of impairments increase; and how a teenager’s income increases with job experience.

Figure 1.1 shows the types of longitudinal data that are regularly encountered by analysts and how this thesis addresses the different types. In this thesis, the highlighted branches are investigated, and annotation on the branches indicate which part of the research is used. For example, data with multivariate binary responses and missing completely at random cases are considered in both exploratory data analysis and models. Some possibilities are suppressed to avoid making the figure too crowded.
Figure 1.1 Some possible combinations of longitudinal data sets. The branches highlighted as "all" will be used in both exploratory data analysis (EDA) and models. Some possibilities are suppressed due to space limitations if they are not direct interest of this thesis.

For example, analysis with univariate binary responses is not included in this thesis.

Time series data, which commonly arises in economics, is also a result of repeated measurements taken over time. However, there are some differences between time series and longitudinal data. Table 1.1 compares some properties of three kinds of data: longitudinal, time series, and cross-sectional data.

1.1.1 Contributions to exploratory data analysis

We investigate the use of interactive graphical tools, particularly linked brushing and identification, to examine structure in longitudinal data. Mean and correlation structure are explored by using these tools with plots such as scatterplots and mosaic plots. Graphics are used before and after modeling. They are used for detecting errors in data collection and reconstruction, for checking model assumptions prior to modeling, for model diagnostics and output.

In Chapter 2, we discuss general principles of EDA for multivariate longitudinal data. We review and emphasize some important findings from the literature, and discuss our findings. Discussion in this chapter is based on several datasets collected throughout this thesis. In Chapter 3, we discuss EDA for
### Table 1.1 Comparison of common datasets: Longitudinal, time series, and cross-sectional data

<table>
<thead>
<tr>
<th></th>
<th>Longitudinal data</th>
<th>Time series data</th>
<th>Cross-sectional data</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>General Goal</strong></td>
<td>Inferences about mean as a function of time by accommodating dependence</td>
<td>Inferences about mean as a function of time by accommodating dependence; forecasting</td>
<td>Inferences about mean (can't measure change over time); prediction</td>
</tr>
<tr>
<td><strong>Properties of series</strong></td>
<td>Short (compared to time series); non-stationary; several series (one for each case)</td>
<td>Long; usually stationary; usually a single series</td>
<td>No series at all</td>
</tr>
<tr>
<td><strong>Measurement times</strong></td>
<td>Varies</td>
<td>Usually equally spaced common time points</td>
<td>Only one time point</td>
</tr>
<tr>
<td><strong>Measurements taken</strong></td>
<td>Response(s); Time; Time-dependent and/or independent covariates</td>
<td>Response(s); Time; Usually no covariates</td>
<td>Response(s); Time-independent covariates</td>
</tr>
</tbody>
</table>
univariate and multivariate response longitudinal data. In chapter 4, graphics are used for visualizing the output of our model.

1.1.2 Contributions to modeling

Models are examined for a smaller subset of possible longitudinal data types: binary responses. We propose models for analyzing the mean and correlation structure of multivariate longitudinal binary data. The model we propose, marginalized transition random effects model (MTREM), consists of a triple of regression models. We attempt to explain (1) the mean response by a marginal logistic regression on covariates, (2) the within-subject time dependence by a transition model, and (3) the multivariate structure at each time by a random effects model.

In Chapter 4, we introduce MTREM(p) and discuss details for the special cases of p, i.e. lag-1 and lag-2. Implementation and estimation are illustrated on a real life dataset.

1.1.3 Data

Many datasets are introduced in Chapter 2 to describe the array of different types of multivariate longitudinal data. Two datasets are used in detailed analysis: A random sample from The Panel Study of Income Dynamics (PSID), and Iowa Youth and Families Project (IYFP).

The PSID data is obtained from an article by Dr. Julian Faraway. IYFP data is collected from the Institute for Social and Behavioral Research of Iowa State University Research Park, Ames, with the permission of Dr. Rand Conger, head of the project. See section 1.2.5 for detailed information on these datasets.

1.1.4 Implementation

The software and languages, GGobi and R/S+ are used for exploratory data analysis. Fortran77 is used to program the model fitting. As of April 20th, 2004, the related code is available at the attached compact disk and from following web page: http://www.public.iastate.edu/~oilk/mtrem.html
1.2 Literature review

1.2.1 Exploratory data analysis

Graphical methods can be used to address the complexities that arise with multivariate longitudinal data, and may be used to examine a variety of aspects of the data. There is a dearth of literature on graphics. Here we describe the short list of existing papers.

To explore the mean structure in longitudinal data, Segal (1994) used regression trees by identifying homogeneous subgroups with respect to both response and covariates. Unfortunately, this approach requires the specification of the correlation structure and is sensitive to this assumption. Becker, Cleveland and Shyu (1996) included a longitudinal data example in their description of trellis displays. These static plots examine the relationship of certain variables conditioned on the other ones. Many panels can be arranged in a nested way to form a trellis display, and data can be displayed by using any univariate or higher dimensional graphical method, such as dot plot or scatterplot. However, as the number of variables or the levels of a conditional variable increase, the number of panels increase, and plot may need be viewed on several pages. Therefore, in longitudinal data applications, if time is used as a conditional variable, visualizing and exploring long time series may be a challenge. A software package for longitudinal data analysis (Bates and Pinheiro, 1997) is available in S-Plus with a specific emphasis on trellis graphics. Koschat and Swayne (1996) illustrated the use of interactive graphics for customer panel data. They applied tools such as case identification, linking multiple views and brushing not only on scatterplots, dot plots and clustering trees, but also on a new plot they introduced; case-profile plot (time series plot of a specific subject). Case-profile plots are also called parallel coordinates plot, interaction plots, and profile plots in the literature. Koschat and Swayne recommended looking at different views of the same data to avoid the false judgment on structure from one possibly spurious view. Although Faraway (1999) introduced a graphical method for exploring the mean structure in longitudinal data, it is not truly a graphical method, but more properly called a parametric modeling approach. He used simple interpolation to estimate the coefficients of a regression model, and then plotted these estimates against time to construct the parametric model.

To explore the correlation structure in longitudinal data, regression trees were also used (Segal, 1992). Although this approach can handle multiple responses, time-varying covariates, unequally spaced measurements and missing values, there are limitations. These limitations include the restriction for
continuous or ordered covariates and possible misspecification in the assumption of parametric form for covariance function. Dawson, Gennings and Carter (1997) demonstrated a Draftman's display (scatter matrix of time points) and parallel coordinate plots for exploring the correlation structure in repeated measures. Unfortunately, these methods fail in the precedence of uncommon measurements. Heagerty and Zeger (1998) introduced lorelogram to explore the association by the use of log-odds ratio for longitudinal data with categorical responses. This display overcomes the problem of marginal variance being a function of mean in categorical responses and the problem of correlation being constrained by the mean in binary responses. However, it is a regression approach, which requires the estimation of a function, and hence is not exactly in exploratory nature.

In this thesis, interactive graphics and related tools such as mosaic plots, scatterplots and linked brushing are commonly used as exploratory tools. Interactive graphics are known to be beneficial in exploring high dimensional data. Even in simple data sets, they accelerate the exploratory process in terms of analyst's time over static graphics. Mosaic plots (Hartigan and Kleiner (1981), Friendly (1999) and Hofmann (2000)) are one of the most common ways of visualizing categorical data. Mosaic plots are extensions of histograms in the sense that more than one variable is used. A histogram is a graphical display of one variable at a time, where counts are represented by the height of a rectangle. Mosaic plots are useful for discovering the dependencies between several variables, and exploring extremely small or high counts. Scatterplots are two dimensional displays of variables. They are used for detecting any marginal or joint outliers, patterns or unusual observations. They are also good for checking the dependence of two variables. Brushing is a direct manipulation technique in which we simply change symbols or colors of a point or a group of points in a plot. Linking is connecting multiple related views simultaneously. In linked brushing, by changing the symbols or colors in one plot, we are able to observe the corresponding changes in linked plots. A brief summary of interactive graphical methods are given by Newton (1978), McDonald (1982), Buja et al. (1996) and Swayne et al. (1998).

1.2.2 Models

Although exploratory data analysis is necessary for understanding data on a deeper level at the early stages of analysis, modeling can provide simple summaries of the main trends, parsimoniously quantify effects of covariates, and allow for prediction of future data. There is a wealth of literature on modeling longitudinal data. Some recent work include Roy and Lin (2000), Verbeke and Molenberghs
(2000), Pourahmadi and Daniels (2002), and Singer and Willett (2003). While models for continuous responses are frequently proposed because of the flexibility of normal distribution assumption, modeling longitudinal data with binary response can be more challenging. Most of the models recommended for discrete response can be categorized into marginal, transitional and/or random effects models. Albert (1999) gives a brief description of these models. However, there is still little research on models for multivariate binary data.

Some of the models proposed for binary multivariate data use random effects models and/or latent variables to exploit the multivariate structure of responses (Bandeen-Roche et al. (1997) and Legler and Ryan (1997)). However, determination and computation of marginal covariate effects with these models can be difficult.

Other models were proposed that directly model the marginal mean (Lang et al. (1999), Fitzmaurice and Laird (1993), and Heagerty (1999 and 2002)). Fitzmaurice and Laird (1993) used a marginal logistic model to explain mean structure and conditional log odds-ratios to model the time dependence in longitudinal binary data. However, since the log odds-ratios considered are conditioned on all observed responses, and not just on the history (previous responses at a given time point), this model does not exploit the longitudinal nature of correlation.

Heagerty (1999 and 2002) introduced marginally specified logistic-normal models and marginalized transition models (MTM) for univariate binary data. In both models, he started with a marginal logistic regression model for explaining the average response. The model specification is completed by adopting a random effects model in the logistic-normal models, and a transition model for MTM for explaining the within-subject dependence which impose constraints on certain parameters. Parameter estimations in both papers were handled by maximum likelihood and/or estimating equations.

We propose models for handling the multivariate longitudinal binary data which exploits the longitudinal structure. Specifically, we propose marginalized transition random effects model, MTREM(p). These models exploit the marginal covariate effects while accounting for a within-subject time dependence and multivariate response dependence. Markov Chain Monte Carlo Methods, specifically Gibbs sampling with Hybrid steps, are used to sample from the posterior distribution of parameters in our model.
In complex models, Markov Chain Monte Carlo (MCMC) methods allow the investigator to simulate from the posterior distribution of parameters. A brief summary of these methods is given in Robert and Richardson (1998), and some computational issues are discussed in the Appendix of this thesis. The Metropolis-Hastings algorithm, Gibbs sampler and Hybrid MC are some examples of these methods. In Metropolis-Hastings (Metropolis et al. (1953) and Hastings (1970)), one selects an easy to simulate candidate density, and then accepts or rejects the candidate state with a probability. In some cases, this is an approximation to the true density. The acceptance rate should be as high as possible here (i.e. close to 1). In random walk version of this approach, the candidate state is sampled from the previous state with some noise. Average acceptance rate should be set to be between 25 to 45%. The Gibbs sampler, (Geman and Geman (1984), Gelfand and Smith (1990)), is one of the best known MCMC methods. It is a special case of Metropolis-Hastings algorithm, where the candidates are always accepted. Each component of the parameter vector is updated by sampling from the full conditional distribution of each parameter conditioned on the current value of the other parameters. Under certain conditions, the stationary distribution is the joint posterior distribution. MC, which was originally introduced in statistical physics (Duane et al., 1987) has been applied to Bayesian analysis by Neal (1993, 1994) and Gustafson (1997). This method makes use of gradient information, suppresses random walk behavior and updates all parameter vector simultaneously; when tuned properly, it explores the parameter space fast. Moreover, even though the calculations are based on true target density, unlike Gibbs sampling full conditionals do not have to be in a parametric form that can be easily simulated. In this method, acceptance probability is suggested to be around 90%. Since using Metropolis-Hastings or Hybrid methods might facilitate the computations when the full conditionals are in a hard to sample form, they can be used within a Gibbs sampler. This will be our approach here.

Diagnostics are an important part of MCMC determining whether methods converge to the target distribution. Although the theory behind MCMC guarantees that convergence will occur, it should be decided when this occurs. In other words, the researcher should decide whether the chain has run long enough that the samples from posterior distribution are representative of the target distribution. Cowles and Carlin (1996) gave a brief comparative review of convergence diagnostics in the literature. They compared 13 diagnostic methods in terms of the ease of use, whether method is graphical or quantitative, and whether method requires single or multiple chains. Free user-friendly software Bayesian Output Analysis (BOA) offers some convergence tests, as well as descriptive analysis. This software
runs under SPlus or R, it is available from http://www.public-health.uiowa.edu/boa

1.2.3 Relationship between exploratory data analysis and models

Even though exploratory data analysis can capture patterns that are too complex to model, models can help make more affirmative decisions and drawing inferences. Exploratory data analysis can be used separately or together with models to analyze the data. When it is used alone, significance of findings is usually questioned. There is some recent development on inferences for data visualization by using permutation tests (Buja, 1999). When they are applied together, after the data is analyzed with graphics, investigator may be tempted to introduce new effects and tests suggested by the data. However, this is sometimes called data snooping (Neter et al. (1996), page 724), and it is usually not recommended since that can introduce bias in the model estimates. Graphical techniques can be introduced before and after fitting models for a couple of reasons. For example checking the validity of assumptions used in models and diagnostic analysis. Applying exploratory data analysis and models together leads to stronger and well justified conclusions.

Tukey (1980) emphasized the importance of using both exploratory and confirmatory analysis together. He pointed out that usually finding the question is harder than finding the answer, and the scientists start with the “idea” of question. This leads to the question and design, which lead to the collection of data, analysis of data, and finally to the answer. He advocated that neither exploratory nor confirmatory analysis are sufficient alone to implement this paradigm.

Although, the importance and necessity of combining exploratory and confirmatory analysis are discussed in the statistics community, there has not been much of an attempt to build a bridge. Building this bridge for longitudinal data is one of the purposes of this thesis.

1.2.4 Missing responses and/or covariates

Missing cases are common in most statistical applications and occur when observations are lost or unavailable for a measurement that was intended to be collected. They occur even more frequently in longitudinal studies because of (long) duration of data collection over time. This should not be confused with incomplete data by the nature of study design (Diggle, et al. 2002). That is, if we decide in
advance to take measurements in unequally spaced time points, we will have unbalanced data, but, they are not missing cases. In the first situation, we may not know why data is incomplete, and moreover, the pattern in missing cases may have an effect on our analysis.

Subjects might dropout or might be missing intermittently. In the latter case, information about the reason of nonresponse can be determined once subjects are back in the study. Missings in panel studies may occur for many reasons. These reasons include:

- Subjects might enter study after data collection started.
- We might lose track of subjects because of death, or illness, or moving out of the project area.
- Subjects might simply be unwilling to cooperate after some time, especially in long term studies.
- Subjects might be unwilling to answer some private questions, such as age and/or income.

Besides the fact that unobserved cases create unbalanced data that is usually harder to analyze, they also create a concern about the reason why they are missing. In other words, whether missing mechanism is related to measurement process is an important question. For example, in a clinical trial assessing the effect of a new drug, did a patient miss a checkup because the drug made him too sick to come to the clinic, or because he had a work related emergency? If it is the first case, nonresponse contains important information, and should not be ignored.

Rubin (2000) classifies missing data mechanisms as missing completely at random (MCAR), missing at random (MAR), and nonignorable (NI) nonresponse (also called informative). If missigness does not depend on the values of missing or observed data, then it is called MCAR. When missigness is assumed to depend only on the observed components of the data, it is called MAR. NI nonresponse mechanism occurs when missingness depends on the missing values of the data. These mechanisms might effect our choice of analysis. For example, generalized estimating equations (GEE), a commonly used estimation method in longitudinal analysis, gives biased results when missing pattern is not completely at random. Robins et al. (1995) proposed an extension of GEE estimators based on inverse probability weighting that can handle MAR. Later, they proposed an extension that allows NI nonresponse (Rotnitzky et al. (1998)). Therefore, if MAR or NI pattern is suspected, one should avoid using GEE, and consider these extensions.
Although it is not possible to extract missing data patterns from observed data, exploratory data analysis is useful in exploring distributions conditional on missing cases. Swayne and Buja (1998) discussed the use of interactive graphics in exploring missing data structure. They suggested that visualizing missing data can inform the investigator on the number of missing cases, the association of missing cases with other variables and the adequacy of imputations. They used two windows, one for recorded data and one for missingness data, and linked these two. An illustration of this method will be given in Chapter 2.

Other work on visualizing missing cases is the implementation in MANET, an interactive software for categorical data. This package treats unavailable cases as a separate category. By examining the empty-bin pattern, one is able to see the proportion of missing cases to observed ones. Illustrations of this technique for the infamous “Titanic data” are given on page 42 of Hofmann (2001).

Model fitting also requires dealing with missing cases. Two common methods for handling unobserved cases are complete case analysis and imputing reasonable values. The first one, which discards all incomplete cases, is usually not recommended since this is a waste of data and typically introduces bias (Little and Rubin, 1987). There is a broad literature on imputation under different missing data patterns (e.g. Rubin (2000)). Laird (1988) gave a brief discussion on missing data in longitudinal studies. In this paper, the emphasis is on the likelihood based methods, such as maximum likelihood estimation under both ignorable and non-ignorable missing patterns.

When closed form of likelihood is difficult to obtain, Bayesian data analysis is frequently used. Data augmentation (Tanner and Wong, 1987) was introduced for the calculation of posterior distributions when there are missing cases. This method is an iterative algorithm consisting of two steps: imputation and posterior. In the imputation step, one generates the missing data given the observed data. The posterior step involves sampling from the model parameters given the augmented data of missing and observed. Algorithm iterates between these two steps. This algorithm assumes MAR, but can be augmented for nonignorable missingness.

In longitudinal studies, missing values in covariates are also common. For normally distributed covariates, least squares methods or model based methods (such as maximum likelihood, Bayesian methods, and multiple imputation) are frequently used. See Little (1992) for a review. However, when
covariates are categorical or mixed, very little literature exists on how to handle missing cases. Maximum likelihood and least square methods are usually not appropriate since imputed values might lie outside the range. To handle missing data problem with mixed covariates, Little and Schluchter (1985) and Schafer (1991) provided EM and Bayesian algorithms, respectively. One of the most recent work on this area is by Ibrahim et al. (2002). Their methods assume a parametric distribution for covariates in Bayesian inferences. One writes the joint distribution of covariates as a product of one-dimensional sequential conditional distributions. This hierarchical structure lets us to handle both continuous and categorical covariates. Although they illustrated their method on examples with only time independent covariates, extension to time-varying covariates is easy, and illustrated in our framework (Chapter 4). Their method also assumes MAR.

1.2.5 Data

Two data sets are used in depth in this thesis: Iowa Youth and Families Project (IYFP), and The Panel Study of Income Dynamics (PSID).

Iowa Farm Crisis and Iowa Youth and Families Project Data (IYFP)

In 1970s, as a result of a big concern for food shortage in the world, government and banks supplied low interest farm loans to increase food production. Farmers used this opportunity to invest in their lands, such as buying new machines. However, these golden years of agriculture lasted only into the 1980's, when the Federal Reserve Board cut its budget for farm loans (Geller, 1986, and Conger, et al., 1994). A bank manager, in the movie “Country,” tells a farmer “If you want loan for traveling around the world or for a Mercedes, it is yours. But, don’t ask for a farm loan.” This rapid change in the economic situation caused a crisis all over the country, not only in the farm sector, but also in all other food production and distribution related sectors. Iowa, especially rural parts of it, was grossly affected by this crisis, popularly known as the Farm Crisis.

Iowa Youth and Families Project (IYFP) is a panel study with an interest on understanding the impacts of economic hardship on family members’ well-being. The project started in 1989 with 451 Iowa families. Targets were 7th graders in 1989, with two married biological parents and a sibling within four years of age. For the data used in this thesis, families were followed through 1992, once each year, and then in 1994, and then once in every other year (1995, 1997, 1999).
The study area of IYFP is selected because of their farm related income, and because of their
closeness to the project center in Ames (Elder et al., 2000). Since data collection for some part of
IYFP requires videotaping, minimizing travel cost and time was an important factor in the selection of
eight counties on the north of Des Moines: Webster, Hamilton, Hardin, Marshall, Humboldt, Wright,
Franklin, Butler. This region was relatively homogenous in terms of cultural, social and economic situations. Twenty percent of the families in the study were full-time farmers, while 25% had no connection
with a farm at all. Among the rest, 10% were classified as part-time farming families (typically em­
ployed off the farm), another 10% as displaced farming (did not continue farming after 1980s crisis),
and 35% were farm-reared farmers (grew up in a farm, but not in a farm anymore).

One of the main purposes of IYFP was to investigate the relationship between economic pressure
and its effect on family relationships and on well-being of family members. Other studies on Iowa
Farm Crisis reported that economic hardship led to life style adjustments, which in turn caused distress
caused by economic crisis was the best predictor of distress. Moreover, the researchers in 1970s and
1980s reported a modest correlation between negative events in subject’s life and distress (Lin and
Ensel, 1989).

The studies on Iowa Youth and Families Project points out similar results to these. Conger et al.
(1994) stated that farm crisis of 1980s had a long term effects on these families in terms of relationships
and individual emotional status. They also concluded economic hardship causes daily hassles and hence
distress on parents, which in turn effects the well-being of their children through harsh parenting.

A large number of studies found that girls are more likely than boys to start feeling depression
between the ages 13 to 15; and adult females are two to three times more likely to show depressive
symptoms compared to males (eg. Culbertson, 1997, Ge et al. 1994). Ge et al., (2001), investigated the
gender differences in depressive symptoms on 5 wave lengths of IYFP data, and reported that signifi­
cant differences were observed between 8th grade and 12th grade subjects. Moreover, they concluded
that early-maturing girls with initial depressive symptoms and recent negative life events were the most
likely ones to be depressed.
Although many studies investigated the Iowa Farm Crisis, not many have the flexibility to examine the social and economical changes over eleven years. This thesis uses IYFP data from 1989 through 1999, with eight observed time points, and observes the changes from adolescent to adulthood. Gender differences, impact of negative events and economic pressure on distress are of high interest over these years. However, our main purpose with this dataset is using its rich and somewhat complex structure in illustrating our statistical methods.

Some of the questions of interest we would like to address with this data include how transition from childhood to adulthood, and how economic or social problems, effect psychological well-being. We are especially interested in exploring the effect of gender and age on distress. More specifically, the hypothesis we are interested in follows:

1. There is a relationship between age of the target and distress.
2. There is a relationship between gender of the target and distress.
3. There is a positive relationship among distress measures (i.e. bivariate and multivariate relationships: depression and anxiety, depression and hostility, anxiety and hostility, depression and anxiety and hostility)
4. There is a positive relationship between economic hardship and distress.

Limitations of the data in this thesis:

As mentioned above, one criteria in the selection of study area was its proximity to research center. Elder et al. (2000) reports that these families are not a representative of rural parts of Iowa or Midwest. They also reported that since the study includes children born only in late 1970’s, it is a single birth cohort study, and for general inference purposes, the study needs to be replicated. However, Conger et al. (1994) mention that even though this sample may not be representative of Iowa farm families, it is rich in terms of variation, and is of interest.

The worst period of Iowa Farm crisis was between 1984 and 1986 (Hook, 1990). Since the data for IYFP is collected starting from 1989, it is hard to conclude that distress observed in our sample is a result of Farm crisis. Moreover, since targets are going through pubertal transitions during these early stages of project, distinguishing the reason of distress becomes more challenging.
In IYFP, multiple self-reports from each family were collected: target, mother, father and sibling of target. Videotapes of families were also available in some waves. In this thesis, we included only questionnaire self-reports from the targets and their mothers. Even though mothers reported more distress and economic pressure than fathers did (Elder, et al. (2000), and Conger et al. (1994)), we preferred to avoid including fathers' responses. Father's responses are only included if mother is not in the dataset (only 8 families over eight waves). Investigating both mother's and father's responses for eight time points would lead to an even more complex data structure than is reasonable to address here.

Some studies recommend including positive events along with negative events (Ge et al. 2001), but, IYFP only recorded negative events. Personal distress might be better explained with both types of data.

As a final limitation, the income variable used in this thesis does not include farm income. Only 20% of families are involved with full-time farming, and income is used as a moderating variable, rather than a primary variable in this analysis. Lorenz, et al. (1994) pointed out that although lower income families and families who had income loss were more in danger of depression, the impact of income on depressive symptoms disappeared once the economic pressure variables were introduced into the path model analysis.

The Panel Study of Income Dynamics Data (PSID)

PSID is an ongoing longitudinal study conducted by the Survey Research Center in University of Michigan. The original data is a representative sample of U.S. population including 4,800 families at the beginning of the study. The main purpose of the study is to examine the change in economic and demographic aspects of the subjects throughout their lives. Between 1968 and 1996, subjects have been contacted once a year for face-to-face or phone interviews. After this time, interviews were conducted once every other year.

In this thesis, a random sample of the full data is used, as analyzed in Faraway (1999). This random sample consists of 85 household heads. The response variable is annual income observed over years 1968-1990. Three covariates, age at onset of study, years of education, and gender are used. Since the data used is a subsample and some important variables are ignored, our purpose is not to draw conclusions about the subject matter, but to illustrate the graphical methods.
1.3 References


2 GENERAL PRINCIPLES OF EXPLORATORY DATA ANALYSIS

FOR

MULTIVARIATE LONGITUDINAL DATA

2.1 Introduction

Longitudinal data are repeated measurements from the same case over time. They are common especially in medical and social sciences. The case can be a patient in a clinical study, families in a social study survey, or location in environmental data.

Longitudinal data comes in very different forms. Measurements can be taken at different and irregular time points. There may be more than one response variables, which might be binary, ordinal or continuous. Covariates may be measured for each case, or for each time point or in the most difficult case completely different time points.

We introduce new language to describe the different types of longitudinal data. We call the data unconstrained when measurements are taken at irregular time intervals. The data is constrained when measurements are recorded at the same time points, and fully constrained when they are at regular intervals.

Both models and visualization methods can differ greatly depending on these types of longitudinal data. Our purpose in this chapter is reviewing and illustrating exploratory data analysis methods for these different types of longitudinal data.
2.2 Types of longitudinal data and examples

This section overviews the diversity in types of longitudinal data, defines the notation for each and gives examples.

For a discussion on creating a longitudinal dataset, we refer the reader to Chapter 2 of Singer and Willett’s book. They discuss in detail the difference between person-level data set and person-period data sets. Their webpage also provides computer codes that can convert one format to another. In person-level data, each person has only one row of records. In person-period data, each person has multiple rows of records, each one corresponding to different measurement times. They recommended use of the latter one for a couple of reasons, including that the person-level data set does not include information about measurement times, and is inefficient for study designs with irregular and uncommon time points. Moreover, they advocate that this format cannot handle time-varying covariates well, since for each such covariate, one needs to include additional columns. However, with person-period format, one needs to duplicate the same information for each wave with time-independent covariates. For instance, one needs to repeat the information that subject is male in each wave (row) he has been in the study. Applying Cleveland’s (1994) data-ink ratio idea on graphics to tables, a person-period data set is a clear disadvantaged.

Although a person-period data set is recommended in our data analysis, we introduce a different format in order to distinguish different types of longitudinal data. Our format is basically a person-level format that includes measurement times as a variable.

Data sets introduced in this section are labeled with names in parenthesis for the purpose of easy recalling in the later sections.

2.2.1 Unconstrained: Irregular, uncommon time points

We call this data unconstrained because in a modeling framework, there would be more parameters needed to define the model than in datasets with regular time points. In this type of data, cases are observed at irregular time points. The data structure is illustrated in Table 2.1. For case $i, i = 1, 2, ..., N,$ responses $Y_i = matrix(Y_{i1}, ..., Y_{im})$ are measured at time points $t_i = vec(t_{i1}, t_{i2}, ..., t_{im}).$ For instance, say, subject 1 is followed up at months 1, 3, and 4, while case 2 is observed at 2nd, 5th, 8th
Table 2.1 Unequally spaced uncommon time points

and 10 th months. In addition to responses, time-dependent covariates, $X_t^{(1)} = \text{matrix}(X_{t1}^{(1)}, ..., X_{tn_i}^{(1)})$, and time-independent covariates, $X_t^{(2)}$ are observed at time points $t_{ik}, k = 1, ..., n_i$.

Specifically, in $k$ th follow-up for case $i$, the response is an $r$-dimensional vector, $Y_{ik} = \text{vec}(Y_{ik1}, ..., Y_{ikr})$, time-dependent covariates is $p_1 \times 1$ vector, $X_{ik}^{(1)} = \text{vec}(x_{ik1}^{(1)}, ..., x_{ikp_1}^{(1)})$, and time-independent covariates is $p_2 \times 1$ vector, $X_{i}^{(2)} = \text{vec}(x_{i1}^{(2)}, ..., x_{ip_2}^{(2)})$.

Example: National Longitudinal Survey of Youth (NLSY)

The purpose of this study is to track the labor experience of high school dropouts. Subjects are males between the ages of 14 and 17 at the beginning of study. Since subjects may drop out of school or change jobs at different times, taking observations at fixed time points is not appropriate. For this reason, interview time is recorded as roughly the year after subject's first day at his first job. The measurement times vary from one individual to another, and also the number of measurements taken varies. For instance, for one subject interviews were held 1.9, 2.8 and 4.3 years after the first day at job, while for another subject 10 interviews ranged between 0.1 and 9.1 years.

For more information and analysis on the data, see page 146 of Singer and Willett (2003). Data (wages*), along with many other longitudinal datasets, is available from the following webpage:
http://www.ats.ucla.edu/stat/examples/alda.htm

The response variable, $Y_{ik}$, is the natural logarithm of wages (LNW) for subject $i$ ($i = 1, ..., 888$) at interview $k$.

The time-varying covariate is $x_{ik}^{(1)}$, unemployment rate in the local geographic area ($URATE_{ik}$) for $i$th person at followup $k$, while the time independent covariates are, $X_i^{(2)}=(BLACK_i, HGC_i)$. BLACK is an indicator for race/ethnicity, and HGC is the highest grade completed. $EXPER_{ik}$ is the time that measurements are taken.

A portion of this data is illustrated in Table 2.2. The measurement times vary from subject to subject, and also the numbers of follow-ups vary. That is, we have irregular uncommon time points.

**Example: Baltimore Longitudinal Study of Aging (BLSA)**

BLSA is an ongoing longitudinal study observing the human aging process. It started in 1958, and subjects are scheduled to be measured once every two years. The study included more than 1400 men. Once the study started, however, this planned schedule could not be followed. Some cases had more than one visit within a year, while other cases had 10 years between two consecutive visits. Among many measurements are the prostate and hearing data discussed below. More detailed description can be found in Verbeke and Molenberghs (2000), pages 10-14.

**The prostate cancer data (Prostate)**

With an aim to detect prostate cancer early, this study investigates a marker, prostate-specific antigen (PSA). Although high PSA level does not necessarily imply cancer, change over time might be an important signal.

Response, $Y_{ik}$, for this dataset is PSA for patient $i$ ($i = 1, ..., 54$) at follow-up year $k$.

There are no time-varying covariates, and the only time-independent covariate is
Table 2.2 A portion of NLSY data

<table>
<thead>
<tr>
<th>Case(i)</th>
<th>Follow-up(k)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>1.49,015,3.2</td>
<td>10.0</td>
</tr>
<tr>
<td>1.43,715,3.2</td>
<td></td>
</tr>
<tr>
<td>1.47,173,3.2</td>
<td></td>
</tr>
<tr>
<td>2.80,596,2.6</td>
<td></td>
</tr>
<tr>
<td>2.13,698,4.8</td>
<td></td>
</tr>
<tr>
<td>2.12,1.21,13</td>
<td></td>
</tr>
<tr>
<td>1.74,1.41,6.7</td>
<td></td>
</tr>
<tr>
<td>2.22,2.07,4.9</td>
<td></td>
</tr>
<tr>
<td>2.13,2.38,4.2</td>
<td></td>
</tr>
<tr>
<td>2.35,2.76,2.2</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>2.03,1.87,9.2</td>
<td></td>
</tr>
<tr>
<td>2.3,2.81,11</td>
<td></td>
</tr>
<tr>
<td>2.48,4.31,6.3</td>
<td></td>
</tr>
<tr>
<td>2.09,5.96,2.6</td>
<td></td>
</tr>
<tr>
<td>2.13,6.98,4.8</td>
<td></td>
</tr>
<tr>
<td>2.09,5.96,2.6</td>
<td></td>
</tr>
<tr>
<td>2.13,6.98,4.8</td>
<td></td>
</tr>
</tbody>
</table>

$X_{i}^{(2)}$: 4 treatment groups (Control group, BPH (benign prostatic hyperplasia) group, L or R (local or regional) cancer cases, metastatic cancer cases)

The hearing data (Hearing)

The purpose of this data is to study the change in hearing thresholds' of elderly males.

$Y_{ik}$: hearing threshold sound pressure levels for male i ($i = 1, ..., 681$) measured at 11 different frequencies on both ears at follow-up year k.

$X_{ik}^{(1)}$: indicator of which ear measurement is taken from (right or left)
Example: Effect of pH on drying rate of holly leaves (Holly)

This experiment was conducted to study the effect of pH level on the drying rate of holly leaves. Forty plants were placed on four different pH levels (ten plants for each level), and leaves were allowed to dry. Their weights were measured at 13 irregular time points over three days. This example is taken from Diggle, et al. (2002), pg 120.

$Y_{ik}$: ratio of current to initial weight of plant $i$ ($i = 1, ..., 40$) at follow-up $k$ ($k = 1, ..., 13$).

$X_i^{(2)}$: (pH level (2.5,3.5,4.5,5.6(control)))

Measurements are taken on irregularly spaced time points.

2.2.2 Constrained: Irregular, common time points

Table 2.3 Unequally spaced common time points

In constrained longitudinal data structure, measurements for each case are taken at same times. Data structure is illustrated in Table 2.3. For case $i$, $i = 1, 2, ..., N$, responses $Y_i = matrix(Y_{i1}, ..., Y_{in})$ are measured at unequally spaced time points $t_i = vec(t_{i1}, t_{i2}, ..., t_{in})$. In addition, at these same time points, we observe time-dependent covariates, $X_i^{(1)} = matrix(X_{i1}^{(1)}, ..., X_{in}^{(1)})$, and time-independent co-
variates, $X_i^{(2)}$.

The constraint introduces some simplification on the notation related to time. The difference from previous type of data is that $t_i$ is same for all $i$. That is, $t_{ik} = t_k$. Note that, for this data $t_i(k+2) - t_i(k+1) \neq t_i(k+1) - t_{ik}$ for some $k$.

Example: Iowa Youth and Families Project Data (IYFP)

IYFP is a longitudinal study with an interest on understanding the impacts of economic hardship on family members' well-being. Project started in 1989 with 451 Iowa families. Targets were 7th graders in 1989, with two married biological parents and a sibling within four years of age. Targets and their families were followed up once a year initially, and then once every two years.

$Y_{ik}$ is a three dimensional vector, where the first component, $Y_{ik1}$, corresponds to anxiety record for case $i$ ($i = 1, ..., 451$) at year $k$ ($k = 1989, 1990, 1991, 1992, 1994, 1995, 1997, 1999$). The second and third components of response vector are hostility and depression records for this case at this particular year.

$X_{ik}^{(1)} = (\text{income, material needs, cutbacks, ends meet, concerns, negative economic events, negative life events, household size})$

$X_i^{(2)} = (\text{gender, indicator of whether target is out of school in 1994})$

A portion of data is given in Table 2.4. Note that, all subjects have measurements taken at the same years. However, while $t_2 - t_1$ is 1, $t_5 - t_7 = 2$. That is, study design is taking measurements at unequally spaced, but common time points.

Example: Vorozole Study (Vorozole)

This data is discussed in Verbeke and Molenberghs (2000), page 15. The data is used to compare the overall life quality of breast cancer patients taking a standard or new drug. The Functional Living Index: Cancer (FLIC) is used to measure quality of life. Higher values of FLIC are higher quality.
Table 2.4 A portion of IYFP data

<table>
<thead>
<tr>
<th>Year</th>
<th>(Anxiety, Hostility, Depression)</th>
<th>(Income,...,Household size)</th>
<th>Gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>1989</td>
<td>(1,1,1.6) (32000,...,6)</td>
<td>(NA,NA,NA) (NA,...,NA)</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>(1,2,1,3,1.5) (10000,...,4)</td>
<td>(16000,...,4)</td>
<td></td>
</tr>
<tr>
<td>1990</td>
<td>(1,1,5,1.4)</td>
<td>(NA,NA,NA) (NA,...,NA)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1,2,1,3,1.5) (10000,...,4)</td>
<td>(16000,...,4)</td>
<td></td>
</tr>
<tr>
<td>1997</td>
<td>(1,5,1,3,2,2) (5500,...,4)</td>
<td>(NA,NA,NA) (NA,...,NA)</td>
<td></td>
</tr>
<tr>
<td>1999</td>
<td>(1,8,2,1.7) (12000,...,3)</td>
<td>(NA,...,5)</td>
<td>1</td>
</tr>
<tr>
<td>1990</td>
<td>(1,2,1,6,1.6) (1900,...,5)</td>
<td>(1,1,1) (40000,...,1)</td>
<td></td>
</tr>
<tr>
<td>1997</td>
<td>(1,1,1) (66500,...,1)</td>
<td>(1,1,1)</td>
<td></td>
</tr>
</tbody>
</table>

Follow-up(k)

Although the response is ordinal in nature, it has more than 70 categories and can be considered to be continuous in the analysis.

\[ Y_{ik} : \] FLIC at baseline, months 1, 2, 4, 6, 8, 10, ..., 44 (i.e. one every other month after month 2).

\[ X_{ik}^{(1)} = \text{dominant site of disease, clinical stage} \]

\[ X_{ik}^{(2)} = \text{treatment(new drug, Vorozole or standard drug, megestrol acetate)} \]

Observations are taken at unequally spaced but common time points until disease progression or death.
Example: Body-weights of cows (Cows)

This data is described in Diggle, et al. (2002), at page 103. Experiment consists of two factors: iron dosing and infection with M. paratuberculosis. Twenty six cows are assigned to one of four possible groups. The response variable is body weight.

\[ Y_{ik} : \text{body weight of cow } i \ (i = 1, ..., 26) \text{ at day } k \]
\[ (k = 122, 150, 166, 179, 219, 247, 276, 296, 324, 354, 380, 445) \]

\[ X_{i}^{(2)} = \text{ (treatment groups (control, iron dosing, infection with M. paratuberculosis, iron dosing and infection))} \]

Observations are taken at unequally spaced but common time points.

2.2.3 Fully constrained: Regular, common time points

Table 2.5 Equally spaced common time points

<table>
<thead>
<tr>
<th>Case(i)</th>
<th>N</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>\ldots</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Follow-up(k)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

An additional constraint of equally spaced time points produces fully constrained longitudinal data. Here, the data is balanced. The data structure is illustrated in Table 2.5. For case \( i \), \( i = 1, 2, ..., N \), responses \( Y_i = \text{matrix}(Y_{i1}, ..., Y_{in}) \) are measured at equally spaced time points \( t_i = \text{vec}(1, 2, ..., n) \). In ad-
dition, at these same time points, we observe time-dependent covariates, \( X^{(1)}_i = \text{matrix}(X^{(1)}_{i1}, \ldots, X^{(1)}_{in}) \), and time-independent covariates, \( X^{(2)}_i \).

The notation is much simpler. Now we have \( t_{ik} = t_k = t = k = (1, 2, \ldots, n) \). For each subject we have the same time points. The index \( i \) is dropped. Note that, we have \( t_{k+2} - t_{k+1} = t_{k+1} - t_k \) for all \( k = (1, \ldots, n) \).

In some cases, time measurements may not be exactly equal to positive natural number set, \( N^+ \). For instance, consider measurements are taken once a year starting from 1990. In such situations, we can use \( t = \text{year} - 1989 \), which satisfies our restriction \( k = t \).

Example: Faraway's Data (random sample from The Panel Study of Income Dynamics) (PSID)

PSID was a longitudinal study which is conducted by the Survey Research Center in University of Michigan. The random sample used in this thesis, which was first used by Faraway (1999), includes only three covariates among many measured.

In this example, we have a univariate response, and all covariates are time independent. Specifically,

\( Y_{it} \) is annual income for case \( i = 1, 2, \ldots, 85 \), at years \( t = 1968, 1969, \ldots, 1990 \) (23 consecutive years).
\( X^{(2)}_i \) is a 3-dimensional vector, where \( x^{(2)}_{1} \) is age in 1968, \( x^{(2)}_{2} \) is years of education, and \( x^{(2)}_{3} \) is the gender of the case.

A portion of the dataset is given in Table 2.6. Note that if we take \( t = \text{year} - 1967 \), then follow-up number is the same with \( t \). Cases are followed at equally spaced common time points.

Example: Acoustic Data (Acoustic)

This is an experiment conducted in the Department of Electronics of Ankara University, Turkey. With the purpose of observing voice disorders or changes in speech spectrum, voices of sixteen speakers are recorded for five vowels /a/, /e/, /i/, /o/, /u/ at two different time points.
Table 2.6 A portion of PSID data

<table>
<thead>
<tr>
<th>Case(i)</th>
<th>Income, Year</th>
<th>Age, Education, Gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6000, 1968</td>
<td>31, 12, 1</td>
</tr>
<tr>
<td>2</td>
<td>7500, 1968</td>
<td>29, 16, 1</td>
</tr>
<tr>
<td>3</td>
<td>NA, 1990</td>
<td></td>
</tr>
<tr>
<td>85</td>
<td>3000, 1968</td>
<td>33, 6, 1</td>
</tr>
<tr>
<td>41000, 1990</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In acoustic analysis, common measurements are pitch (average fundamental frequency), jitter (the relative evaluation of the period-to-period variability of the pitch within the analyzed segment), shimmer (the relative evaluation of the period-to-period variability of the peak-to-peak amplitude within the analyzed segment), and HNR-harmonic-to-noise ratio (average ratio of the harmonic spectral energy in the frequency range 70 to 4500 Hz to the average ratio of the inharmonic spectral energy in the frequency range 1500 to 4500 Hz).

The four components of response vectors are pitch, jitter, shimmer, and HNR measurements for speaker \(i (i = 1, ..., 16)\), at follow-up \((t = 1, 2)\).

\(X_{i}^{(2)}\): vowel
Table 2.7 A portion of acoustic data

<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(.411,156,10.307,7.504),1</td>
<td>(.335,115,12.571,7.203),2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(.486,297,10.08,7.506),1</td>
<td>(.441,148,11.817,1.17),2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>(.447,126,13.625,4.787),1</td>
<td>(.804,213,9.602,5.181),2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(.619,635,7.312,4.56),2</td>
<td>(.488,594,9.954,4.45),2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Case(i)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>(.399,179,11.503,6.039),1</td>
<td>(.396,305,10.832,6.309),2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(.615,225,10.315,4.718),1</td>
<td>(.562,335,11.241,5.317),2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Follow-up(k)

A small portion of the Acoustic data is illustrated in Table 2.7. Recordings are taken at equally spaced common time points.

Example: Longitudinal Studies of Aging (LSOA)

This was a survey study that is conducted by the National Center for Health Statistics and the National Institute on Aging. The data is collected by the U.S. Bureau of the Census. LSOA is reported to be the representative of noninstitutionalized U.S. population who are 70 years and older in 1984.

The first component of the response vector, $Y_{it1}$, is the indicator value of whether the husband in family $i$ ($i = 1, ..., 258$) used home health care at time $t$, where $t = 1984, 1985, ..., 1991$ (eight consecutive...
years). The second component, \( Y_{it2} \), corresponds to wife's response.

\[ X_{it}^{(1)} = \text{(any ADL or IADL impairment, number of ADL impairment, indicator variables of available help, whether case is near death)}, \]

where ADL stands for Activities of Daily Living and IADL stands for Instrumental Activities of Daily Living. ADL is related to personal care such as showering, dressing, and eating, whereas IADL is related to independent living, such as preparing meals, managing money, and shopping.

\[ X_{it}^{(2)} = \text{(gender, age in 1984, year of earliest observed ADL or IADL impairment, year of death, year of spouse death)} \]

Table 2.8 A portion of LSOA data

(Husband's HHC use, Wife's HHC use), (Number of ADL's for husband, Number of ADL's for wife).

<table>
<thead>
<tr>
<th>Year</th>
<th>Case(i)</th>
<th>(Age of husband in 1984, Age of wife in 1984)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>258</td>
<td>(0.0)(0.3), 1984, (0.0)(&quot;.&quot;), 1985</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>8</td>
<td></td>
</tr>
</tbody>
</table>

A small portion of the data is given in Table 2.8. In this example, each case is a couple. Although,
the subjects within each case are related, different cases are independent. The number of ADL impairments is measured only in “even” years. For other years, the corresponding cells are empty (marked as * here). Note that, we didn’t mark them as missing, since these cells are empty as a result of study design, and shouldn’t be treated as missing. We have equally spaced and common time points.

Example: Growth Data (Growth)

This data investigates the growth change for 11 girls and 16 boys. It was first introduced by Potthof and Roy (1964). The dataset and description is available in Verbeke and Molenberghs (2000) on pages 16 and 17.

\[ Y_{it} \] is the height of child \( i (i = 1, ..., 27) \) at ages 8, 10, 12, 14.

\[ X_{it}^{(1)} = \text{age}, \quad X_{i}^{(2)} = \text{gender}. \]

Data is collected at equally spaced common time points.

Example: Heights of Schoolgirls (Height)

This data, taken from Verbeke and Molenberghs (2000), page 16, questions whether the height of preadolescent girls is related to their mother’s height.

\[ Y_{it} \] is the height of girl \( i (i = 1, ..., 20) \) yearly observed between ages 6 to 10.

\[ X_{i}^{(2)} \] is the height of the mother (small, medium, tall)

We have equally spaced common time points.

Example: Mastitis in Dairy Cattle (Mastitis)

This data looks into the relationship between the milk yield and occurrence of mastitis. Mastitis is an infectious disease in cows. In the first wave of data collection all cows are expected to be free of the

\[ Y_{it} \] is milk yield in thousands of liters for cow \( i \) (\( i = 1, \ldots, 107 \)) at year \( t \) (\( t = 1, 2 \)). Measurements are taken in two consecutive years.

\[ X_{i}^{(2)} \]: mastitis occurrence in cow \( i \)

We have equally spaced common time points.

**Example: CD4+ cell numbers (CD4+)**

This dataset is discussed throughout the book of Diggle et al. (2002), and introduced on page 3. It includes 369 infected men enrolled in Multicenter AIDS Cohort Study. The purpose is studying the change in a marker for HIV. HIV attacks an immune cell called CD4+. An uninfected person has around 1100 cells per milliliter of blood, and this number decreases over time for an infected person.

\[ Y_{it} \]: CD4+ cell numbers for men \( i \) (\( i = 1, \ldots, 369 \)) at \( t \) years since seroconversion, where time since seroconversion is the time when HIV becomes detectable.

\[ X_{it}^{(1)} = (\text{smoking (packs per day)}, \text{recreational drug use (yes/no)}, \text{number of sexual partners}, \text{depressive symptoms (on CESD scale)}) \].

Although patients did not report to the center at the same time, time points are pulled together for simplicity of analysis. Therefore, there are common points that are equally spaced.

**Example: The Rat Data (Rats)**

The purpose of this data, described in Verbeke and Molenberghs (2000) page 7 and 8, is to investigate the effect of therapeutic use of hormones on growth. This is an experiment on Wistar rats, conducted in the Department of Orthodontics of the Catholic University of Leuven, Belgium.

\[ Y_{it} \]: distances (in pixels) between well-defined points on X-ray pictures of the skull of the rat \( i \).
$(i = 1, ..., 50)$ at days 50, 60, 70, 80, 90, 100, 110.

$X_{i}^{(2)}$: treatment method (control, low, high)

Observations are taken at the equally spaced time points. But, most rats died before the end of the study.

As it should be clear from the examples in this section, it is not common to see unconstrained data in the literature. We searched far and wide to find our examples for this type. This indicates a problem. We would expect that the most unstrained data would be the more common type of data collected, and we suspect it is not often seen in the literature because it is difficult to model. In our proposals for EDA and graphics we were conscious that the methods in next chapter could work with unconstrained data, because it is important to be able to digest information existing in even the most complex types of data.

2.3 Questions of interest and how to address them using exploratory methods

In this section, we will discuss some questions of scientific interest in longitudinal data. Possible approaches used in the literature and our approaches to answer these questions will be discussed. The approaches will be illustrated on datasets discussed in the previous section. The corresponding SPlus/R code used to generate plots will be supplied when available. Some functions and tools, such as histograms for categorical data (hist.factor), are not available in R. On the other hand, mosaic plots are only available in R. That is why we used both R and SPlus in the analysis. To distinguish the codes from these two software, "R >" is used for R code.

Responses:

1a.

* Question: What are the frequencies or distribution of a response at time $t$? That is, at a specific time, how many individuals possess each value of a response?
**Approach:** The distribution of responses can be explored by a histogram of response at time $t$.

**Example:** In IYFP data, how many individuals admitted feeling at least some anxiety? Figure 2.1 gives the histogram of logarithmic transformation of mean anxiety of subjects in year 1994 for the IYFP study. It is a skewed distribution with almost half of the people reporting no anxiety.

**Code:**

```r
> hist(log(iyfp[iyfp[,20]==94,13]),xlab="Log(Mean Anxiety) in 1994")
```

Figure 2.1 IYFP Data: Histogram of log-mean anxiety in 1994. Almost half of the subjects reported feeling no anxiety.

**Question:** How does response $j$ change over time?

**Approach:**

Scatterplot of response $j$ at time $t$ versus $t'$ for each $t \neq t'=1,2,...,T$, can be used. That is, we can use scatterplot matrix for equally spaced time points. Although Dawson et al. (1997) discussed
scatterplot matrix display as an effective method even for studies with unequally spaced time points, in Diggle et al. (2002, page 34) variogram is suggested to be more effective for this case.

Scatterplot matrices are hard to display when there are many time points (eg, Faraway's data with 23 time points, and Vorozole data with 24 time points). Especially in these cases, an alternative is scatterplot of time versus response j, with maybe a smoothing curve to investigate mean structure (see Figure 2.23 for an example). Connected symbol plot, that is scatterplot of response versus time with lines connected for subjects, is another alternative for addressing this question. However, scatterplots with connected lines for every subject can get crowded very easily. On the other hand, if observations for subjects are not connected, longitudinal aspect of the data is not emphasized, and time trend can be misinformative. The following plot is copied from page 2 of Diggle et al. (2002) illustrating a hypothetical data. It addresses relationship between the age and reading ability of children. Data in Figure 2.2 (a) is assumed to be from a cross-sectional study (each observation is taken from a different child), while (b) and (c) are assumed to be from longitudinal studies (two observations are taken from each child). In the first and third plots, overall trend is that as age increases reading ability decreases. Moreover, in the third plot, for each individual child, reading ability decreases as he/she gets older. However, longitudinal aspect of second plot shows an increase in ability over time within each child. Therefore, in a longitudinal study, if we don't connect the observations with lines, we will observe a plot like (a), whereas the true time trend could be the one in plot (b).

![Figure 2.2](image)

**Figure 2.2** Copied from Diggle et al. (2002, page 2). Only plot (a) is from cross sectional study. Not using connected lines might mislead the time trend within subjects.
Dubin et al. gives some suggestions to handle the overcrowded look in a scatterplot of response versus time. One can connect the records of only a randomly chosen subset of subjects or use light gray for subject specific lines and darker gray for population average. However, in the first case, selected subjects may not be representative of population and the second case is not easy to obtain with most graphic devices. Finally, they suggest calculating a robust statistic for each subject (eg. median record), and display the curves of only interesting quantiles (eg. min, 5%, 10%, 50%, 90%, 95%, max).

Interactive graphics come in handy with this problem. Even though it is still inconvenient to display connected lines for all subjects, exploring subject-specific trends is much easier and time efficient with dynamic graphics over static ones. Subject traces and multivariate time trends with interesting observations can easily be explored using linked brushing by identity numbers.

Figure 2.3 Copied from Wainer (1984, page 142). Right panel is redone from left panel. Plots should reflect the unequal time spaces.

For unequally spaced data designs, the corresponding scatterplot should show the reflection of these time spaces. Otherwise, trend could seem quite different. Wainer, 1984, cautions that the scale of the graph should not be changed in the middle of the axes. Figure 2.3 is adapted from his article (page 142). It demonstrates the misleading graphic implying linear growth trend (left panel) because of unequally spaced measurements treated as equal, and reproduced one with correct exponential trend (right panel).
Trellis display of histograms conditioned on years is also useful to see the change over time.

- **Example:** In IYFP data, does anxiety increase/decrease over time?

  Figure 2.4 illustrates a trellis display in which the lower left corner panel of the graph corresponds to response at the beginning of study. The second panel from left on the lower graph corresponds to second year, 1990. As we go from left to right, the time variable increases by one. When we are finally at the end of bottom panels, display continues on the top panel. One should follow the top panel from left to right, too. That is, again as we go from left to right, time variable increases by one. In this trellis display, one can observe that there is a noticeable difference in responses between 1994 (high school graduation year) and 1995. Almost twice as many people reported no anxiety in the later year.

- **Code:**

```r
> trellis.device(color=F)
> histogram("log(iyfporiginal[,13])|time",type="Count",xlab="Log(Mean Anxiety)")
```

- **Question:** What is the association between response $j$ and response $j'$ at time $t$? In other words, how likely is it to observe the occurrence of two responses together for a specific subject at a specific time?

- **Approach:** Scatterplot of response $j$ versus response $j'$, where $j \neq j' = 1, 2, ..., r$, for each time point separately is a good starting point. Mosaic plot is an alternative for categorical responses.

Another way to answer this question is linked brushing. Link time series plots of responses $j$ and $j'$ (Figure 2.6), and brush low and high observations in one response. Observe if responses behave the same way or opposite.
Figure 2.4 IYFP Data: Trellis display of histograms of log-mean anxiety over years. Almost twice as many people reported no anxiety in 1995 compared to 1994 (high school graduation year).

- **Example:** In IYFP data, what is the odds of a person having hostility if he/she also has depression?

Figure 2.5 illustrates the trellis display of scatterplots for IYFP data in which we can observe the positive correlation between log-mean hostility and log-mean depression scores.

Figure 2.6 illustrates the link brushing approach. Low depression scores are brushed into red, high ones are brushed into blue. One can observe that marginal distribution of anxiety and marginal distribution of hostility behaves the same way as depression does. That is, those who recorded no depression (527 observations over 8 waves) recorded lower levels of anxiety and hostility. Similarly, high values of depression records correspond to high values of other distress measurements. In Figure 2.7, depression plot is linked to the scatterplot of anxiety versus hostility to observe joint distribution of these two responses conditional on depression. Now not only the positive correlation between anxiety and hostility is revealed, but also we can observe that low (and high) scores of depression are associated with low (and high) values of other two responses.
Figure 2.5 IYFP Data: Positive correlation between two responses; hostility and depression.

- Code:

```r
> xyplot(log(jitter(iyfp[,15])) ~ log(jitter(iyfp[,17]))) | time, xlab="Log of Mean Hostility (jittered)", ylab="Log of Mean Depression (jittered)", aspect=1
```

- Question: What is the correlation structure among cases that belong to the same group for response j? That is, if subjects are connected in some way, are their responses likely to be related too?

- Approach: Our assumption in notation and parameterization part states that different cases are independent. However, a case may involve more than one unit. For example, a case can be a household, or a region in a spatial data. Exploring correlation between individuals that belong to the same case could be interesting.

Scatterplot of response for unit 1 versus response for unit 2, or linked time series plots could prove helpful. Another approach would be plotting the empirical growth curves of subjects in the same group on the same plot. Using different shapes for each person in the group will enable us to compare. If the number of subjects is large, one can randomly choose a sample to visualize. These
Figure 2.6 IYFP Data: Marginal distributions of other two responses conditional on depression. High (low) values of depression associates with high (low) values of anxiety and hostility.
Figure 2.7 IYFP Data: Joint distribution of anxiety and hostility conditional on depression. Positive correlation between anxiety and hostility, and positive correlation between (anxiety,hostility) and depression.

Plots will contain the information on responses for more than one subject, and information on time.

- **Example:** In LSOA data, what is the odds of wife using home health care given the husband is using it, too?

Figure 2.8 displays the responses by time for 20 randomly chosen families in LSOA study. Wife’s home health care is shown by an ‘o’. Within each family, given a specific year, the overlap of most data implies high correlation within household. That is if one of the members of couple used care then the other one was likely to use it, too.

**Covariates:**

2a.

- **Question:** What is the frequency of a subgroup? What is the distribution of a covariate? In simpler words, in the group of our interest, how often do we observe a subgroup or how does the characteristics of an observed variate change?
Figure 2.8 LSOA Data: Home health care use of wife (o) and husband (x) in 20 randomly chosen households. There is a high correlation between couples within each household.
• **Approach:** To address this question, one can take steps in visualizing the data depending on whether covariate is time-dependent or not.

If covariate is time-independent, the number of cases in one subgroup does not change from one time point to another. Therefore, histogram with classes as subgroups at one time point is useful. For time-varying covariates, although histogram of all observations at every time point gives an overall look at the covariate, trellis displays of histograms over years could be more informative. With longitudinal data, we have dependent observations over time. That is, observations taken from cases at second follow-up is not independent from the first one, since they are taken from the same case. Trellis displays conditioned on time slice up the data by time points and enhance exploring this dependence structure.

• **Example:** How many females are there in IYFP study? How many families did experience negative life events at a particular time point?

In Figure 2.9, time-independent covariate gender is displayed by using the numbers in only one time point. The numbers in y-axis, therefore, inform us about how many male and female there are in the study.

Figure 2.10 gives histogram (on left) and trellis displays of histograms (on right) for time-varying covariate negative life events. Although from first plot, we can observe that over all years majority of the subjects reported small number of negative life events, we cannot observe the change over time which is obvious in the later one; number of subjects who reported no negative life events increases over years, and almost symmetric shape in the first four years leaves its place to skeweness to right towards the last years.

• **Code:**

```r
> hist(iyfp[,18],breaks=c(0:28)/28,xlab="Negative Life Events")
> histogram(~iyfp[,18]|time,type="Count",xlab="Negative Life Events")
> hist.factor(iyfp[iyfp[,20] == 89,3],names=c("Male","Female"),xlab="Gender")
```
2b.

- **Question:** For time-dependent variables, is there a tendency over time? If it is possible for an observed characteristic to change from one time to another, how does it behave?

- **Approach:** One way to observe the time trend is boxplots of covariate at each time point. Boxplots are useful for displaying quantiles, variation and outlying observations in the data. Another way to observe the time trend would be trellis displays with one panel for each time point, just like Figure 2.10.

- **Example:** In IYFP data, is the number of people living in the household increase/decrease over time?

Boxplot in Figure 2.11 indicates that household size remains very similar over the years until 1994, with a median of 5 people. Starting from 1994, which is the year of graduation from high school, the houses that our subjects are living in gets less crowded on the average.

Another look at the same data with aim to answer the same question is attempted by trellis display of histogram for that variable with time conditioned. As discussed in 2a., trellis displays can be informative to recognize the time effect. However if the sole purpose is to see the time tendency, using boxplot could be better. For trellis display given in Figure 2.11, one should shift
his/her eye from one plot to another to see the time change, and comparing the location of average for each time point could be challenging. One solution to this is changing the layout of the graphic. Figure 2.11 is redone with a new layout format and illustrated in Figure 2.12 (empty display for year 1995 is ignored for space convenience). Now it is easier to see the location and scale differences in different time points. However, when the number of time points is large, it is not possible to fit this trellis display into one page, and exploring trends in different pages can be difficult. Moreover, identifying outlying observations is easier for boxplots.

- Code:

```R
> boxplot(iyfp89[,4], iyfp90[,4], iyfp91[,4], iyfp92[,4], iyfp94[,4], iyfp95[,4], iyfp97[,4], iyfp99[,4], names=c("1989", "1990", "1991", "1992", "1994", "1995", "1997", "1999"), boxwex=1, outpch=1, outline=F, boxcol=-1, medcol=1, ylab="Household size")


> histogram(~ iyfp89[,4]|time2, layout=c(1,8), type="Count", xlab="Household size", ylab="Negative Life Events")
```
Question: What is the association among covariates? Is there a multicollinearity problem? Is the occurrence of an event related to the occurrence of another? In other words, is observing one characteristic related to observing another? Is it possible and maybe even better to explain the questions at hand by observing just one of them?

Approach: When looking at the association between two covariates, one should choose different visualization techniques depending on whether covariates are time-varying and whether they are categorical. A mosaic plot is an effective method for categorical covariates, while scatterplot of one covariate against other is useful for continuous variables. For mixed data (when one covariate is continuous and other categorical) scatterplot with categorical variable jittered is one method. An alternative would be categorizing the continuous variable, though this may cause some information loss.

If at least one of the covariates is time-varying, displaying one plot for each separate time points could enable us to explore change over time. However, even for moderate amount of time points, displaying all plots can get crowded. If association is observed to be constant over time, only one plot at one time point could be displayed. Our suggestion to this challenge is calculating and displaying association measurements for each time point, and displaying only few plots that correspond to the highest and maybe lowest association.

Diggle et al. (2002, page 52) suggest the use of log-odds ratio related measures for binary data to estimate association, and the use of correlation-based approaches for continuous variables. They discuss the extension of log-odds statistic for ordinal and nominal variables. Although there is no simple built-in function in Splus or R for calculating odds ratio, the effort to write ones own function might be worthwhile.
Figure 2.11 IYFP Data: Exploring household size change over years. Distribution is almost same for the first four years, less crowded houses afterwards. Temporal change is easier to see from boxplot.
Figure 2.12  IYFP Data: Exploring household size change over years. Previous trellis display redone by using a different layout.
Figure 2.13  IYFP Data: $\chi^2$ and spearman correlation traces, and mosaic plots of cutbacks vs concerns corresponding to highest year. Horizontal axis corresponds to cutbacks (labels: 0=None, 1=less than or equal to 5, 2=more than 5), and vertical corresponds to concerns (labels: 0= None, 1=some, 2= lots).
Example: In IYFP data, are the four variables that measure economic pressure highly correlated? In other words, is there a multicolinearity problem across cutbacks, concern, material needs and ends meet variables?

Figure 2.13 illustrates the association between cutbacks and concerns in IYFP data. $\chi^2$ (top plot) and Spearman's correlations (bottom plot) for each year are calculated and displayed along with the mosaic plots for the highest correlated year.

In these mosaic plots, we are able to observe the conditional distribution of concerns given cutbacks, hence able to observe association between these two variables. The positive correlation between covariates suggests, for example, subjects who had more than 5 cutbacks are likely to have lots of concerns too.

Note that although $\chi^2$ and Spearman's correlations suggest different years to be plotted, trace plots of statistics show quite similar patterns: The first three measurements being highest, and the last three lowest. One disadvantage of using these kinds of association measurements, of course, is the underlying assumptions. For $\chi^2$ test to be valid, expected cell counts are suggested to be at least 5.

Figure 2.14 illustrates a scatterplot with the aim of exploring association between ends meet and material needs. Although both variables are ordinal, material needs have 26 categories. Therefore scatterplots with jitter is preferred over mosaic plots here. In this plot, higher levels of answers in both items are supposed to indicate more problems. Since a person who has problems with meeting the ends would also have problems with material needs, we would expect a positive correlation. The negative correlation observed in this plot is actually an indicator of data reconstruction error. When this dataset first arrived, there were 1944 columns of items. I reconstructed this data to obtain 20 meaningful columns of variables. While going through the codebook, I didn't notice that one of the ends meet item is negatively stated. After detecting this issue, negatively stated question is recoded.

Code:
R > csq<-NULL
R > for(i in c(90:92,94,95,97,99))
R > csq[i]<-chisq.test(iyfp[iyfp[,20]==i,10],iyfp[iyfp[,20]==i,11])$statistic
R > csq<-csq[c(90:92,94,95,97,99)]
R > plot(c(90:92,94,95,97,99),csq,xlab="Years",ylab="Chi-Square",cex=1.5,pch=16)
R > lines(c(90:92,94,95,97,99),csq,lty=2)
R > par(pty="s")
R > tcsq<-table(iyfp92[,10],iyfp92[,11])
R > mosaic.plot(tcsq)

R > csp<-NULL
R > for(i in c(90:92,94,95,97,99))
R > csp[i]<-cor.test(iyfp[iyfp[,20]==i,10],iyfp[iyfp[,20]==i,11],
method = "spearman",conf.level = 0.95)$estimate
R > csp<-csp[c(90:92,94,95,97,99)]
R > plot(c(90:92,94,95,97,99),csp,xlab="Years",ylab="Spearman's correlation",cex=1.5,pch=16)
R > lines(c(90:92,94,95,97,99),csp,lty=2)

2d.

- **Question:** What is the correlation structure among cases that belong to the same group such as family? Sometimes observations are taken from units that have a common characteristic, such as subjects related genetically, or locationwise. In this situations, how related are the observations taken from these units?

- **Approach:** At a given time point t, exploring the scatterplot of one person's record versus the covariate of other people with the same ID might reveal correlation structure. An alternative is linked brushing by ID—Link the scatterplots of time vs first person's variable with the scatterplot
of time vs second person’s variable. Then, brush in one plot by using brushing by ID.

- **Example:** In IYFP study, suppose records were taken from both the mother’s and father’s of the target. How correlated mother’s and father’s cutback record are?

Because of time limitations, we did not use data obtained from father’s in this thesis. Therefore, we don’t have an illustration for this question.

**Covariates and Response**

3a.

- **Question:** Does response j of subjects differ conditional on their covariates at a given time point? Given the value of a covariate, how does subject’s answer vary?

- **Approach:** One dimensional responses are easy to compare. Histograms of responses, one for each subgroup, can be plotted side by side. One should be careful about using identical scales
Figure 2.15  IYFP Data: Histograms of hostility for two gender groups. In the margin, 'H' stands for hostility. Females are more likely to report hostility compared to males.

Multivariate response data introduces complexity even for such a simple question. However, along with complexity, interesting features appear. Scatterplots (or mosaic plots) or grand tour of multiple responses can be used with different colors or shapes for each subgroup for comparison purposes. If there is overplotting problem, one plot for each subgroup can be displayed side by side. However, more interesting issue is the question of how to divide the space, which will be discussed below with an example.

- **Example**: In IYFP data, is distress level higher/lower for females compared to males?

Figure 2.15 illustrates the binary hostility scores for males and females in IYFP study. There are about 8% more females that reported feeling hostile.

To answer a similar question, suppose now we look at joint distribution of hostility and depression conditional on gender. Here, we will discuss two different plots that divide the space in the scatterplot of two continuous responses in two different ways.
First, consider dividing the scatterplot by two lines, one parallel to x-axis and other parallel to y-axis (Figure 2.16). That is, dichotomize each variable according to whether they are greater than a value or not. For instance, if hostility is equal to 1, subject shows no hostility symptom. Otherwise, subject is classified as feeling hostile. The corresponding mosaic plots are given in Figure 2.17. Note that, there is not much gender difference observed from these mosaic plots. We observe that females report more distress, both hostility and depression, which was observed from histograms of univariate response anyway. However, the joint distribution of responses do not differ much. In fact, for females probability of observing depression but no hostility is 0.1407, while this probability for males is 0.1314. For females probability of observing hostility, but no depression is 0.023, and for males it is 0.0464.
Figure 2.17  IYFP Data: Mosaic plot of Depression versus Hostility for females and males separately. Joint distributions for two gender seems similar.
Figure 2.18  IYFP Data: Scatterplot of Depression versus Hostility for females and males separately. Dividing subjects into 2 within each gender group.
Figure 2.19 IYFP Data: Histograms for comparing hostility and depression for male and female separately. Data is divided into two by responses: Hostility<Depression, and Hostility ≥ Depression. Gender differences observed. Males are more likely to report feeling hostile compared to feeling depressed. Exactly opposite is observed for females.
Figure 2.20  IYFP Data: Joint distribution of three responses (MA: log mean anxiety, MH: log mean hostility, MD: log mean depression) conditional on gender. Subjects tend to report similarly on all three responses.
Second, consider dividing the scatterplot of continuous variables by just one line, in the x=y direction (Figure 2.18). Now, we divide subjects according to whether they are more likely to feel hostile than depressed or not. Histograms are given in Figure 2.19. Depression scores of females are likely to be higher than their hostility scores, and exactly opposite is observed for males.

Note that, first partition of space attempts to answer the question of what is the joint probability of having depression and hostility conditional on gender, i.e. \( P(D = \text{Yes}, H = \text{Yes}|\text{Gender} = g) \) where \( g = \text{Female}, \text{Male} \). On the other hand, second partition looks at the probability of having a higher hostility than depression given gender, i.e. \( P(D < H|\text{Gender} = g) \). In statistical analysis, of course, both questions are of interest.

To sum up, from univariate response results, we observe that females are more distressed than males, even in the hostility measurements. However, when we look at multivariate responses, once males admit feeling distress, it is more likely to be feeling hostile rather than depressed.

These analyses invite new questions in the subject matter. Can the differences be explained by response bias by gender? Can females be more willing to show emotion, and males more likely to suppress it? For the purpose of this thesis, we will avoid drawing such conclusions, and content with advising to look at the data from different perspectives.

Finally, we explore the three dimensional plots in IYFP data. To explore the joint distribution of r-dimensional responses conditional on a covariate, we can link r-D plot of responses to conditioned covariate.

In Figure 2.20 the gender of targets is conditioned to observe the differences in distress levels. Although this does not reveal new pattern, this may not be the case for other datasets. Note that, subjects tend to report similarly on all three responses. That is, if a subject reports high score on anxiety, then he/she reports high hostility and depression as well. We will explore multivariate response trends more in the next section.
• Code:

```r
par(mfrow=c(1,2))
par(cex=1.2)
hist.factor(iyfp[,3]==2,14], xlab="Male", prob=T, ylim=c(0,1), names=c("H=No","H=Yes"))
hist.factor(iyfp[,3]==2,14], xlab="Female", prob=T, ylim=c(0,1), names=c("H=No","H=Yes"))
```

Scatterplots:

```r
par(mfrow=c(1,2))
par(pty="s")
plot(jitter(iyfp[,3]==2,15]), jitter(iyfp[,3]==2,17]),
xlab="Hostility", ylab="Depression", main="Female", xlim=c(1,5), ylim=c(1,5))
abline(h=1,v=1)
plot(jitter(iyfp[,3]==1,15]), jitter(iyfp[,3]==1,17]),
xlab="Hostility", ylab="Depression", main="Male", xlim=c(1,5), ylim=c(1,5))
abline(h=1,v=1)
```

Mosaic plots:

```r
par(mfrow=c(1,2))
par(pty="s")
tfemale<-table(iyfp[,3]==2,14], iyfp[,3]==2,16])
dimnames(tfemale)[[1]]<-c("H=N","H=Y")
dimnames(tfemale)[[2]]<-c("D=N","D=Y")
mosaic.plot(tfemale)
title("Female")
```

```r
par(pty="s")
tmale<-table(iyfp[,3]==1,14], iyfp[,3]==1,16])
dimnames(tmale)[[1]]<-c("H=N","H=Y")
```
Scatterplots:

```r
R > dimnames(tmale)[[2]] <- c("D=N", "D=Y")
R > mosaic.plot(tmale)
R > titleCMale"

R > par(mfrow=c(1,2))
R > par(pty="s")
R > plot(jitter(iyfp[iyfp[,3]==2,15]), jitter(iyfp[iyfp[,3]==2,17]),
        xlab="Hostility", ylab="Depresslon", main="Female", xlim=c(1,5), ylim=c(1,5))
R > abline(0,1)
R > plot(jitter(iyfp[iyfp[,3]==1,15]), jitter(iyfp[iyfp[,3]==1,17]),
        xlab="Hostility", ylab="Depresslon", main="Male", xlim=c(1,5), ylim=c(1,5))
R > abline(0,1)
```

Histograms:

```r
> iyfpdh1_ifelse(iyfp[,15]>=iyfp[,17],1,0)
> iyfpdh_cbind(iyfp,iyfpdh1)

> par(mfrow=c(1,2))
> par(cex=1.2)
> hist.factor(iyfpdh[iyfpdh[,3]==2,21],names=c("Host<Depr","Host->Depr"),xlab="Female",probability=T,ylim=c(0,0.7))
> par(cex=1.2)
> hist.factor(iyfpdh[iyfpdh[,3]==1,21],names=c("Host<Depr","Host->Depr"),xlab="Male",probability=T,ylim=c(0,0.7))
```

3b.

- Question: How does response j change over time conditional on covariates? Given the value of a covariate, how does subject’s answer vary from one time to the next?
Figure 2.21 PSID Data: Scatterplot with smooth curves and dot plot related to gender. Males earn more than females.

- **Approach:**

  Linked brushing is an effective tool for this question. Link the dotplot of covariate to scatterplot of time versus response $j$. Other way around to look at the problem is brushing the change in response from one year to another, and linking this to the plots of covariates.

- **Example:** In PSID data, do males earn more than females? If so, is this difference consistent over time?

  Linked brushing approach is illustrated in Figure 2.21. Here, we can observe income depending on gender. Males earn more than females with a slight decrease over time.

**3c.**

- **Question:** How does a change over time in a covariate effect the change in the response? Does response $j$ at time $t$ depend on the covariate at time $t-1$? In other words, how related are the temporal changes in covariate and response?
• **Approach:** Suppose we are interested in the effect of loss in income on distress. Scatterplot of difference between income at time $t$ and $t-1$ versus income at time $t-1$ would display the change in income. Points in the lower right corner of this plot will correspond to subjects that experienced income loss. When brushed and linked to r-dimensional plot of responses, this points would explore the effect of income loss on distress.

• **Example:** In IYFP data, how does distress level of case change with a 10% income loss from one year to another? Does distress level increase if person had negative events a year ago?

We applied this approach to answer this question. However, no interesting pattern is observed.

3d.

• **Question:** Is there any outstanding/outlier group or unexpected pattern in response $j$? Is there a relationship between this outstanding/outlier group and other responses or covariates? May this pattern be because of an error in the process of collecting or typing data? eg, all outliers come from one interviewer at one time point? If not, what is the implication of this finding, eg. interaction effect?

In simpler words, do some observations stand out compared to others? Is this because of a mistake or does it have a statistical implication? For example, do different levels of a covariate have a different effect on response?

• **Approach:** In most statistical applications, this is an important and frequently asked question. This might be of interest in data inspection and model checking steps. Exploratory analysis may point out a mistake in data collection (see below and Figure 2.22 for an example) or an effect that is not recognized by modeling (see below and Figure 2.23).

Scatterplot of response or residuals against time and linked brushing is proven to be very effective in answering this question. One can link the scatterplot of time versus response $j$ to covariates, and other responses.
Figure 2.22 Acoustic Data: All extreme shimmer responses and most borderline responses correspond to vowel /u/. Recording mistake caught by exploratory data analysis.

- **Example:** In Acoustic Data, are the extreme shimmer values related to any covariate? Why is vowel /u/ behaving different than other vowels?

In the right side of Figure 2.22, extreme shimmer cases are brushed into dark red colored big circles, and shimmer cases which have borderline values (around 0.4) are brushed into lighter red colored pluses. This scatterplot is linked to splineplot of vowels, which reveals that almost all high shimmer values comes from one vowel, /u/. Corresponding back to experts, it was noticed that this was in fact a result of recording mistake. When collecting acoustic data for vowel /u/, shimmer recordings were shifted by mistake. Although this is a small dataset, researchers did not notice the mistake until exploratory data analysis is handled.

**Example 2:** In PSID data, why do some subjects earn highest incomes although they belong to the low education group?

Figure 2.23 is a static output of an interactive graphic in GGobi. While exploring this dataset, we noticed that the extreme observations (highest and lowest income groups) are an indication of an
interaction effect between gender and education on income. This interaction effect was overlooked by modeling tools (see Chapter 3 for a more detailed discussion).

3e.

- **Question:** What is the missing value pattern? Is there a reason to believe response j is more likely to be missing for some subgroup? In other words, is it likely nonresponses to be systematic?

- **Approach:** As mentioned before, missing cases are very common in longitudinal studies. Although there is no way to prove the randomness of missing pattern from observed data, exploratory analysis is useful to reveal some interesting patterns. To explore missing pattern, we suggest the use of linked brushing. For our datasets in this thesis, we explored the dotplot or scatterplot or r-dimensional plot of response(s) with missing cases brushed and linked to other covariates.
Figure 2.24  IYFP Data: Exploring missing value pattern. Males are more likely to have missing responses compared to females. There is no missing case at baseline.

Figure 2.25  IYFP Data: Missing values in 1990 and 1999. In 1990, missing ratio for males and females are almost equal. In 1999, males are at least twice more likely to be missing.
**Example:** In IYFP data, is there a change in missing pattern over time? Also, is there a bias in gender in terms of missingness in reporting distress? In other terms, are males more likely not to report compared to females or vice versa?

Although we didn’t observe any disturbing departures from our assumptions of missing at random, Figure 2.24 demonstrates an interesting example from IYFP data. In the top very right panel, missing cases in the 3-dimensional plot of responses are brushed into red, and it is linked to spline plots of time and gender variables. These plots point out a couple of interesting features. First, there is no missing case in responses in the first year. Second, missing ratio increases over years from 1990 through 1992, and suddenly drops in 1994 (high school graduation year), and starts increasing again. Last but not least males are more likely to have missing distress records compared to females.

Figure 2.25 illustrates similar plots for specific years, 1990 (two plots on the left) and 1999 (two plots on the right). In these plots, observations corresponding to other years are deleted, and then the missing cases in responses are brushed (plot not shown). Now we can observe the variation in the pattern over years. In 1990, almost half of the missing responses correspond to females, while in 1999, less than 1/3 corresponds to females.

2.4 Limitations of existing software

Longitudinal data is usually visualized with the visualization software designed for cross sectional data. A few existing software packages for this type of data (eg, S-Plus package of Bates and Pinheiro, 1997) are designed to create static graphics. While these graphics are useful for displaying the data, the complex structure of panel data can best be explored by dynamic graphics. Theus (1995) compared the interactive graphics against static trellis displays. He discussed programming one’s own functions versus the ease of mouse click systems.

While software such as GGobi and MANET are useful interactive graphic software, they can be limited for exploring repeated measures. In this section we will discuss some of these limitations, and in the next section we will suggest software designs for this particular data type.
Figure 2.26 Hypothetical Data: Time trace brushing. Subject profiles of those with highest baseline response are brushed to observe their temporal change.

- Time trace brushing by variable: In longitudinal data, time traces of individuals can contain important information. Observing subject-specific changes, as well as population mean changes, over time is of particular interest in this area. Displaying all subject-specific lines can get crowded easily as the number of subjects increases. See 1b. in section 2.3 for the discussion. Time trace brushing by variable, specifically by the identification number of cases, can be useful to follow interesting individuals. For instance, how did subjects with extremely high or low responses behave in other time points or in the past? Figure 2.26 illustrates trace brushing for a hypothetical data. In GGobi, this can be done only by subsetting the subjects in the dataset.

- Built-in trend curves: To investigate the population mean changes over time, one commonly used method is fitting and displaying smooth curves. This is a nonparametric method, and it is a
quick attempt to eyeball the population average trend. Parametric methods, such as, smoothing the trend by ordinary least squares regression (Singer and Willett, 2003, page 28), are out of the scope of this thesis. While XGobi, an earlier version of GGobi, has built in nonparametric trend curves, in GGobi, there is not one ready to use function. Data analyst needs to fit the smooth curve in R, then include points and edges in the xml file of GGobi. Although this is not exactly a limitation, having built-in trend curves would be more practical.

- Mixed data: One of the biggest challenge of software is the integration of categorical and continuous variables into one software. Most datasets contain a mixture of categorical and continuous data, and exploring these together can be a challenge even for cross sectional data. GGobi basically handles continuous data, and MANET handles categorical data. In GGobi, we can jitter categorical variables, but not having mosaic plots and its derivatives (eg, fluctuation plot) limits our ability to explore proportions. Being able to brush and link interesting features in a continuous variable to an interactive mosaic plot with queries would be helpful. For instance, the following task requires interaction between scatterplots and mosaic plots: brushing the change in response from one year to another, and linking this to the plots of covariates (Figure 2.27). In IYFP data, we can plot the difference in depression records at time $t$ and $t-1$ versus depression at time $t-1$, and brush the cases that has lower depression records in the current year. Linking this to mosaic plots of negative events in these two years could help us determine whether change in the mood of target is related to what is going on in his/her life. Although this can be handled in MANET with some effort, a visualizing software that is strong in handling mixed data is still of need.

Now that we discussed some limitations of software, we will suggest software package designs for longitudinal data in the next section.

### 2.5 Suggestions for software design

Exploring the complex nature of longitudinal data requires good software development. Considering the existing visualization software are mostly for cross sectional data, or for static displays of repeated measure, in this section, we discuss some properties preferred in software and suggest some developments.
Figure 2.27 Hypothetical Mixed Data: Assumes continuous response and categorical covariates. Scatterplot measuring the change in response is linked to mosaic plot of covariate.
● Smooth curves and 45% banking: One of the most common questions in longitudinal data analysis is how the subject profiles change over time. Another related question is how the average profile changes over time. Since nonparametric smooth curves are frequently used for answering these questions, built-in smooth curves with 45% banking property in a software can be practical to visualize univariate responses. Cleveland (1994, page 66) discussed the importance of 45% banking in visualizing the shape of the curve correctly. He emphasizes that 45% banking maximizes the efficiency of judgments in slopes of smooth curves to perceive the rate of change of one variable versus another.

● Time trace brushing by variable: In an interactive software such as GGobi, adding the feature of time trace brushing by variable is necessary. This will enable us to observe individual time traces even in high dimensional graphics.

● Direction of time traces and visualizing data conditional on history: Even with the time trace brushing, tracking the time information can still be difficult. When time is used as one axis of a 2-dimensional plot, it is easy to see the change process as a function of time. In high dimensions, however, the direction of trace can be difficult to follow. Therefore time trace brushing feature can be enhanced by brushing the edges and points at different time points into different colors or shades, such as objects brushed into darker shade for later times.

This can be helpful for displaying conditional distributions of responses on history. Since at a given time point, the future observations are not yet taken, using this information in decision making does not make much sense. To explore the longitudinal nature of data, software should consider conditioning on the history rather than all observations. At any current software, this can be handled by deleting future time points, but this requires the deletion and addition of different observations at each time point.

● Sorting by brushed percentages: Another interesting feature in software would be able to sort the levels of a categorical variable by brushed percentage. For example, after brushing the most severe level in depression variable, we can sort the subject ID barcharts. This will enable us to
easily identify cases with highest responses, and look at their time trace plots of response and covariates. Do highly depressed subjects share the fate of having multiple negative events? This idea arose from a discussion with Dr. Martin Theus. This feature is available in Mondrian.

All in all, we should remember that no matter how advanced and well crafted software designs are, it is the user that should ask the questions of interest, that should carefully choose the approach, and recognize the patterns.

2.6 Summary of findings and discussion

- **Amount of information:** Great aspect of graphics is you can show large amount of data in a small area. When you plot the scatterplot of response against time, each single point corresponds to two table cells worth of information: response and time. This number can be enhanced by using different color or glyph to encode more information. Linked brushing is an approach, for example, that enables us to have additional variable information. When response is conditioned on a covariate by linked brushing, there are three cells worth of information: response, time and covariate.

  For multivariate $r$ response, each single point corresponds to $(r + 1)$ data points ($r$ responses and time) encoded in 2 dimensional screen/paper. While the eye of most people is not trained to observe such amount of information from tables, graphics are much easier to digest, communicate and detect patterns.

- **Dimension and partitioning of space:** Exploring data, especially multivariate data, requires looking at the data from many different perspectives. Joint distributions may include important information that cannot be simply observed from marginal plots. Moreover, two different plots, which both use joint distributions, can reveal different information. For example, in question 3a., we compared two plots, that both explores gender differences on hostility and depression, and emphasized the effect of dividing space in different directions. This brings us to the next item, which is probably one of the most important messages of this chapter.
• What is the question?: Answering different questions usually require different plot constructions. Therefore the exploration and visualization method will and should change according to the intended question in mind.

• Time points: It is not possible to think or find a longitudinal data without “time” information. Making a good use of this information is a must. In this chapter, we attempted to distinguish longitudinal datasets according to time structure, and emphasized that the analysis may differ according to this structure. For example, in question 1b., we emphasized that in the presence of irregular time points, plots should reflect this irregularity.

Moreover, slicing data for different time points can give information more than the overall look at the data. In longitudinal data, measurements taken from the same subject at different time points are dependent. Not slicing the data for different time points implies treating all observations as independent. Trellis plots are good examples of slicing data with respect to time measurements.

• Objects in scatterplots: Emphasizing time effect in scatterplots may require connecting the points by lines. Although this might clutter the plot, not using lines may be misleading (eg. Figure 2.2).

• Layout: Default options in software may need to be changed in order to serve the question of interest better. For instance, the layout of Figure 2.11 was changed to one column layout, since the temporal change of central location is better observed with the second one.

2.7 Glossary of terms

box plots: earlier known as schematic plot. useful for displaying quantiles, variation and outlying observations. Tukey (1977), Cleveland (1994, chapter 3.3, page 139).

Draftman’s display: scatterplot matrix of time points. useful for exploring association structure. preferred for data with equally spaced common time points. fails in the precedence of uncommon measurements, and in binary responses (even in categorical response with few categories, even with jittering). To remove the effects of explanatory variables, might use residuals $r_{ik} = y_{ik} - \hat{x}_{ik} \hat{\beta}$. Dawson,

**lorelogram**: useful for exploring association in categorical response. For binary data, range of correlations is constrained by the marginal means. Lorelogram uses log-odds ratio to avoid that. Heagerty and Zeger (1998).

**parallel coordinate plots**: useful for displaying association structure. fails in the precedence of uncommon measurements. Dawson, Gennings and Carter (1997). Parallel coordinate plots are also known as case-profile plots (time series plots of a specific subject) and interaction plots, and profile plots in the literature. Koschat and Swayne (1996).

**regression trees**: used both for exploring mean structure and association structure. display that explores association structure can handle multiple responses, time-dependent covariates, unequally spaced measurements, and missing values. display that explores association structure is sensitive to the assumption of parametric form for covariance function. Segal(1992, 1994).

**time series plots**: has different styles: connected symbol plot, symbol plot, connected plot, vertical line plot. Cleveland (1994, Chapter 3.8, pages 181-182).

**trellis displays**: displays relationship of certain variables conditioned on the other ones. Harder to display as the dimension increases. Since they are static plots, it is slower to explore. Becker, Cleveland and Shyu (1996).

**variogram**: used for displaying association. with irregular time points data, variogram is recommended. Diggle (1990).

### 2.8 References


3  EXPLORING THE MEAN STRUCTURE IN LONGITUDINAL DATA
WITH GRAPHICS: LINKED BRUSHING APPROACH

A paper yet to be submitted

Ozlem Ilk and Dianne Cook
Department of Statistics, Iowa State University, Ames, IA, 50011-1210
E-mail: oilk@iastate.edu

3.1  Abstract

The complex structure of longitudinal data with covariates begs for new visual methods that enable interactive exploration. In this paper we discuss using linked brushing for studying mean structure in longitudinal data. The methods are first applied to univariate data set, analyzed in Faraway (1999), called the Panel Study of Income Dynamics (PSID). We compare the results obtained by Faraway's analysis with those uncovered using graphics. Then the approach for exploring multivariate longitudinal data is described using data from the Iowa Youth and Families Project (IYFP). It is possible to reveal the unexpected, to explore the interaction between responses and covariates, and understand structure in multiple dimensions using these methods.

Key Words: Linked brushing, longitudinal data, visual methods, statistical graphics, exploratory data analysis, data mining.

3.2  Introduction

Longitudinal data can be considered to be complex. Repeated measures induce a correlation structure that includes a within-subject component, and a between-subject component. Measurements may be taken at infrequent and unequal time points. There are commonly missing values that may occur in a non-random manner. These complexities provide difficulties for modeling the data, although
graphical methods can often be used. Here we describe how to explore the mean structure (or trend) in longitudinal data in relation to the correlation structure of the different components. Direct manipulation graphics are used. The analyst displays several plots on the computer screen and directly manipulates elements of one plot which simultaneously changes corresponding elements of other plots. The relationships between multiple measured variables can be assessed.

Koschat and Swayne (1996) illustrated the use of direct manipulation for customer panel data. They applied tools such as case identification, linking multiple views and brushing on scatterplots, dot plots and clustering trees, and a new plot they called the case-profile plot (time series plot of a specific subject). Case-profile plots are also known as parallel coordinates plot, interaction plots, and profile plots in the literature. Koschat and Swayne recommended looking at different views of the same data to avoid the false judgment on structure from one possibly spurious view.

Faraway (1999) introduced what he called a graphical method for exploring the mean structure in longitudinal data. He used the Panel Study of Income Dynamics (PSID) data. His approach fits a regression model and uses graphical displays of the coefficients as a function of time. It is not truly a graphical method, but more properly called a parametric modeling approach. To argue this point, this paper begins with a truly graphical exploration of the mean structure in the data used by Faraway, and exposes features missed by his methods.

The purpose of this paper is to describe the use of direct manipulation graphics, primarily linked brushing, to "slice-and-dice" multivariate longitudinal data, to explore the inherent multiple aspects of the trend. There is no literature on this that we have found with the exception of some small treatment in Sutherland et al. (2000) where a tour is used to examine multiple responses in relation to time context. In this paper, we discuss exploring mean trend in the response(s) over time, in the presence of covariates. The first part describes the approach on the univariate longitudinal data used by Faraway, showing what he missed. The second part describes how to use the approach with multivariate longitudinal data. The software used is R (www.R-project.org) in association with GGobi (www.ggobi.org). The appendix contains R scripts and examples of the xml file structure used to describe longitudinal data.
3.3 Data notation

Longitudinal data comes in very different forms. Measurements can be taken at different and irregular time points. The most general form is illustrated in Table 3.1. In this type of data, which we call unconstrained, cases are observed at irregular uncommon time points.

Table 3.1 Unequally spaced uncommon time points.

<table>
<thead>
<tr>
<th>Case(i)</th>
<th>1</th>
<th>2</th>
<th>.</th>
<th>.</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Follow-up(k)</td>
<td>1</td>
<td>2</td>
<td>.</td>
<td>.</td>
<td>n,</td>
</tr>
</tbody>
</table>

For case i, i=1,2,...,N, responses $Y_i = matrix(Y_{i1},...,Y_{in})$ are measured at time points $t_i = vec(t_{i1},t_{i2},...,t_{in})$. In addition to responses, time-dependent covariates, $X^{(1)}_i = matrix(X^{(1)}_{i1},...,X^{(1)}_{in})$, and time-independent covariates, $X^{(2)}_i$ are observed at time points $t_{ik}, k=1,...,n_i$.

Specifically, in k th follow-up for case i, the response is an r-dimensional vector, $Y_{ik} = vec(Y_{ik1},...,Y_{ikr})$, time-dependent covariates are $p_1 \times 1$ vector, $X^{(1)}_{ik} = vec(x^{(1)}_{ik1},...,x^{(1)}_{ikp_1})$, and time-independent covariates are $p_2 \times 1$ vector, $X^{(2)}_i = vec(x^{(2)}_{i1},...,x^{(2)}_{ip_2})$.

The data is constrained when measurements are recorded at the same time points, and fully constrained when they are at regular intervals. In other words, for constrained data, we introduce the following notation. For case i, i=1,2,...,N, responses $Y_i = matrix(Y_{i1},...,Y_{in})$ are measured at unequally spaced time points $t_i = vec(t_{i1},t_{i2},...,t_{in})$. In fully constrained situation, for case i, i=1,2,...,N, responses $Y_i = matrix(Y_{i1},...,Y_{in})$ are measured at equally spaced time points $t_i = vec(1,2,...,n)$. The method we discuss in this paper is applicable to all of these data types.
3.4 Exploratory data analysis of univariate longitudinal data

We start our discussion with the simple situation, data with univariate response. The approach is illustrated on a random sample from a much broader study of the Panel Study of Income Dynamics Data, used in an article by Faraway. Exploring mean trend conditionally by covariate and comparison of linked brushing with Faraway's method follows.

3.4.1 Faraway's approach

Faraway (1999) suggested an exploratory method for finding an appropriate mean structure in longitudinal data analysis with an application on data from an income study. A random sample of 85 heads of household is included from a much broader study, The Panel Study of Income Dynamics (PSID). Along with a response (annual income between 1968 and 1990), the covariates age at onset, years of education and gender are recorded. The subjects in this random subset had at least 11 years of complete data. Income is transformed into natural logarithmic scale (i.e. log based e) since it is skewed. Out of those 85 household heads, 39 are female. Age of the subjects range between 25 and 39, and forty of the subjects are younger than 33 years old. Most subjects are either high-school or college graduates (34 and 13 subjects, respectively). For this random sample, we do not have the occupation information. However, for general PSID sample, majority of subjects are reported to be craftsmen, operatives, sales workers, farm workers or farmers, professionals and managers, or not in labor force in the previous year. Faraway’s work on this dataset did not attempt to draw conclusions about the subject matter, but to illustrate his method. We will also primarily use this data to illustrate the graphical methods.

Faraway used simple interpolation to estimate the coefficients of a regression model, and then plotted these estimates against time to construct the parametric model. Measurements are provided on a yearly time scale, allowing for simple modeling due to the equal time intervals. The model he used for individual $i$ is:

$$\log(\text{income})_{i}(t) = \beta_{0}(t) + \beta_{g}(t)\text{gender}_{i} + \beta_{e}(t)\text{education}_{i} + \beta_{a}(t)\text{age}_{i} + \epsilon_{i}(t)$$
Figure 3.1  PSID Data: Scatterplot with Lowess Smooth Curves and Barchart related to Gender. Males (blue) earn more than females, while the difference slightly decreases by time.

He concluded that females earned substantially less than males but the difference decreased with time. An approximately linear (increasing over time) effect of education was observed. There was no important age effect, but some indication of quadratic age effect was reported.

3.4.2 What did Faraway miss?

Figure 3.1 illustrates the linked brushing procedure. The barchart of gender is brushed and linked to the scatterplot of year versus log(income). That is, male and female subjects are brushed into different colors in barchart which makes us able to distinguish corresponding points in the scatterplot, and hence able to compare 1) the income changes over time for these two groups separately, and 2) the income of these two groups with each other. Lowess smoothed curves are fitted for each group to illuminate the trend. The interpretation is that males earn more than females, and the difference slightly decreases by time.

Table 3.2 gives the median values for raw incomes for some selected years. The overall median annual income increases from 4,530$ to 21,000$ in twenty years. The ratio of male income to female
Table 3.2 Median of raw incomes in PSID data over selected years.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>4530</td>
<td>6525</td>
<td>7500</td>
<td>9375</td>
<td>15850</td>
<td>21000</td>
</tr>
<tr>
<td>Male</td>
<td>5995.5</td>
<td>8701</td>
<td>12000</td>
<td>14400</td>
<td>22300</td>
<td>28750</td>
</tr>
<tr>
<td>Female</td>
<td>2450</td>
<td>2950</td>
<td>3676</td>
<td>6240</td>
<td>9800</td>
<td>15000</td>
</tr>
<tr>
<td>M/F</td>
<td>2.45</td>
<td>2.95</td>
<td>3.26</td>
<td>2.31</td>
<td>2.28</td>
<td>1.92</td>
</tr>
</tbody>
</table>

income (M/F) is approximately 2 to 3.

Figure 3.2 displays the scatterplot of year versus log(income) where the different levels of education are brushed into different colors. The year of education in the original dataset ranges between 3 and 16, but most of the subjects (47/85) have 12 or 16 years of education. Subjects with 16 years of education are considered to be college graduates, while subjects with 12 years are assumed to have high school diploma. Suppose our purpose is to compare the incomes of high school graduates with the incomes of college graduates over years (Figure 3.2). Note that this plot does not distinguish the gender of the subjects. We can conclude that subjects with higher education earn more than those with low education. Now, take a closer look. An interesting observation is that some of the lowest incomes correspond to the college graduates, and some of the highest incomes correspond to the high school graduates. The identity of some subjects is probed (5, 20, 45, 49, 59), and the corresponding linked gender dot plot is examined. The lowest incomes correspond to females (eg. subjects 5 and 45) even though they had college educations, and highest incomes with high school education correspond to males (eg. subjects 20, 49 and 59). There are more than 5 subjects with this pattern, but only these are marked avoiding a messy graphic. Therefore, the effect of education on income appears to be different according to gender suggesting a gender-education interaction. Faraway's work only considered main effects for this dataset.

Figure 3.3 illustrates the linked plots for exploring gender and education interaction effect. We used four different colors, one for each level of interaction. The size of the glyph is also increased, just to make plots more readable in white-black prints. Female subjects are brushed into pink if they are college graduates, and into red if they are high school graduates. Male with college diploma are brushed into lighter blue. Note how high the incomes of high school graduate males are. Figure 3.4 displays the smooth curves where both gender and education are conditioned. For males the curves are almost parallel over all years with a slight decrease in the difference during the last few years, which says that
Figure 3.2  PSID Data: Scatterplot with Smooth Curves and Dot plot related to Education. Subjects with higher education earn more than those with low education. The lowest incomes correspond to females (eg. subjects 5 and 45) even though they had college education, and highest incomes with high school education correspond to males (eg. subjects 20, 49 and 59).
Figure 3.3  PSID Data: Scatterplots related to interaction effects between gender and education. Female subjects are brushed into pink if they are college graduates, and into red if they are high school graduates. Males with college diploma are brushed into lighter blue. Some of the highest incomes belong to males with high school diplomas. There is little variability observed within college graduated males.
college educated males uniformly earn more than high school educated males. Females earn lower on average than males, which we had observed earlier, but the incomes of college-educated females seem to be particularly volatile.

Figure 3.5 displays smooth curves for gender and age interaction. Age is effectively categorized here, < 33, ≥ 33. One of the benefits of interactive plots is that we can discretize covariates “on-the-fly”, and explore different subgroups easily. In Figure 3.5, it appears as though there is no income difference between age groups for male subjects, a slight difference between age groups is observed for females, with younger women earning on average more.

In both plots (Figures 3.4 and 3.5), a nonuniform difference is observed in females. Note that uniform distribution is disturbed between years 1977 and 1983. Females that are older and those with high school diploma earned somewhat more than their expected income between those 7 years. If we had a landmark data, we could have drawn conclusions such as, this might be a result of political changes at the time.

In Figure 3.1, the smooth curves related to gender seem quite parallel to each other. A similar structure is observed in Figure 3.2 for education curves. They make this analyst question if the difference between the logarithm of incomes for male and female (or high school and college graduates) remain constant over time. Let us take a look at the plots of differences in smooth curves against year to answer this question better. Figure 3.6 gives the exponentiated differences in smooth curves for gender, age and education. Since the analysis in this section are based on a logarithm scale, the difference of smooth curves on each plot are indeed log-ratio of two subgroups. By exponentiating, we are able to observe the income ratio for these two groups over years. The ratio of male income to female income ranges between 2.2 and 2.9 over years. That is, males earn almost 3 times more than females in years between 1975 and 1978.

The difference plots for education and age effects show that the marginal effects of these variables are less significant. The income ratio of high school graduates to college graduates ranges from 0.5 to 0.85. The incomes for two age groups are very close to each other in terms of the ratio being close to 1 for most years. It is changing between 0.8 and 0.95.
Figure 3.4  PSID Data: Smooth curves related to interaction effects between gender and education. Female subjects are brushed into pink if they are college graduates, and into red if they are high school graduates. Males with college diploma are brushed into lighter blue. College educated males uniformly earn more than high school educated males. The incomes of females seem to be volatile.

Testing the significance of the results reported in this paper were attempted by data visualization based on permutation tests. Because of space limitation, method description and plots are not given in this paper. However, it is worth mentioning that the main effects of gender, and education, and the interaction effect of gender and education, on income are observed to be significant.

3.5 Exploratory data analysis of multivariate longitudinal data

3.5.1 About the data

The Iowa Youth and Families Project started in 1989, with the goals of understanding the effect of economic hardship and social changes on family members and of improving family life in Iowa during changes. Economic stress, such as that experienced in rural parts of Midwest during 1980's, and negative life events are expected to be related to emotional distress. For more information about this project, see Elder and Conger (2000).
Data in this paper is part of 11-year follow up of 451 families from the eight counties of north central Iowa. Measurements are taken at 8 time points: 1989, 1990, 1991, 1992, 1994, 1995, 1997, and 1999. Targets were selected to be seventh graders, with an average age of 12.7 years at the start of the project, with two married biological parents and with a sibling within four years of age. There are 215 male and 236 female subjects in the study.

The response variables, anxiety, hostility and depression, were measured by using a symptom check list. Original responses were ordinal, ranging from 1 to 5, with 1 corresponding to no problem. Due to skewness, responses are transformed into logarithm scale with base 10. Some of the symptoms for distress include nervousness or shakiness, an urge to break things, or feeling low in energy.

The main question of interest is the effect of specific covariates on the emotional distress. The covariate included in this study is gender (1 = male, 2 = female) of targets.
Figure 3.6  PSID Data: Exponentiated difference plots for main effects. Due to the logarithmic scale of response, exponentiated differences are actually ratios of responses for two subgroups. Males earn two to three times more than females. The marginal effects of education and age are less significant.
3.5.2 Exploring mean trend conditionally by covariate gender

In this section, we investigate the gender differences on distress level. That is, the main question of interest is whether distress level is higher/lower for females compared to males.

A large number of studies found that girls are more likely than boys to start feeling depression between the ages 13 to 15; and adult females are two to three times more likely to show depression symptoms compared to males (e.g. Ge et al. 1994). Ge et al. 2001, investigated the gender differences in depressive symptoms on 5 wave lengths of IYFP data, and reported that significant differences were observed between 8th grade and 12th grade subjects.

We start exploring the gender differences on each response univariately. Each response on logarithmic scale is plotted over time, and conditioned on the gender of subject (Figure 3.7). Node at edges on plots corresponds to the mean at that specific year. Females are presented with red glyph. Distress levels of both gender decreases over time. Average anxiety and hostility scores for both gender are close to each other, slightly higher for females in some years. However, females clearly report higher depression scores compared to males. Note that, observing less significant gender differences at baseline coincides with Ge et al. (2001)'s finding. Therefore, univariate response scores suggest that females in this study are more likely to be distressed than males, maybe even in the hostility measurements.

Next, we will look at bivariate response plots. Scatterplot of hostility versus depression is given in the first panel of Figure 3.8. Edges shown are the overall mean trend. Mean at first time point is highlighted with orange point. Second panel in Figure 3.8 is a close up look at the mean trend. These plots reveal many interesting features. As one would expect, depression and hostility scores have a high positive correlation; there is a large amount of individual variation (look how mean trend concentrates in overall data points in the first panel.) Moreover, hostility and depression scores jointly decrease over time. That is, on average, subjects are less likely to be feeling depressed and hostile as they grow up. However, the zig-zag pattern of the curve in early years reveals more. From second time point (1990) to the fifth (1994), there is a huge jump in depression level, while hostility scores stay almost stable. Depression in 1994, in which year students graduate from high school, is as high as it is at baseline.

Third panel in Figure 3.8 plots the mean curves for two genders separately. Male and females have similar mean trends. However, there are some distinctive features. For one thing, mean scores for
Figure 3.7 IYFP Data: Scatterplots of anxiety, hostility, and depression by time for two gender groups. Females are brushed into red, and males are into blue. Females are more likely to report distress, especially depression.
males have a much larger variability in both responses compared to females. From marginal point of view, females have much higher depression scores. While hostility reports of males are higher at the beginning, they seem to report lower than females by the end of the study. To sum up, when we look at marginal and joint distribution of responses, females seem to be more likely to report distress in most years.

Finally, we explore the three dimensional plots in IYFP data. To explore the joint distribution of 3-dimensional responses conditional on a covariate, we can link 3-D plot of responses to conditioned covariate. In Figure 3.9 the gender of targets is conditioned to observe the differences in distress levels. Two different views are captured here. Now we can observe the change in the mean trend in 3 dimensions. The mean trend is not flat, making jumps in different directions which can only be observed from high dimensional plots. For instance, in the first view, look at fifth and sixth time points for females. From 1994 (high school graduation year) to 1995, anxiety decreases dramatically while hostility and depression also decreases but slightly. For males, anxiety and depression increase just before high school graduation. From the second view, we can observe that the mood of women is changing for the first four years, and after high school graduation, distress is decreasing linearly in all three dimensions.

3.6 Conclusion and discussion

Visualizing data often provides information that is different to the information obtained by the numerical summaries, since if cleverly crafted, plots provide an easy to digest complex summary. Because of the complex nature of longitudinal data, developing exploratory tools should be an essential primary part of analysis.

The linked brushing is a simple and flexible method, and yet an excellent tool for understanding the underlying structures. It requires neither assumptions nor model fitting. It enables the investigator to detect the unexpected responses and effect interactions, briefly to understand the underlying phenomena, which is usually not the case with other tools, especially with model fitting. It helps to answer questions like 'what is happening when?' and possibly 'why?', such as when and why do high school graduates earn more than college graduates?

Note that reporting results is easier when the covariates conditioned are categorical, especially bi-
Figure 3.8  IYFP Data: Scatterplots of hostility against depression (top left), with overall mean zoomed in (top right) and for two gender groups separately (bottom). Mean at the first time point is highlighted with a larger glyph. Females are brushed into red, and males are into blue. Hostility and depression scores jointly decrease over time. There is a huge increase in depression at the fifth time point (high school graduation year). Males are less likely to report hostility and depression in most years. Larger variability in male trend over years is observed.
Figure 3.9  TYFP Data: Two views of joint distribution of three responses conditional on gender. Female are brushed into red, and male into blue. Mean at the first time point is highlighted with a larger glyph. Mean trend has jumps in all 3 dimensions. After high school graduation (i.e. after fifth year), distress decreases in all 3 responses.
nary. Main issue with continuous variables or categorical variables with large number of categories is
deciding on the number of categories to divide the variable into. If data is divided into small number of
groups, then it is possible and likely to lose some information. On the other hand, dividing into large
number of groups will lead to unrealistic judgments because of small number of people in each group.
Because of limited paper space, for the first data, we only reported the results of age and education
variables categorized into binary classes. However, investigating and demonstrating interactive plots
for other situations are not a challenge at all.

The linked brushing approach for the first data set in this paper could be effectively done with
static plots in S-plus/R/SAS since it is a simple data set. However, linked identification, which is
one of the advantages of interactive plots over static ones, was helpful in detecting possible interaction
effects. The greater benefits of linked brushing and identification arise when exploring more complex
longitudinal data such as in which multiple responses are recorded at a given time point, or in which
subjects are observed at ragged time points (unequally or uncommon spaced time points). The visual-
ization methods proposed in this paper can be applied to any of these simple or complex structures,
provided there is enough waves and data points at each time point. Linked views are also reported
to be beneficial in exploring missing values in GGobi (Swayne and Buja (1998)). Even in simple data
sets, interactive graphics accelerates the exploring process in terms of analyst's time over static graphics.

As we have already mentioned, our purpose with the first data was not to draw conclusions on
subject but to illustrate our methods. Therefore, we are not disturbed with the fact that this is a small
sample with a few covariates recorded. Some of the subgroups are even smaller, such as we have only
115 observations on 5 female college graduates. Now that we have an idea how to analyse such datasets,
we can explore larger ones.

In this paper, we illustrated how to explore the mean trends in longitudinal datasets. To display
variance, one can include ± standard deviations or (simultaneous) confidence intervals/regions around
mean. Visual significance tests based on permutation tests can also be applied (Buja, 1999).

The datasets in this paper are illustration of two types of longitudinal data we introduced before.
Time points in PSID are fully constrained, and the ones in IYFP are constrained. Applying this method
to unconstrained data as well as to time-varying covariates are still to be investigated.
XGobi, an older version of GGobi, has a built-in smooth curve estimation feature. However, XGobi codes to create smoothing curves are actually a repetition of corresponding R/S+ functions. Since this creates an unnecessary redundancy, this feature is removed in GGobi. Implementing different smooth algorithms through R-GGobi link is an area of future work. This feature will allow users to experiment and explore different smoothing methods. In the meantime, one can create edges in R/S+ and include these into the xml file to obtain nonparametric curves in GGobi. Note that, this requires updating xml file each time one changes the smoothing method. We give an illustration of constructing edges in GGobi in the Appendix.

3.7 Acknowledgements

The authors would like to thank Julian J. Faraway for making the first data available. The second data used in this paper is provided by Institute for Social and Behavioral Research, ISU Research Park, Ames. We would like to thank Dr. Becky Burzette, Ph.D. and Dr. Frederick O. Lorenz, Ph.D. for their help in data collection and clarifying some issues with data.

3.8 References


3.9 Appendix

The analysis in this paper are handled in GGobi, which is a data visualization system for interactive graphics written by Swayne, Cook and Buja. The software is freely available from:

http://www.ggobi.org

For the analysis in this paper, lowess smoothed curves and mean trends are calculated in R, and these edges are added to the xml file. R scripts and a part of xml file for IYFP data are given below.

Creating edges for IYFP data:

iyfp<-matrix(scan("C:/Program Files/ggobi/data/mydata/iyfp/iyfp.dat"),
ncol=20,byrow=T)
dimnames(iyfp)<-list(NULL,c("ID","TOS","Gender","HS","Income","PercapitaIncome",
"MN","EM","NEE","CUTBACKS","CONCERNS","CANXIETY","Manxiety","CHOSTILITY",
"Mhostility","CDEPRESSION","Mdepression","NLE(adj)","PARENT","TIME"))

# defining time range
year<-c(89,90,91,92,94,95,97,99)

# calculating anxiety smooth curve
smeananx<-NULL
for(i in 89:99)
smeananx[i]<-mean(iyfp[iyfp[,20]==i,13],na.rm=T)
smeananx<-smeananx[year]

# calculating hostility smooth curve
smeanhost<-NULL
for(i in 89:99)
smeanhost[i]<-mean(iyfp[iyfp[,20]==i,15],na.rm=T)
smeanhost<-smeanhost[year]
# calculating depression smooth curve
smeandepr<-NULL
for(i in 89:99)
smeandepr[i]<-mean(iyfp[iyfp[,20]==i,17],na.rm=T)
smeandepr<-smeandepr[year]

# calculating anxiety smooth curve for females
smeananxf<-NULL
for(i in 89:99)
smeananxf[i]<-mean(iyfp[iyfp[,3]==2&iyfp[,20]==i,13],na.rm=T)
smeananxf<-smeananxf[year]

# calculating hostility smooth curve for females
smeanhostf<-NULL
for(i in 89:99)
smeanhostf[i]<-mean(iyfp[iyfp[,3]==2&iyfp[,20]==i,15],na.rm=T)
smeanhostf<-smeanhostf[year]

# calculating depression smooth curve for males
smeandeprf<-NULL
for(i in 89:99)
smeandeprf[i]<-mean(iyfp[iyfp[,3]==2&iyfp[,20]==i,17],na.rm=T)
smeandeprf<-smeandeprf[year]

# calculating anxiety smooth curve for males
smeananxm<-NULL
for(i in 89:99)
smeananxm<-smeananxm[year]

# calculating hostility smooth curve for males
smeanhostm<-NULL
for(i in 89:99)
smeanhostm[i]<-mean(iyfp[iyfp[,3]==1&iyfp[,20]==i,15],na.rm=T)
smeanhostm<-smeanhostm[,year]

# calculating depression smooth curve for males
smeandeprm<-NULL
for(i in 89:99)
smeandeprm[i]<-mean(iyfp[iyfp[,3]==1&iyfp[,20]==i,17],na.rm=T)
smeandeprm<-smeandeprm[,year]

smoothdepr<-cbind(500,1,2,1,1,1,1,1,1,1,1,1,smeanzmx,1,smeanhost,1,smeandepr,1,1,year)

smoothdeprf<-cbind(501,1,2,1,1,1,1,1,1,1,1,1,smeananxf,1,smeanhostf,1,smeandeprf,1,1,year)

smoothdeprm<-cbind(502,1,1,1,1,1,1,1,1,1,1,1,smeananxm,1,smeanhostm,1,smeandeprm,1,1,year)

smooth<-rbind(smoothdepr,smoothdeprf,smoothdeprm)

# including smoothed curves into original data as if they are new subjects
iyfpmean<-rbind(iyfp,smooth)

# creating data
for (i in 1:dim(iyfpmean)[[1]])
cat("<record label="",iyfpmean[i,1],"id="",i,"">
iyfpmean[i,1],iyfpmean[i,2],iyfpmean[i,3],iyfpmean[i,4],iyfpmean[i,5],iyfpmean[i,6],iyfpmean[i,7],iyfpmean[i,8],iyfpmean[i,9],iyfpmean[i,10],iyfpmean[i,11],iyfpmean[i,12],iyfpmean[i,13],iyfpmean[i,14],iyfpmean[i,15],iyfpmean[i,16],iyfpmean[i,17],iyfpmean[i,18],iyfpmean[i,19],iyfpmean[i,20],"</record>
\n"
# calculating the length of each edge
# since we have 3 smooth curves, we should calculate 3 edges

dim1 <- NULL
for (k in 1:3)
dim1[k] <- dim(iyfp)[1] + (k-1)*length(year)

dim2 <- NULL
for (k in 1:3)
dim2[k] <- dim(iyfp)[1] + k*length(year) - 2

# including edges for only smooth curves

for (i in c(dim1[1]:dim2[1], dim1[2]:dim2[2], dim1[3]:dim2[3]))
cat("<record source=", i, "destination=", i+1,"/>
",
    file="C:/Program Files/ggobi/data/mydata/iyfp/iyfpmean", append=T)

A portion of XML file for IYFP data:

```
<category>
    <levels count="454">
        <level value="1"> 1 </level>
        <level value="2"> 2 </level>
        <level value="3"> 3 </level>
        ...
        <level value="460"> 460 </level>
        <level value="500"> 500 </level>
    </levels>
</category>
```
<description>
Overall mean
</description>

<level value="501"> 501 </level>
<description>
Mean for females
</description>
<level value="502"> 502 </level>
<description>
Mean for males
</description>

<levels>
</categoricalvariable>

<records count="3632" glyph="fc 2" color="7" missingValue="NA">
<record label="1" id="1">
1 0 2 6 32000 5333.333333 1 3 1 2 NA 0 1 0 1 1
1.1666666667 1 0 89 0 0 0
</record>
<record label="2" id="2">
2 0 2 5 10000 2000 2.5 3.5 1 2 NA 1 2.2 1
1.6666666667 1 2.9166666667 2 0 89 0 1 0
</record>
<record label="3" id="3">
3 0 2 4 16500 4125 1.5 3 1 2 NA 1 1.4 1 1.5 1.5
2 0 89 0 0 0
</record>
<record label="500" id="3609">
500 1 2 1 1 1 1 1 1 1 1 1.503646 1 1.6 1
1.618508 1 1 89
</record>
<record label="500" id="3610">
500 1 2 1 1 1 1 1 1 1 1 1.385613 1 1.498899 1
1.485688 1 1 90
</record>

...
<record label="502" id="3631"> 502 1 1 1 1 1 1 1 1 1 1 1 1.176842 1 1.284211 1 1.313596 1 1 97 </record>
<record label="502" id="3632"> 502 1 1 1 1 1 1 1 1 1 1 1 1.124022 1 1.206331 1 1.239292 1 1 99 </record>

</records>
</data>

<data name="IYFP edges">
<records count="21">

<record source="3609" destination="3610"> </record>
<record source="3610" destination="3611"> </record>
<record source="3630" destination="3631"> </record>
<record source="3631" destination="3632"> </record>

</records>
</data>
</ggobidata>
4 MARGINALIZED TRANSITION RANDOM EFFECTS MODELS (MTREM(p)) FOR MULTIVARIATE LONGITUDINAL BINARY DATA

A paper yet to be submitted

Ozlem Ilk\(^1\) and Michael Daniels\(^2\)
\(^1\) Department of Statistics, Iowa State University, Ames, IA 50011
\(^2\) Department of Statistics, University of Florida, Gainesville, FL 32611

4.1 Abstract

Generalized linear models with random effects and/or serial dependence are commonly used to analyze longitudinal data. However, determination and computation of marginal covariate effects can be difficult. Heagerty has proposed marginally specified logistic-normal models (1999) and marginalized transition models (2002) for longitudinal binary and categorical data in which the marginal mean is modeled explicitly in the presence of random effects and serial dependence, respectively. In this paper, we extend his work to handle multivariate longitudinal binary response data by proposing a framework consisting of a triple of regression models, which permits subject-specific inferences, while modeling the marginal mean response taking into account dependence across time via a transition structure and across responses within a subject for a given time via random effects. Markov Chain Monte Carlo Methods, specifically Gibbs sampling with Hybrid steps, are used to sample from the posterior distribution of parameters. Methods are illustrated on a real life dataset from Iowa Youth and Families Project (IYFP).

Key words: Bayesian hierarchical model, Hybrid MC.
4.2 Introduction

Longitudinal data occurs when repeated measurements from the same subject are observed over time. This kind of data introduces a within-subject correlation. When more than one response is measured, multivariate longitudinal data introduces a second correlation structure: correlation among responses. The purpose of this paper is to introduce a model for multivariate binary data that takes these two correlations into account, but also directly models marginal covariate effects.

There is a wealth of literature on modeling longitudinal data. Some recent work include Roy and Lin (2000), Verbeke and Molenberghs (2000), Pourahmadi and Daniels (2002), and Singer and Willett (2003). Models for multivariate binary data is a currently developing area with still limited literature. Some of the models proposed for binary multivariate data use random effects models and/or latent variables to exploit the multivariate structure of responses (Bandeen-Roche et al. (1997) and Legler and Ryan (1997)). However, determination and computation of marginal covariate effects with these models can be difficult.

Fitzmaurice and Laird (1993) proposed a likelihood based method that directly models the marginal mean. They used a marginal logistic model to explain mean structure and conditional log odds-ratios to model the time dependence in longitudinal binary data. However, since the log odds-ratios considered are conditioned on all observed responses, and not just on the history (previous responses at a given time point), this model does not allow straightforward modeling of correlation structure.

Heagerty (1999 and 2002) introduced marginally specified logistic-normal models and marginalized transition models (MTM) for univariate binary data. In both models, he started with a marginal logistic regression model for explaining the average response. The model specification is completed by adopting a random effects model in the logistic-normal models, and a transition model for MTM for explaining the within-subject dependence which imposes constraints on certain parameters. Parameter estimation in both papers was handled by maximum likelihood and/or estimating equations. Recently, Miglioretti and Heagerty (2003) discussed marginalized multilevel models for binary data in the presence of time-varying covariates.

In this paper, we extend Heagerty's work to handle multivariate longitudinal binary response data. Estimation in our model is handled by MCMC methods, specifically Gibbs sampling with Hybrid steps.
Hybrid MC, originally introduced in statistical physics (Duane et al., 1987), has been applied to Bayesian analysis by Neal (1993, 1994) and Gustafson (1997). This method makes use of gradient information, suppresses random walk behavior and updates the entire parameter vector simultaneously.

Section 4.3 introduces the motivating dataset from Iowa Youth and Families Project (IYFP). Proposed model is introduced in Section 4.4. In Section 4.5, methods and model checking are demonstrated on IYFP data. Discussion follows.

4.3 Motivating example: Iowa youth and families project (IYFP) data

The Iowa Youth and Families Project, started in 1989, is a longitudinal study with the goals of understanding the effect of economic hardship and social changes on family members, and help improve family life in Iowa during changes. Economic stress, such as that experienced in rural parts of Midwest during 1980's, and negative life events are expected to be related to emotional distress.

Data in this paper is part of 4-year follow up of 451 families from the eight counties of north central Iowa. Targets were selected to be seventh graders, with an average age of 12.7 years at the start of the project, with two married biological parents and with a sibling within four years of age. There are 215 male and 236 female subjects in the study.

The response variables, anxiety, hostility and depression, were measured by using a symptom checklist. Responses were dichotomized as whether they were feeling at least one of the physical symptoms of distress. Some of the symptoms for distress include nervousness or shakiness, an urge to break things, or feeling low in energy.

One of the main purposes of IYFP was to investigate the relationship between economic pressure and its effect on family relationships and on well-being of family members. Some of the questions of interest include how the economic or social problems effect the human beings. The effect of gender and age on distress is also of interest.

Conger et al. (1994) stated that farm crisis of 1980s had a long term effects on these families in terms of relationships and individual emotional status. They also concluded economic hardship causes daily
hassles and hence distress on parents, which in turn effects the well-being of their children through harsh parenting. Ge et al., (2001), investigated the gender differences in depressive symptoms on 5 waves of IYFP data, and reported that significant differences were observed between 8th grade and 12th grade subjects.

To examine the possible predictors of emotional distress, in our model, we included information from targets and their parents. Covariates from targets used were their gender, and whether they experienced any negative life events (NLE) in the last 12 months, such as having a close friend move away. Household information incorporated included whether they had any negative economic events (NEE), such as changing jobs for a worse one, and whether they needed to have cutbacks. Responses and covariates were examined over time. To explore the time and response item effects, we allow regressions to depend on time and response type. For more information about this project, see Elder and Conger (2000).

4.4 Model: Marginalized transition random effects models (MTREM)

4.4.1 Complete data model

In this section, we will discuss the complete data model for MTREM(p) for p=1 (1st order) and 2 (second order). In subsequent section, handling missing cases under missing at random (MAR) will be discussed.

General model: MTREM(p)

Multivariate binary longitudinal data introduces two kinds of correlation structure: a within-subject time dependence and multivariate response dependence. To explain these dependencies, we propose a model consisting a triple of regression models: a marginal logistic regression to explain the mean response, a transition model to explain within-subject time dependence for each response, and a random effects model for multivariate response structure at each time.

Let $Y_{itj}$ be the response at $j$th response type for the $i$th subject at time $t$, and $X_{itj}$ be the corresponding, possibly time dependent, covariates.
\[
\logit P(Y_{ij} = 1) = X_{ij} \beta \\
\logit P(Y_{ij} = 1|y_{i-1,j}, \ldots, y_{i-p,j}, X_{itj}) = \Delta_{ij} + \gamma_{ij1,y_{i-1,j}} + \ldots + \gamma_{ijp,y_{i-p,j}} \\
\logit P(y_{itj} = 1|y_{i-1,j}, \ldots, y_{i-p,j}, X_{itj}, b_{it}) = \Delta^*_{ij} + \lambda_j b_{it}
\]

where \( b_{it} \sim N(0, \sigma^2_b) \), and \( \lambda_1 = 1 \). We can write \( b_{it} = \sigma_t z_i \), where \( z_i \sim N(0,1) \). We will also let \( \gamma_{ij,m} = \alpha_{t,m} Z_{itj,m} + \ldots + \alpha_{tk,m} Z_{itj,k,m} \) for \( m = 1, \ldots, p \), where \( Z_{itj} \) are covariates. Typically, \( Z_{itj} \) is a subset of \( X_{itj} \). Note that, in the second level of the model, we can assume different correlation structures between response items by including an indicator variable in \( Z \).

We call the multilevel model given by equations (1)-(3) a marginalized transition random effects model of order \( p \), MTREM(\( p \)). Note that the model specified by (1) together with (2) for each response \( j \) is actually a marginalized transition model of order \( p \) (MTM(\( p \)) proposed by Heagerty(2002) for the univariate response case. MTREM introduces the third level to explain the multivariate response dependence at each time.

Since our model is marginally specified, but can handle subject-specific inferences at the same time, we can analyze both individual variation and variation between subgroups. Marginal regression coefficients \( \beta \) contrast the log odds of success for different values of a covariate, \( X_{itj} \), by averaging over individual variation, hence comparing subgroups. For instance, we can compare the log odds of observing depression in females versus males. On the other hand, the parameters at the second and third level of our model handle individual variation. The second level parameters, \( \gamma \), capture the serial dependence within each response, i.e. how \( y_{itj} \) depends on \( y_{i-1,j}, \ldots, y_{i-p,j} \). By letting \( \gamma \) to be a function of covariates \( \gamma_{ij} = \alpha_i Z_{itj} \), we allow serial dependence to differ by covariates. For instance, we can test whether subjects having negative life events in the previous year have a significant effect on this year’s distress. Note, we might also allow serial dependence to differ by response type by simply including dummy variables of response types in \( Z \) matrix. The random effects, \( b_{it} \), models the correlation across responses at a given time point, while \( \lambda \) allows these correlations to differ by response type. The first element of \( \lambda \) is set to be 1 for identifiability.

**Time-varying covariates**

MTREM assumes that the conditional mean of responses given the entire set of covariates is equal to the conditional mean given the covariate history. That is, at a time point \( t \), \( E(Y_{ij} | X_{itj}, q = 1, \ldots, n) = \)

This is a typical assumption, but should be exercised with caution. If this assumption does not hold, lagged values of covariates can be included into the model. A time dependent covariate is called exogenous if it is conditionally independent from past responses. Miglioretti and Heagerty (2003) discuss analysis with time-varying covariates for marginalized multilevel binary models. We will discuss more on this with our application.

First order model: MTREM(1)

The first order MTREM takes the following form:

\[
\text{logit} P(Y_{itj} = 1) = X_{itj} \beta \\
\text{logit} P(Y_{itj} = 1|y_{it-1j}, X_{itj}) = \Delta_{itj} + \gamma_{itj} y_{it-1j} \\
\text{logit} P(Y_{itj} = 1|y_{it-1j}, X_{itj}, b_{it}) = \Delta_{itj}^* + \lambda_j b_{it}
\]

where, \( b_{it} \sim N(0, \sigma_b^2) \), and \( \lambda_1 = 1 \). Let \( b_{it} = \sigma_1 z_{it} \), where \( z_{it} \sim N(0, 1) \). We will also let \( \gamma_{itj,1} = \alpha_{tt,1} + \alpha_{tt,1} Z_{itj,1} + \cdots + \alpha_{tt,k} Z_{itj,k,1} \), where \( Z_{itj} \) are covariates.

Initial state

Since there is no history data available for the initial state, we cannot use the second level of MTREM, and we need to assume a different model for \( t=1 \). Moreover, in longitudinal data, more variability is expected at baseline, and marginal means for fixed effects are often different than the rest of the time points. Therefore, we specify the model below. For the distribution of response at first time point, we assume the following:

\[
\text{logit} P(Y_{i1j} = 1) = X_{i1j} \beta^* \\
\text{logit} P(Y_{i1j} = 1|X_{i1j}, b_{i1}) = \Delta_{i1j}^* + \lambda^*_j b_{i1}
\]

where \( b_{i1} \sim N(0, \sigma^2) \), and \( \lambda^*_1 = 1 \). Let \( b_{i1} = \sigma_1 z_i \), where \( z_i \sim N(0, 1) \).

That is, the model assumed for initial state is a marginalized logistic normal model defined in Heagerty (1999).

The MTREM(1) model is diagrammed in Figure 4.1. For simplicity, we assumed bivariate response
in this diagram. The numbers in circles show which level of the model captures which correlation or mean. For example, level 2 measures the correlation between responses at two consecutive time points.

![Diagram of MTREM(1)](image)

Figure 4.1 Diagram of MTREM(1).

Constraints are needed for (1)-(3) to be a valid probability model. Specifically, \( \Delta_{itj} \) is determined given the other parameters. To obtain a value for \( \Delta_{itj} \) in p-order model, we solve the marginal constraint equation below (we suppress dependence on parameters for simplicity.)

\[
P(Y_{itj} = 1) = \sum_{y_{it-1}, \dots, y_{it-pj}} P(Y_{itj} = 1|y_{it-1j}, \dots, y_{it-pj})P(y_{it-1j}, \dots, y_{it-pj})
\]

(4)

For the special case of \( p=1 \), the marginal constraint equation is given by

\[
\frac{e^{X_{itj} \beta}}{1+e^{X_{itj} \beta}} = \sum_{y_{it-1j}=0}^{1} \frac{e^{X_{itj} \beta + X_{it-1j} \gamma_{it-1j}}}{1+e^{X_{itj} \beta + X_{it-1j} \gamma_{it-1j}}} \frac{e^{X_{it-1j} \gamma_{it-1j} \beta}}{1+e^{X_{it-1j} \gamma_{it-1j} \beta}}
\]

(4.1)
However, notice that for $t=2$, marginal distribution of previous response is obtained from initial state model, and is given by:

$$
\frac{X_{2i,j}}{1+e^{X_{2i,j}}} = \sum_{y_{1j}=0}^{1} \frac{e^{x_{1j}^*+\gamma_{12j}y_{1j}}}{1+e^{x_{1j}^*+\gamma_{12j}y_{1j}}} \frac{X_{1i,j}}{1+e^{X_{1i,j}}} \delta^*
$$

(4.2)

$\Delta_{itj}^*$ is obtained by solving the convolution equation

$$
P(Y_{itj} = 1|y_{it-1j}, \ldots, y_{it-pj}, X_{itj}) = \int P(Y_{itj} = 1|y_{it-1j}, \ldots, y_{it-pj}, X_{itj}, h_{it})dF(h_{it})
$$

(5)

Similarly, convolution equation takes different forms depending on time. For $t \geq 2$, the convolution equation for $p=1$ becomes

$$
P(Y_{itj} = 1|y_{it-1j}, \ldots, y_{it-pj}, X_{itj}) = \int P(Y_{itj} = 1|y_{it-1j}, \ldots, y_{it-pj}, X_{itj}, h_{it})dF(h_{it})
$$

(5.1)

To approximate this 1-dimensional integral, we used Gauss-Hermite quadrature.

$$
\int \frac{e^{\alpha_{ij}^*+\lambda_{ij}^*z_{ij}}}{1+e^{\alpha_{ij}^*+\lambda_{ij}^*z_{ij}}} \phi(z_i)dz_i \approx \sum_{k=1}^{20} w_k \frac{e^{\alpha_{ij}^*+\lambda_{ij}^*z_{ik}}}{1+e^{\alpha_{ij}^*+\lambda_{ij}^*z_{ik}}} \phi(z_i)dz_i
$$

(5.2)

where $(w_k, z_k)_{k=1,\ldots,20}$ are given in Abramowitz and Stegun (1972) as in Heagerty (1999).

The convolution equation for $\Delta_{itj}^*$ takes the following form:

$$
P(Y_{11j} = 1|X_{11j}) = \int P(Y_{11j} = 1|X_{11j}, h_{11})dF(h_{11})
$$

(5.3)

$$
\frac{X_{11j}}{1+e^{X_{11j}}} = \int \frac{e^{\alpha_{11j}^*+\lambda_{11j}^*z_{11j}}}{1+e^{\alpha_{11j}^*+\lambda_{11j}^*z_{11j}}} \phi(z_i)dz_i
$$

(5.4)

From these equations, we can see that $\Delta$ and $\Delta^*$ are deterministic functions of other model parameters. Specifically, we see that $\Delta_{itj}^*$ is an implicit function of $\beta^*, \lambda^*_t$, and $\sigma_t$. $\Delta_{itj}$ is a function of $\beta$, $\beta^*$ and $\gamma_{itj}$. $\Delta_{itj}$ for $t \geq 2$ is an implicit function of $\gamma_{itj}$ and $\beta$, while $\Delta_{itj}^*$ for $t \geq 2$ is an implicit function of $\gamma_{itj}$, $\lambda_j$, $\Delta_{itj}$ (hence $\beta$ and $\beta^*$) and $\sigma_t$. Both marginal constraints and convolution equations are solved by using Newton-Raphson methods. See Appendix A for details.
Prior distributions

The model introduced in section 4.4.1 is completed by specifying prior distributions on the parameters. Specifically, we specify diffuse proper priors for $\beta$, $\beta^*$ and $\alpha$; multivariate normal distributions with means 0 and large variances $\sigma_2^2 I$, $\sigma_3^2 I$ and $\sigma_4^2 I$ respectively. More informative priors are specified for $\lambda$ and $\lambda^*$, a normal distribution with mean 1 and variance 2; the priors were centered at 1 which corresponds to equal correlation among the responses at a given time. We assumed the priors for $\sigma_i^2$ to be proportional to $\frac{1}{(1+\sigma_l^2)^2}$.

Posterior sampling

We use a Gibbs sampler to sample from the posterior distribution of $(b_{il}, \beta, \beta^*, \alpha_l, \lambda_j, \lambda_j^*, \sigma_i^2)$. The full conditional distributions for $b_{il}, \beta, \beta^*$ and $\alpha_l$ are sampled using Hybrid MC. The full conditionals of $log(\sigma_i^2)$, $\lambda_j$ and $\lambda_j^*$ are sampled using Hybrid MC after random effects are integrated out to facilitate convergence. The algorithmic details are given in Appendix A.

Initial estimates for marginal regression coefficients were obtained from fitting independence models in SAS. These naive estimates ignore the correlation structures, therefore are biased, but they are good starting points. We fit independent logistic regression for fixed effects. Although starting points for variance estimates were educated guesses, we tested the estimation by running several chains with different starting values. All chains converged to the same portions of the parameter space. Convergence issues will be discussed in more detail in section 4.5.2.

Second order model: MTREM(2)

The second order model takes a similar form to first order model. However, at the second level of model, we regress response on the previous two time points. Specifically we define the second order MTREM as follows:

\[
\text{logitP}(Y_{ij} = 1) = X_{ij}\beta \\
\text{logitP}(Y_{ij} = 1|y_{it-l_j}, y_{it-2j}, X_{ij}) = \Delta_{ij} + \gamma_{itj,1}y_{it-l_j} + \gamma_{itj,2}y_{it-2j} \\
\text{logitP}(Y_{ij} = 1|y_{it-l_j}, y_{it-2j}, X_{itj}, b_{il}) = \Delta_{ij}^* + \lambda_j b_{il}
\]

where $b_{il} \sim N(0, \sigma_l^2)$, $\lambda_l = 1$, and $\gamma_{itj,m} = \alpha_{it,m}Z_{itj,m} = \alpha_{it,1}Z_{itj1,m} + \ldots + \alpha_{it,k,m}Z_{itjk,m}$, $m = 1, 2$. 

Level (1) Level (2) Level (3)
For initial states, we assume marginally specified logistic-normal model for $t = 1$, and MTREM(1) for $t = 2$. More specifically, we assume

$$\text{logit} P(Y_{1ij} = 1) = X_{1ij} \beta^*$$
$$\text{logit} P(Y_{1ij} = 1 | X_{1ij}, b_{11}) = \Delta_{1ij}^* + \lambda_1^* b_{11}$$

$$\text{logit} P(Y_{2ij} = 1) = X_{2ij} \tilde{\beta}$$
$$\text{logit} P(Y_{2ij} = 1 | y_{1ij}, X_{2ij}) = \Delta_{2ij} + \tilde{\gamma}_{2i1} y_{1ij}$$
$$\text{logit} P(Y_{2ij} = 1 | y_{1ij}, X_{2ij}, b_{12}) = \Delta_{2ij}^* + \tilde{\lambda}_j b_{12}$$

$\lambda_1^*$ and $\tilde{\lambda}_j$ are again set to be 1. Details for $p=2$ are given in Appendix B.

**4.4.2 Missing responses and/or covariates**

We use data augmentation (Tanner and Wong, 1987) to handle missing cases under an MAR assumption. For missing covariates we extend ideas in Ibrahim et al. (2002) for time-varying covariates. This algorithm assumes parametric distributions on covariates. It makes use of the sequential one-dimensional conditional distributions to allow both continuous and categorical covariates to be missing. See appendix C for details on the covariate model and for the computational algorithm for data augmentation.

In IYFP data, there are no missing cases in responses and covariates at baseline. The percentages of missing responses in 1990, 1991, 1992 are 5.9 %, 9.8 % and 10.6 % respectively. The percentage of cases with at least one year of missing response or covariate is 15%. That is, complete case analysis of this dataset would require the removal of 15% of subjects. There is both intermittent missingness and dropouts. If one of the responses is missing, the other two responses are also missing. Missing covariates are observed on some of the time-varying covariates, NIE, NEE, and cutbacks. The percentages of missing cases in NLE, NEE, and cutbacks are 7.3%, 7.9%, and 8.6% overall years.

Lorenz et al. (1997) reported that there was no difference in negative life events or distress level between dropouts and completers in IYFP data. Ge and Conger (1999) stated that there were also no
significant differences in demographic or study variables between these two groups. We also investigated the missing data mechanism by using exploratory analysis, and did not observe any serious problem. We also discussed this issue with experts in the center data was collected. They explained us that missigness is mostly related to relocations of subjects (such as due to work); and is not related to subjects being more/less distressed. Therefore, we assume MAR.

4.5 Analysis of IYFP data

In this section, MTREM(l) is illustrated on IYFP data. Estimation results and model checking are discussed.

4.5.1 Results

Descriptive and exploratory analysis of the data suggest that there is moderate amount of temporal correlation between responses, and also a moderate amount of correlation among distress measures at a given year. Specifically, lag-1 Spearman’s correlation for each response range between 0.11 and 0.42. Mean (over years) correlation between any two distress measure (e.g. anxiety and depression) was approximately 0.41. Observed Spearman’s correlations and lorelograms (log odds ratios versus lags) are given in Figures 4.2 and 4.3. We encounter the first kind of correlation information at the second level of our hierarchical model, while the second correlation structure is introduced at the third level. The impact of gender, negative events, and cutbacks on distress are introduced at the first level.

More specifically, covariates included in our model on the first level were the gender of subjects ($X_{ij2} : 1= male, 2= female$), and whether they experienced any negative life events (NLE) in the last 12 months (categorical variable with three levels; $X_{ij3} : 1$ if they didn’t have any negative events; 0 otherwise, and $X_{ij4} : 0$ if they had lots of negative events; 1 otherwise). Moreover, household information was incorporated; whether they had any negative economic events (NEE) ($X_{ij5} : 0 = no, 1= yes$), and whether they had cutbacks ($X_{ij6} = 1$ if they had no cutbacks; 0 otherwise, and $X_{ij7} = 0$ if they had more than 5 cutbacks last year; 1 otherwise). Responses and covariates were examined over four years ($1 = 1989, 2 = 1990, 3 = 1991, 4 = 1992$). We allow different intercepts by including the following indicator variables: Resp1 ($X_{ij8} = 1$ if response=anxiety; 0 o.w.), Resp2 ($X_{ij9} = 1$ if response=hostility; 0 o.w.), Timel ($X_{ij10} = 1$ if Year=1991; 0 if Year=1990,1992), Time2 ($X_{ij11} = 1$ if Year=1992; 0 if
Table 4.1 Posterior summaries for parameters in MTREM(1) and GEE estimates and standard errors for baseline and later time points

<table>
<thead>
<tr>
<th></th>
<th>MTREM(1)</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2.5%</td>
<td>50%</td>
<td>97.5%</td>
<td>posterior mean</td>
</tr>
<tr>
<td>Inter</td>
<td>1.50</td>
<td>2.00</td>
<td>2.50</td>
<td>2.00</td>
</tr>
<tr>
<td>Gender</td>
<td>-1.28</td>
<td>-0.55</td>
<td>0.10</td>
<td>-0.55</td>
</tr>
<tr>
<td>NLE1</td>
<td>-2.61</td>
<td>-0.87</td>
<td>1.08</td>
<td>-0.85</td>
</tr>
<tr>
<td>NLE2</td>
<td>-1.14</td>
<td>-0.39</td>
<td>0.30</td>
<td>-0.40</td>
</tr>
<tr>
<td>NEE</td>
<td>-0.41</td>
<td>0.34</td>
<td>1.09</td>
<td>0.35</td>
</tr>
<tr>
<td>Cuts1</td>
<td>-1.00</td>
<td>-0.36</td>
<td>0.27</td>
<td>-0.36</td>
</tr>
<tr>
<td>Cuts2</td>
<td>-0.13</td>
<td>0.34</td>
<td>0.85</td>
<td>0.35</td>
</tr>
<tr>
<td>Resp1</td>
<td>-3.39</td>
<td>-2.03</td>
<td>-0.96</td>
<td>-2.06</td>
</tr>
<tr>
<td>Resp2</td>
<td>-2.85</td>
<td>-1.55</td>
<td>-0.56</td>
<td>-1.59</td>
</tr>
<tr>
<td>NLE1*NEE</td>
<td>-1.71</td>
<td>0.50</td>
<td>2.52</td>
<td>0.46</td>
</tr>
<tr>
<td>NLE2*NEE</td>
<td>-1.29</td>
<td>-0.37</td>
<td>0.46</td>
<td>-0.39</td>
</tr>
<tr>
<td>Gender*Resp1</td>
<td>0.01</td>
<td>0.68</td>
<td>1.44</td>
<td>0.69</td>
</tr>
<tr>
<td>Gender*Resp2</td>
<td>-0.26</td>
<td>0.38</td>
<td>1.12</td>
<td>0.39</td>
</tr>
<tr>
<td>$log(\sigma_f^2)$</td>
<td>0.77</td>
<td>1.46</td>
<td>2.40</td>
<td>1.47</td>
</tr>
<tr>
<td>$\lambda_2$</td>
<td>0.66</td>
<td>1.40</td>
<td>2.99</td>
<td>1.50</td>
</tr>
<tr>
<td>$\lambda_3$</td>
<td>0.69</td>
<td>1.31</td>
<td>2.42</td>
<td>1.37</td>
</tr>
</tbody>
</table>
Table 4.1 (Continued)

<table>
<thead>
<tr>
<th></th>
<th>MTREM(1)</th>
<th>GEE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2.5%</td>
<td>50%</td>
</tr>
<tr>
<td>Inter</td>
<td>1.22</td>
<td>1.47</td>
</tr>
<tr>
<td>Gender</td>
<td>0.66</td>
<td>1.09</td>
</tr>
<tr>
<td>NLE1</td>
<td>-1.53</td>
<td>-0.92</td>
</tr>
<tr>
<td>NLE2</td>
<td>-1.80</td>
<td>-1.28</td>
</tr>
<tr>
<td>NEE</td>
<td>-0.80</td>
<td>-0.30</td>
</tr>
<tr>
<td>Cuts1</td>
<td>-0.47</td>
<td>-0.14</td>
</tr>
<tr>
<td>Cuts2</td>
<td>-0.15</td>
<td>0.12</td>
</tr>
<tr>
<td>Resp1</td>
<td>-0.30</td>
<td>0.34</td>
</tr>
<tr>
<td>Resp2</td>
<td>-0.23</td>
<td>0.47</td>
</tr>
<tr>
<td>Time1</td>
<td>-0.61</td>
<td>-0.14</td>
</tr>
<tr>
<td>Time2</td>
<td>-0.44</td>
<td>0.09</td>
</tr>
<tr>
<td>NLE1*NEE</td>
<td>-0.22</td>
<td>0.54</td>
</tr>
<tr>
<td>NLE2*NEE</td>
<td>0.08</td>
<td>0.65</td>
</tr>
<tr>
<td>Gender*Resp1</td>
<td>-1.01</td>
<td>-0.59</td>
</tr>
<tr>
<td>Gender*Resp2</td>
<td>-1.15</td>
<td>-0.71</td>
</tr>
<tr>
<td>Resp1*Time1</td>
<td>-0.48</td>
<td>-0.03</td>
</tr>
<tr>
<td>Resp1*Time2</td>
<td>-0.98</td>
<td>-0.50</td>
</tr>
<tr>
<td>Resp2*Time1</td>
<td>-0.38</td>
<td>0.04</td>
</tr>
<tr>
<td>Resp2*Time2</td>
<td>-0.89</td>
<td>-0.42</td>
</tr>
<tr>
<td>$\alpha_2$</td>
<td>0.61</td>
<td>1.00</td>
</tr>
<tr>
<td>$\alpha_3$</td>
<td>0.89</td>
<td>1.32</td>
</tr>
<tr>
<td>$\alpha_4$</td>
<td>0.74</td>
<td>1.08</td>
</tr>
<tr>
<td>$log(\sigma^2_2)$</td>
<td>0.55</td>
<td>1.16</td>
</tr>
<tr>
<td>$log(\sigma^2_3)$</td>
<td>0.13</td>
<td>0.78</td>
</tr>
<tr>
<td>$log(\sigma^2_4)$</td>
<td>0.13</td>
<td>0.78</td>
</tr>
<tr>
<td>$\lambda_2$</td>
<td>0.84</td>
<td>1.19</td>
</tr>
<tr>
<td>$\lambda_3$</td>
<td>1.09</td>
<td>1.74</td>
</tr>
</tbody>
</table>
Figure 4.2 Spearman’s correlations for the observed data. $dt$ is the lag difference. Marks at $dt=0$ are correlations for that specific response with the other two response items. For instance, in the first plot (titled ‘Anxiety’), mark $H$ stands for the correlation between anxiety and hostility.

Year=1990,1991). We also included the following interaction effects: NLE1*NEE ($X_{ij12}$), NLE2*NEE ($X_{ij13}$), gender*Resp1 ($X_{ij14}$), gender*Resp2 ($X_{ij15}$), Resp1*Time1 ($X_{ij16}$), Resp1*Time2 ($X_{ij17}$), Resp2*Time1 ($X_{ij18}$), Resp2*Time2 ($X_{ij19}$).

This parameterization assumes that covariates are independent of the history of covariates and previous responses. That is, it assumes that there is no need to include the lagged values of covariates into design matrix. We need to check how realistic this assumption is. To check the exogeneity assumption, we regressed time dependent covariates (negative life and economical events and cutbacks) on the history of those covariates and previous responses, adjusting for other covariates (gender and time). From SAS estimates, we found that none of the time varying covariates were predicted by responses. For instance, the odds ratio for cuts regressed on anxiety is 1.175 with a 95% CI between 0.841 and 1.640. All other confidence intervals for OR included 1. Hence we assumed the exogeneity assumption of covariates and did not include any lagged values of covariates.

From now on, “X model” stands for the model which we use to generate missing covariates and related parameters. “Y model” stands for the model that we use to sample from missing responses and related parameters, given the X. We first ran the “X model” to impute 1,000 sets of covariates, and randomly picked five of these. Then, we ran five separate Gibbs sampling algorithms for “Y model” each of 2,100 iterations, and discarded the first 100 in each chain as burn-in. The use of Hybrid MC facilitated fast convergence. The use of five separate Gibbs samples leads to a more reasonable and manageable algorithm in terms of computational time; and allows us to use different tuning parameters.
across different X's if necessary. We treated these 5 chains as independent, and after samples were obtained from all chains, they were combined to obtain one long chain. Only every fifth sample is retained to avoid possible autocorrelation problems. Mixing and convergence of the chain was inspected by graphical tools, such as trace plots, and quantitative tools, such as Heidelberger and Welch criteria and Geweke criteria. These inspections will be discussed more in model checking section.

Results are given in Table 4.1. 95 percent credible intervals, median, mean and posterior standard deviations from our model are displayed. We also include mean and robust sandwich standard error estimates from generalized estimating equations (GEE), obtained from SAS. Note that the mean of the parameters from our model is similar to working independent model estimates (GEE), but differences in standard errors are observed by as much as 50%. The largest differences are pointed out in the tables with a bold font. However, note that GEE estimates are based on complete case data.

The first part of this table corresponds to estimates at baseline. Only response indicators and one of the gender by response interactions are significant at 5% level. That is the corresponding 95% credible intervals excluded 0. The probability of having depression is higher compared to anxiety and hostility at baseline.

For later time points, gender, negative life events, time effects and some of the interaction effects are significant. Females and those who had negative events are more likely to be distressed. For instance, females are exp(1.09) = 2.97 times more likely to feel depressed, and exp(1.09+0.47-0.71) = 2.34 times...
more likely to feel hostile compared to males at a given time point. Note that, our findings of gender being significant after the first year of study coincide with the literature on this data (Ge et al., 2001). The probability of being distressed slightly decreases over time.

The parameters that associate the previous response with the current one, $\alpha_t$, change slightly from one year to another. The odds of observing distress in 1991 are $\exp(1.32) \approx 3.8$ times greater if subjects also experienced distress in 1990, compared to subjects that did not. Large standard deviation estimates of random effects suggest large subject to subject variation. Note that, $\lambda$ estimates assign higher probability to depression for an individual at a given time point. However, $\lambda^*$ estimates assign higher probability to hostility for an individual at baseline.

Note that, 95% interval estimates are not necessarily symmetric around mean. Bayesian inferences enables us to model the data without using asymptotic results for inference.

4.5.2 Convergence diagnostics

We used both graphical and quantitative methods for convergence checking. In the initial stages of our analysis, we used trace plots to check mixing and convergence of chain visually. Quantitative diagnostic tests in Bayesian Output Analysis (BOA) software are used to test convergence. BOA is a publicly available free software written by Brian J. Smith. All parameters passed both the Geweke convergence test and Hiedelberger and Welch convergence test.

To explore the correlation within parameters in the chain, autocorrelation plots were used. An example of these plots is given in Figure 4.4. For space limitations, we present the autocorrelation plots of only a subset of parameters. The plots do not indicate any autocorrelation problems.

To explore the correlation between parameters, scatterplots of two parameters were examined. To explore the correlation across more than two parameters, grand tour would be useful (Buja et al., 1996). We used scatterplots in GGobi to look at joint distributions of parameters. Although these scatterplots can easily be constructed in static plot environments, interactive graphics facilitates the analysis a great deal. Even when random effects estimates are ignored in scatterplots, we still have more than 30 parameters to plot against each other. With interactive graphics, looking at the scatterplot matrices
Bayesian Output Analysis

Figure 4.4 Autocorrelation plots for a subset of parameters. There is no autocorrelation problem.

of these more than 200 plots is simple.

By the use of scatterplots, we noticed that fixed effects parameter corresponding to gender is highly (negatively) correlated with the intercept term (Figure 4.5, left panel). After covariates are centered to avoid this collinearity problem, we rerun the Gibbs sampler. In this second simulation, not only the dependence between parameters was eliminated (Figure 4.5, right), but also the simulation was almost 20% more efficient computationally (in terms of CPU time) because of faster mixing.

4.5.3 Diagnostics for model fit

Posterior predictive checks (Gelman et al., 1995, Chapter 6) were used as quantitative methods for model comparison and checking.

Posterior predictive checks were used to check if our model has captured the correlation structure of responses reasonably. The idea with posterior predictive checks is that if the model fits well, then the data generated from this model and the observed data should look similar. We describe the checks
we used here below. Let \( \theta \) denote the vector of all parameters in the model, and \( y^{obs} \) denote the observed response data matrix. Let \( y^{rep} = (y^{rep,obs}, y^{rep,mis}) \) denote the replicated data matrix of \( N \times n \times r \) dimensions. For each iteration in Gibb's sampling, we obtain a realization of \( \theta \). Given \( \theta \) we draw \( y^{rep} \) from Bernoulli distribution with probabilities a function of \( \theta \). Given \( y^{rep} \), we compute spearman's correlation and log of odds ratios (LOR), which we denote as \( T \). Actually, we compute twenty four test statistics: \( T_1 \) through \( T_{18} \) measure correlation between response and past responses for each time and response type to check the correlation structure in the second level of our model; \( T_{19}, T_{20}, T_{21} \) are pairwise correlations between three items of responses to check the correlation structure in the third level of our initial model; \( T_{22}, T_{23}, T_{24} \) are pairwise correlations between three items of responses to check the correlation structure in the third level of our general model. We compare these correlations to correlations obtained from observed data \( y^{obs} \) supplemented with the current posterior predictive values for the missing, \( y^{rep,mis} \). For our dataset, observed correlations between response and previous response differ substantially across responses and time; eighteen test statistics are computed to account for this variation.

Gelman et al. (1995) stated that for most problems graphical comparisons are useful. For instance, one can draw the histogram of these test statistics, and draw a vertical line at the observed quantity. When more than one test quantity is used, scatterplots or grand tour plots can be applied. For "less blatant discrepancies", they also recommend the use of a p-value. The estimated p-value is given by
Figure 4.6 Some plots used in model checking. The distributions of log odds ratios are plotted for both observed and posterior predictive values on the same plot. They correspond to log odds ratios for anxiety between times 1 and 2; anxiety between times 1 and 3; anxiety between times 3 and 4; anxiety and depression after baseline; anxiety and hostility after baseline; hostility and depression after baseline in that order.

Table 4.2 and Figure 4.6 provides the p-values and some of the distributions of test statistics for our example, respectively. They suggest some problems in model fit in the second level of our model. It looks like this model has trouble capturing the serial correlation, especially later time points of hostility and anxiety responses. This misfit could be a result of misdefinition of Z matrix in the transition model. In the model used in this analysis, the transition parameters are same across response type. Including covariates, such as response type indicators, i.e. a separate 'transition' parameter for each response type, might improve the model fit. Another reason of misfit could be the use of lag 1 order model. Fitting MTREM(2) might improve the model fit. The third level of the model captures the correlation...
Table 4.2  p-values for posterior predictive checks in MTREM(1)

<table>
<thead>
<tr>
<th></th>
<th>COR</th>
<th>LOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>anx</td>
<td>t = 1, 2</td>
<td>0.436</td>
</tr>
<tr>
<td>anx</td>
<td>t = 2, 3</td>
<td>0.002</td>
</tr>
<tr>
<td>anx</td>
<td>t = 3, 4</td>
<td>0</td>
</tr>
<tr>
<td>anx</td>
<td>t = 1, 3</td>
<td>0.172</td>
</tr>
<tr>
<td>anx</td>
<td>t = 2, 4</td>
<td>0</td>
</tr>
<tr>
<td>anx</td>
<td>t = 1, 4</td>
<td>0.307</td>
</tr>
<tr>
<td>host</td>
<td>t = 1, 2</td>
<td>0.331</td>
</tr>
<tr>
<td>host</td>
<td>t = 2, 3</td>
<td>0</td>
</tr>
<tr>
<td>host</td>
<td>t = 3, 4</td>
<td>0</td>
</tr>
<tr>
<td>host</td>
<td>t = 1, 3</td>
<td>0.17</td>
</tr>
<tr>
<td>host</td>
<td>t = 2, 4</td>
<td>0.09</td>
</tr>
<tr>
<td>host</td>
<td>t = 1, 4</td>
<td>0.654</td>
</tr>
<tr>
<td>depr</td>
<td>t = 1, 2</td>
<td>0.328</td>
</tr>
<tr>
<td>depr</td>
<td>t = 2, 3</td>
<td>0.082</td>
</tr>
<tr>
<td>depr</td>
<td>t = 3, 4</td>
<td>0.166</td>
</tr>
<tr>
<td>depr</td>
<td>t = 1, 3</td>
<td>0.558</td>
</tr>
<tr>
<td>depr</td>
<td>t = 2, 4</td>
<td>0.053</td>
</tr>
<tr>
<td>depr</td>
<td>t = 1, 4</td>
<td>0.257</td>
</tr>
<tr>
<td>anx, host</td>
<td>t = 1</td>
<td>0.565</td>
</tr>
<tr>
<td>anx, depr</td>
<td>t = 1</td>
<td>0.579</td>
</tr>
<tr>
<td>host, depr</td>
<td>t = 1</td>
<td>0.435</td>
</tr>
<tr>
<td>anx, host</td>
<td>t &gt; 1</td>
<td>0.125</td>
</tr>
<tr>
<td>anx, depr</td>
<td>t &gt; 1</td>
<td>0.167</td>
</tr>
<tr>
<td>host, depr</td>
<td>t &gt; 1</td>
<td>0.289</td>
</tr>
</tbody>
</table>
structure of data. We expect an improvement on third level of general model by the improvement on the second level of model.

4.5.4 An attempt to improve model fit

We attempt to improve the model fit by including a separate transition parameter for each response type. That is, now we have Resp1 and Resp2 indicators (defined as above) included on the second level of our model.

Posterior quantities under this model are given in Table 4.3. Since GEE estimates are same with the ones in Table 4.1, they are not included here. Note that, marginal covariate effects ($\beta$ and $\beta^*$) are close to the estimates in Table 4.1. That is, first level parameters are insensitive to the change in the second level parameters. Heagerty (2002) proved that $\beta$ and $\alpha$ are orthogonal in MTM(1), and hence marginal parameter estimates are consistent regardless of possible misspecification of second level model. Since first two levels of MTREM(1) are MTM(1)’s, we expect similar robustness.

Parameter interpretations are same as the previous model except for $\alpha$. $\alpha_{tk}$, $k = 1, 2, 3$ are transition parameters at time $t$. $k = 1$ parameters correspond to regressing response on the previous one. $k = 2$ and $k = 3$ allows the regression to differ by Resp1 and Resp2, respectively. Note that, none of the newly introduced parameters are significant. Posterior predictive checks applied to this model showed no improvement in the model fit (Table 4.4).

Adjusting for the previous response, we calculated the observed spearman’s correlations between $y_{ttj}$ and $y_{t-2j}$. This correlation is 0.2 if $y_{t-1j}$ is 0; and it is 0.119 if $y_{t-1j}$ is 1. Since there is some indication that current response is correlated to the response of two years ago, we attempt to fit MTREM(2) model. However, sampling from MTREM(2) is still ongoing at the time this thesis was written.
Table 4.3  Posterior summaries for parameters in updated MTREM(1) for baseline and later time points

<table>
<thead>
<tr>
<th>Baseline</th>
<th>2.5%</th>
<th>50%</th>
<th>97.5%</th>
<th>posterior mean</th>
<th>posterior SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inter</td>
<td>1.55</td>
<td>2.04</td>
<td>2.64</td>
<td>2.05</td>
<td>0.29</td>
</tr>
<tr>
<td>Gender</td>
<td>-1.47</td>
<td>-0.6</td>
<td>0.12</td>
<td>-0.62</td>
<td>0.39</td>
</tr>
<tr>
<td>NLE1</td>
<td>-2.98</td>
<td>-1.27</td>
<td>0.92</td>
<td>-1.22</td>
<td>0.99</td>
</tr>
<tr>
<td>NLE2</td>
<td>-1.13</td>
<td>-0.43</td>
<td>0.29</td>
<td>-0.43</td>
<td>0.36</td>
</tr>
<tr>
<td>NEE</td>
<td>-0.39</td>
<td>0.36</td>
<td>1.11</td>
<td>0.35</td>
<td>0.39</td>
</tr>
<tr>
<td>Cuts1</td>
<td>-0.97</td>
<td>-0.34</td>
<td>0.25</td>
<td>-0.35</td>
<td>0.31</td>
</tr>
<tr>
<td>Cuts2</td>
<td>-0.15</td>
<td>0.38</td>
<td>0.95</td>
<td>0.38</td>
<td>0.28</td>
</tr>
<tr>
<td>Resp1</td>
<td>-3.74</td>
<td>-2.17</td>
<td>-0.98</td>
<td>-2.21</td>
<td>0.69</td>
</tr>
<tr>
<td>Resp2</td>
<td>-3.16</td>
<td>-1.71</td>
<td>-0.49</td>
<td>-1.74</td>
<td>0.67</td>
</tr>
<tr>
<td>NLE1*NEE</td>
<td>-1.63</td>
<td>0.93</td>
<td>3.07</td>
<td>0.85</td>
<td>1.15</td>
</tr>
<tr>
<td>NLE2*NEE</td>
<td>-1.24</td>
<td>-0.34</td>
<td>0.54</td>
<td>-0.33</td>
<td>0.46</td>
</tr>
<tr>
<td>Gender*Resp1</td>
<td>0.05</td>
<td>0.74</td>
<td>1.64</td>
<td>0.76</td>
<td>0.40</td>
</tr>
<tr>
<td>Gender*Resp2</td>
<td>-0.30</td>
<td>0.44</td>
<td>1.25</td>
<td>0.46</td>
<td>0.39</td>
</tr>
<tr>
<td>log(σ^2)</td>
<td>0.74</td>
<td>1.43</td>
<td>2.35</td>
<td>1.45</td>
<td>0.40</td>
</tr>
<tr>
<td>λ^2</td>
<td>0.70</td>
<td>1.43</td>
<td>3.10</td>
<td>1.53</td>
<td>0.61</td>
</tr>
<tr>
<td>λ^3</td>
<td>0.67</td>
<td>1.31</td>
<td>2.51</td>
<td>1.39</td>
<td>0.49</td>
</tr>
</tbody>
</table>
Table 4.3 (Continued)

<table>
<thead>
<tr>
<th></th>
<th>2.5%</th>
<th>50%</th>
<th>97.5%</th>
<th>posterior mean</th>
<th>posterior SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inter</td>
<td>1.10</td>
<td>1.45</td>
<td>1.77</td>
<td>1.44</td>
<td>0.17</td>
</tr>
<tr>
<td>Gender</td>
<td>0.64</td>
<td>1.06</td>
<td>1.53</td>
<td>1.06</td>
<td>0.23</td>
</tr>
<tr>
<td>NLE1</td>
<td>-1.53</td>
<td>-0.88</td>
<td>-0.29</td>
<td>-0.89</td>
<td>0.32</td>
</tr>
<tr>
<td>NLE2</td>
<td>-1.77</td>
<td>-1.27</td>
<td>-0.78</td>
<td>-1.27</td>
<td>0.25</td>
</tr>
<tr>
<td>NEE</td>
<td>-0.83</td>
<td>-0.30</td>
<td>0.19</td>
<td>-0.31</td>
<td>0.26</td>
</tr>
<tr>
<td>Cuts1</td>
<td>-0.43</td>
<td>-0.13</td>
<td>0.18</td>
<td>-0.13</td>
<td>0.16</td>
</tr>
<tr>
<td>Cuts2</td>
<td>-0.17</td>
<td>0.11</td>
<td>0.38</td>
<td>0.11</td>
<td>0.14</td>
</tr>
<tr>
<td>Resp1</td>
<td>-0.36</td>
<td>0.32</td>
<td>0.94</td>
<td>0.32</td>
<td>0.33</td>
</tr>
<tr>
<td>Resp2</td>
<td>-0.20</td>
<td>0.42</td>
<td>1.05</td>
<td>0.43</td>
<td>0.32</td>
</tr>
<tr>
<td>Time1</td>
<td>-0.64</td>
<td>-0.14</td>
<td>0.31</td>
<td>-0.15</td>
<td>0.25</td>
</tr>
<tr>
<td>Time2</td>
<td>-0.49</td>
<td>0.06</td>
<td>0.58</td>
<td>0.06</td>
<td>0.27</td>
</tr>
<tr>
<td>NLE1*NEE</td>
<td>-0.25</td>
<td>0.52</td>
<td>1.34</td>
<td>0.52</td>
<td>0.41</td>
</tr>
<tr>
<td>NLE2*NEE</td>
<td>0.14</td>
<td>0.66</td>
<td>1.21</td>
<td>0.67</td>
<td>0.28</td>
</tr>
<tr>
<td>Gender*Resp1</td>
<td>-1.04</td>
<td>-0.57</td>
<td>-0.17</td>
<td>-0.57</td>
<td>0.22</td>
</tr>
<tr>
<td>Gender*Resp2</td>
<td>-1.11</td>
<td>-0.69</td>
<td>-0.30</td>
<td>-0.69</td>
<td>0.21</td>
</tr>
<tr>
<td>Resp1*Time1</td>
<td>-0.47</td>
<td>-0.02</td>
<td>0.40</td>
<td>-0.03</td>
<td>0.21</td>
</tr>
<tr>
<td>Resp1*Time2</td>
<td>-0.95</td>
<td>-0.46</td>
<td>0.00</td>
<td>-0.47</td>
<td>0.24</td>
</tr>
<tr>
<td>Resp2*Time1</td>
<td>-0.38</td>
<td>0.03</td>
<td>0.43</td>
<td>0.03</td>
<td>0.21</td>
</tr>
<tr>
<td>Resp2*Time2</td>
<td>-0.90</td>
<td>-0.40</td>
<td>0.08</td>
<td>-0.40</td>
<td>0.24</td>
</tr>
<tr>
<td>$a_{21}$</td>
<td>0.53</td>
<td>0.97</td>
<td>1.44</td>
<td>0.97</td>
<td>0.23</td>
</tr>
<tr>
<td>$a_{31}$</td>
<td>0.85</td>
<td>1.28</td>
<td>1.71</td>
<td>1.28</td>
<td>0.22</td>
</tr>
<tr>
<td>$a_{41}$</td>
<td>0.65</td>
<td>1.04</td>
<td>1.46</td>
<td>1.04</td>
<td>0.21</td>
</tr>
<tr>
<td>$a_{22}$</td>
<td>-1.00</td>
<td>-0.14</td>
<td>0.79</td>
<td>-0.13</td>
<td>0.45</td>
</tr>
<tr>
<td>$a_{32}$</td>
<td>-0.81</td>
<td>0.02</td>
<td>0.84</td>
<td>0.03</td>
<td>0.41</td>
</tr>
<tr>
<td>$a_{42}$</td>
<td>-0.35</td>
<td>0.47</td>
<td>1.28</td>
<td>0.48</td>
<td>0.42</td>
</tr>
<tr>
<td>$a_{23}$</td>
<td>-0.50</td>
<td>0.33</td>
<td>1.28</td>
<td>0.35</td>
<td>0.46</td>
</tr>
<tr>
<td>$a_{33}$</td>
<td>-0.15</td>
<td>0.67</td>
<td>1.52</td>
<td>0.67</td>
<td>0.42</td>
</tr>
<tr>
<td>$a_{43}$</td>
<td>-0.67</td>
<td>0.10</td>
<td>0.86</td>
<td>0.10</td>
<td>0.39</td>
</tr>
<tr>
<td>$log(\sigma^2_2)$</td>
<td>0.53</td>
<td>1.17</td>
<td>1.82</td>
<td>1.18</td>
<td>0.33</td>
</tr>
<tr>
<td>$log(\sigma^2_3)$</td>
<td>0.19</td>
<td>0.81</td>
<td>1.50</td>
<td>0.81</td>
<td>0.34</td>
</tr>
<tr>
<td>$log(\sigma^2_4)$</td>
<td>0.22</td>
<td>0.82</td>
<td>1.48</td>
<td>0.83</td>
<td>0.32</td>
</tr>
<tr>
<td>$\lambda_2$</td>
<td>0.80</td>
<td>1.18</td>
<td>1.85</td>
<td>1.21</td>
<td>0.27</td>
</tr>
<tr>
<td>$\lambda_3$</td>
<td>1.05</td>
<td>1.74</td>
<td>3.00</td>
<td>1.82</td>
<td>0.51</td>
</tr>
</tbody>
</table>
Table 4.4  p-values for posterior predictive checks in MTREM(1) (repeated here for easy comparison) and in updated MTREM(1)

<table>
<thead>
<tr>
<th>MTREM(1)</th>
<th>Updated MTREM(1)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>COR</td>
</tr>
<tr>
<td>anx</td>
<td>t = 1,2</td>
</tr>
<tr>
<td>anx</td>
<td>t = 2,3</td>
</tr>
<tr>
<td>anx</td>
<td>t = 3,4</td>
</tr>
<tr>
<td>anx</td>
<td>t = 1,3</td>
</tr>
<tr>
<td>anx</td>
<td>t = 2,4</td>
</tr>
<tr>
<td>anx</td>
<td>t = 1,4</td>
</tr>
<tr>
<td>host</td>
<td>t = 1,2</td>
</tr>
<tr>
<td>host</td>
<td>t = 2,3</td>
</tr>
<tr>
<td>host</td>
<td>t = 3,4</td>
</tr>
<tr>
<td>host</td>
<td>t = 1,3</td>
</tr>
<tr>
<td>host</td>
<td>t = 2,4</td>
</tr>
<tr>
<td>host</td>
<td>t = 1,4</td>
</tr>
<tr>
<td>depr</td>
<td>t = 1,2</td>
</tr>
<tr>
<td>depr</td>
<td>t = 2,3</td>
</tr>
<tr>
<td>depr</td>
<td>t = 3,4</td>
</tr>
<tr>
<td>depr</td>
<td>t = 1,3</td>
</tr>
<tr>
<td>depr</td>
<td>t = 2,4</td>
</tr>
<tr>
<td>depr</td>
<td>t = 1,4</td>
</tr>
<tr>
<td>anx, host</td>
<td>t = 1</td>
</tr>
<tr>
<td>anx, depr</td>
<td>t = 1</td>
</tr>
<tr>
<td>host, depr</td>
<td>t = 1</td>
</tr>
<tr>
<td>anx, host</td>
<td>t &gt; 1</td>
</tr>
<tr>
<td>anx, depr</td>
<td>t &gt; 1</td>
</tr>
<tr>
<td>host, depr</td>
<td>t &gt; 1</td>
</tr>
</tbody>
</table>
4.6 Conclusion and discussion

The model proposed in this paper allows for both marginal covariate effects and individual variation for multivariate binary response data. Our model can handle both time dependent and independent covariates, and introduces heterogeneous random effects variance components.

MTREM(p) captures the marginal covariate effects via a marginal logistic regression and temporal dependence via a transition model as in MTM(p) by Heagerty (2002). However, it has an additional level with random effects to capture the correlation across responses at a given time point. That is, it is an extension of MTM to multivariate longitudinal setting.

Missing values in responses are handled by data augmentation. For missing cases in covariates we proposed a parametric approach that accounts for the time-varying structure of covariates. This approach allows researcher to 'impute' both categorical and continuous covariates.

Model checking was assessed by posterior predictive checks. Posterior predictive checks highlighted that regressing responses on the immediate previous response (i.e. model with p=1) was not adequately capturing the temporal correlation structure of our data. The extension to p=2 is proposed as a remedy.

In Iowa Youth and Families Project, we concluded that, there is a significant effect of gender, age, and negative events (both social and economical) on distress. Females and subjects who experienced negative events are more likely to feel distressed. The distress decreased over time. However, subjects who felt distress in the previous year were almost 4 times more likely to feel distress in the current year as well.

Priors for most parameters in the model are assumed to be diffuse noninformative priors. More informative priors are assumed for the correlation parameters on the third level. Sensitivity analysis to these priors is still to be investigated. Sensitivity in imputation to the order of covariates entered into the “X model” is also left as a future work. We will also investigate the propriety of the posterior when using improper priors on variances and regression coefficients.

Calculations and analyses in this paper are based on special cases of p=1 and p=2 models. Extension to higher orders is possible. Another interesting extension would be introducing temporal dependence
in the random effects and/or cross-temporal dependence between responses.

4.7 Acknowledgements

The data used in this paper is provided by Institute for Social and Behavioral Research, ISU Research Park, Ames. We would like to thank Dr. Becky Burzette, Ph.D. and Dr. Frederick O. Lorenz, Ph.D. for their help in data collection and clarifying some issues with data.

4.8 References


4.9 Appendix A: Details on sampling from posterior distribution of parameters in MTREM(l)

In Appendix A, we provide the computational algorithm and details for sampling from the posterior distribution in MTREM(l). We first lay out the algorithm. The details on how to determine $\Delta$ and $\Delta^*$ as functions of other model parameters are discussed next. Full conditional distributions used in Gibbs sampling are given. Hybrid MC requires the derivatives of loglikelihoods. We conclude Appendix A with these derivatives.

Computational Algorithm for \( p=1 \)

Let $\theta$ be the vector of all parameters in the model, i.e. \((b_1, \beta, \beta^*, \alpha_1, \lambda_j, \lambda_j^*, \sigma_t)\). Also, let $\theta_{-\delta}$ be the vector of all parameters except $\delta$.

The following algorithm is used to obtain a sample from the posterior distribution of $\theta$.

1. Assign starting values to $\theta$.
2. Calculate $\Delta_{itj}$ from marginal constraint equation by using Newton-Raphson method.
3. Calculate $\Delta^*_{itj}$ from convolution equations by using Newton-Raphson and Gauss-Hermite Quadrature.
4. Sample from the full conditional for \((\log(\sigma_t^2), \lambda_j, b_{it})\) using the following two steps:
   a. Integrate out $b_{it}$ by Gauss-Hermite Quadrature (1-dimensional integration), and sample \((\log(\sigma_t^2), \lambda_j | \theta_{-b}, \theta_{-\sigma}, \theta_{-\lambda}, Y)\) by using Hybrid MC. Update $\Delta^*_{itj}$ for $t \geq 2$ (Since $\Delta^*_{itj}$, $t \geq 2$, are functions of $\sigma_t$ and $\lambda_j$, one needs to update them each time $\sigma_t$ and $\lambda_j$ changes).
   b. Sample from the full conditional of $b_{it}$ by using Hybrid MC.
5. Sample from the full conditional for \((\log(\sigma_t^2), \lambda_j^*, b_{i1})\) using the following two steps:
   a. Integrate out $b_{i1}$ by Gauss-Hermite Quadrature (1-dimensional integration), and sample \((\log(\sigma_t^2), \lambda_j^* | \theta_{-b}, \theta_{-\sigma}, \theta_{-\lambda}, Y)\) by using Hybrid MC. Update $\Delta^*_{ij}$.
   b. Sample from the full conditional of $b_{i1}$ by using Hybrid MC.
6. Sample from the full conditional of $\beta$ by using Hybrid MC. Update $\Delta_{itj}$ and $\Delta^*_{itj}$.
7. Sample from the full conditional of $\beta^*$ by using Hybrid MC. Update $\Delta_{i2j}$, $\Delta^*_{ij}$, $\Delta^*_{i2j}$.
8. Sample from the full conditional of $\alpha_t$ by using Hybrid MC. Update $\Delta_{itj}$ and $\Delta^*_{itj}$ for $t \geq 2$.
9. Repeat steps 4 through 8.
Steps 4-8 constitutes the Gibbs sampling algorithm to sample from posterior distribution of \( \theta \).

**Calculation of \( \Delta_{itj}, \Delta_{itj}^* \) and their derivatives:**

\( \Delta_{itj} \) and \( \Delta_{itj}^* \) are determined through equations (4) and (5) by using Newton-Raphson. Numerical solution of the integrals in (5.1) and (5.4) are obtained by a 20-point Gauss-Hermite Quadrature.

- For the special case of \( p=1 \), we have two marginal constraint equations; one for \( t = 2 \), and other for \( t > 2 \). Specifically, for \( t = 2 \), we have

\[
\frac{\partial f(\Delta_{itj})}{\partial \Delta_{itj}} = -\frac{X_{itj}^\beta}{1 + e^{X_{itj}^\beta}} + \sum_{y_{itj}=0}^{1} \frac{e^{\Delta_{itj} + \gamma_{itj} y_{itj}} X_{itj}^\beta}{1 + e^{X_{itj}^\beta}} = 0
\]

For later time points, \( P(Y_{it-1j}) \) depends on \( \beta \) and not on \( \beta^* \). Hence, we have,

\[
\frac{\partial f(\Delta_{itj})}{\partial \Delta_{itj}} = -\frac{X_{itj}^\beta}{1 + e^{X_{itj}^\beta}} + \sum_{y_{it-1j}=0}^{1} \frac{e^{\Delta_{itj} + \gamma_{itj} y_{it-1j}} X_{itj}^\beta}{1 + e^{X_{itj}^\beta}} = 0
\]

To use Newton-Raphson, we require the derivatives of these equations. For \( t = 2 \), this derivative is:

\[
\frac{\partial f(\Delta_{itj})}{\partial \Delta_{itj}} = \sum_{y_{itj}=0}^{1} \frac{e^{\Delta_{itj} + \gamma_{itj} y_{itj}} X_{itj}^\beta}{(1 + e^{X_{itj}^\beta})^2} \frac{e^{\delta_{itj}} X_{itj}^\beta}{1 + e^{X_{itj}^\beta}}
\]

For later time points, this derivative takes the following form:

\[
\frac{\partial f(\Delta_{itj})}{\partial \Delta_{itj}} = \sum_{y_{it-1j}=0}^{1} \frac{e^{\Delta_{itj} + \gamma_{itj} y_{it-1j}} X_{itj}^\beta}{(1 + e^{X_{itj}^\beta})^2} \frac{e^{\delta_{itj}} X_{itj}^\beta}{1 + e^{X_{itj}^\beta}}
\]

- From equation (5.4), we have the convolution equation for \( \Delta_{itj}^* \):

\[
f(\Delta_{itj}^*) = -\frac{X_{itj}^\beta}{1 + e^{X_{itj}^\beta}} + \int \frac{e^{\Delta_{itj} + \gamma_{itj} z_i} X_{itj}^\beta}{1 + e^{X_{itj}^\beta}} \phi(z_i) dz_i = 0
\]

For Newton-Raphson, we use
All integrals in functions and derivatives are approximated by using Gauss-Hermite Quadrature.

The convolution equation of $\Delta_{ij}^*$ gives

$$f(\Delta_{ij}^*) = \frac{-\alpha_{ij}^* + \lambda^* \tau_{ij}^*}{1 + e^{\alpha_{ij}^* + \lambda^* \tau_{ij}^*}} \int \phi(z_i) dz_i = 0$$

In Newton-Raphson step, we use

$$\frac{\partial f(\Delta_{ij}^*)}{\partial \Delta_{ij}^*} = \int \frac{-\alpha_{ij}^* + \lambda^* \tau_{ij}^*}{1 + e^{\alpha_{ij}^* + \lambda^* \tau_{ij}^*}} \phi(z_i) dz_i$$

Full conditional distributions:

- $f(\log(\sigma^2), \lambda|\theta_{-b}, \theta_{-\sigma}, \theta_{-\lambda}, \chi, Y) \propto \prod_{t=2}^{n} \prod_{\lambda_t} \prod_{\log(\sigma_t^2)} \prod_{P(Y_{it}|\lambda_{it-1}, \theta)} f(b_{it}|\sigma_t) db_{it}$

$$\propto \prod_{t=2}^{n} \prod_{\lambda_t} \prod_{\log(\sigma_t^2)} \prod_{P(Y_{it}|\lambda_{it-1}, \theta)} \frac{e^{-\Lambda_{ij}^* \Delta_{ij}^*}}{2\pi \chi^2} \prod_{t=2}^{n} \prod_{\lambda_t} \prod_{\log(\sigma_t^2)} \prod_{P(Y_{it}|\lambda_{it-1}, \theta)} \frac{e^{-\Lambda_{ij}^* \Delta_{ij}^*}}{2\pi \chi^2} \prod_{t=2}^{n} \prod_{\lambda_t} \prod_{\log(\sigma_t^2)} \prod_{P(Y_{it}|\lambda_{it-1}, \theta)} \frac{e^{-\Lambda_{ij}^* \Delta_{ij}^*}}{2\pi \chi^2}$$

$D_{ii} = \sum_{j=1}^{r} \left[ \frac{e^{\Lambda_{ij}^* \Delta_{ij}^*}}{1 + e^{\Lambda_{ij}^* \Delta_{ij}^*}} \right] \prod_{t=1}^{n} \prod_{\lambda_t} \prod_{\log(\sigma_t^2)} \prod_{P(Y_{it}|\lambda_{it-1}, \theta)} \frac{e^{-\Lambda_{ij}^* \Delta_{ij}^*}}{2\pi \chi^2}$

- $f(\log(\sigma^2), \lambda|\theta_{-b}, \theta_{-\sigma}, \theta_{-\lambda}, \chi, Y) \propto \prod_{t=2}^{n} \prod_{\lambda_t} \prod_{\log(\sigma_t^2)} \prod_{P(Y_{it}|\lambda_{it-1}, \theta)} f(b_{it}|\sigma_t) db_{it}$

$$\propto \prod_{t=2}^{n} \prod_{\lambda_t} \prod_{\log(\sigma_t^2)} \prod_{P(Y_{it}|\lambda_{it-1}, \theta)} \frac{e^{-\Lambda_{ij}^* \Delta_{ij}^*}}{2\pi \chi^2} \prod_{t=2}^{n} \prod_{\lambda_t} \prod_{\log(\sigma_t^2)} \prod_{P(Y_{it}|\lambda_{it-1}, \theta)} \frac{e^{-\Lambda_{ij}^* \Delta_{ij}^*}}{2\pi \chi^2} \prod_{t=2}^{n} \prod_{\lambda_t} \prod_{\log(\sigma_t^2)} \prod_{P(Y_{it}|\lambda_{it-1}, \theta)} \frac{e^{-\Lambda_{ij}^* \Delta_{ij}^*}}{2\pi \chi^2}$$

$D_{ii} = \sum_{j=1}^{r} \left[ \frac{e^{\Lambda_{ij}^* \Delta_{ij}^*}}{1 + e^{\Lambda_{ij}^* \Delta_{ij}^*}} \right] \prod_{t=1}^{n} \prod_{\lambda_t} \prod_{\log(\sigma_t^2)} \prod_{P(Y_{it}|\lambda_{it-1}, \theta)} \frac{e^{-\Lambda_{ij}^* \Delta_{ij}^*}}{2\pi \chi^2}$
where \( D_{11} = \sum_i w_i \prod_{j=1}^r \left[ \frac{\sum_{l=1}^{\lambda_i} \phi(z_i^{(l)})^2}{1 + e^{-\Delta z_i^{(j)} + \lambda_i z_i^{(j)}}} \right] \)

- \( f(b_{it}[\theta_{-b}, Y]) \propto \prod_{j=1}^r [P(Y_{itj} | \theta)] P(b_{it} | \sigma_t) \)

\[ \propto \prod_{j=1}^r \left[ \frac{\sum_{l=1}^{\lambda_i} \phi(z_i^{(l)})^2}{1 + e^{-\Delta z_i^{(j)} + \lambda_i z_i^{(j)}}} \right]^{\nu_{itj}} \left( \frac{1}{1 + e^{-\Delta z_i^{(j)} + \lambda_i z_i^{(j)}}} \right)^{1-\nu_{itj}} \exp \left( -\frac{\beta_i^2}{2 \sigma_i^2} \right) \]

- \( f(b_{it}[\theta_{-b}, Y]) \propto \prod_{j=1}^r [P(Y_{itj} | Y_{it-1j}, \theta)] P(b_{it} | \sigma_t) \)

\[ \propto \prod_{j=1}^r \left[ \frac{\sum_{l=1}^{\lambda_i} \phi(z_i^{(l)})^2}{1 + e^{-\Delta z_i^{(j)} + \lambda_i z_i^{(j)}}} \right]^{\nu_{itj}} \left( \frac{1}{1 + e^{-\Delta z_i^{(j)} + \lambda_i z_i^{(j)}}} \right)^{1-\nu_{itj}} \exp \left( -\frac{\beta_i^2}{2 \sigma_i^2} \right) \]

- \( f(\beta^* | \theta_{-b}, Y) \propto \prod_{i=1}^N \prod_{j=1}^r [P(Y_{itj} | Y_{itj-1j}, \theta)] \Pi(\beta^*) \)

\[ \propto \prod_{i=1}^N \prod_{j=1}^r \prod_{t=2}^l \left[ \frac{\sum_{l=1}^{\lambda_i} \phi(z_i^{(l)})^2}{1 + e^{-\Delta z_i^{(j)} + \lambda_i z_i^{(j)}}} \right]^{\nu_{itj}} \left( \frac{1}{1 + e^{-\Delta z_i^{(j)} + \lambda_i z_i^{(j)}}} \right)^{1-\nu_{itj}} \exp \left( -\frac{\beta_i^2}{2 \sigma_i^2} \right) \]

- \( f(\beta | \theta_{-b}, Y) \propto \prod_{i=1}^N \prod_{j=1}^r [P(Y_{itj} | Y_{it-1j}, \theta)] \Pi(\beta) \)

\[ \propto \prod_{i=1}^N \prod_{j=1}^r \prod_{t=2}^l \left[ \frac{\sum_{l=1}^{\lambda_i} \phi(z_i^{(l)})^2}{1 + e^{-\Delta z_i^{(j)} + \lambda_i z_i^{(j)}}} \right]^{\nu_{itj}} \left( \frac{1}{1 + e^{-\Delta z_i^{(j)} + \lambda_i z_i^{(j)}}} \right)^{1-\nu_{itj}} \exp \left( -\frac{\beta_i^2}{2 \sigma_i^2} \right) \]

- \( f(\alpha_t | \theta_{-\alpha}, Y) \propto \prod_{i=1}^N \prod_{j=1}^r [P(Y_{itj} | Y_{it-1j}, \theta)] \Pi(\alpha_t) \)

\[ \propto \prod_{i=1}^N \prod_{j=1}^r \prod_{t=2}^l \left[ \frac{\sum_{l=1}^{\lambda_i} \phi(z_i^{(l)})^2}{1 + e^{-\Delta z_i^{(j)} + \lambda_i z_i^{(j)}}} \right]^{\nu_{itj}} \left( \frac{1}{1 + e^{-\Delta z_i^{(j)} + \lambda_i z_i^{(j)}}} \right)^{1-\nu_{itj}} \exp \left( -\frac{\beta_i^2}{2 \sigma_i^2} \right) \]

Derivatives of full conditional distributions:

For Hybrid MC, we need derivatives of full conditionals, as well as derivatives of \( \Delta_{itj} \) and \( \Delta_{itj}^* \). Necessary derivatives are obtained by chain rule and implicit differentiation. For simplicity of notation, define,

\[ A_{itj} = \int \frac{\sum_{l=1}^{\lambda_i} \phi(z_i^{(l)})^2}{1 + e^{-\Delta z_i^{(j)} + \lambda_i z_i^{(j)}}} \phi(z_i) dz_i \]

For \( t > 1 \),
\[ A_{ij} = \int \frac{e^{\Delta_{ij} + \lambda_j \phi(z_i)}}{(1 + e^{\Delta_{ij} + \lambda_j \phi(z_i)})^2} \phi(z_i) dz_i \]

\[ B_{ij} = \frac{e^{\Delta_{ij} + \lambda_j \phi(z_i) - \frac{\lambda_j}{2}}}{(1 + e^{\Delta_{ij} + \lambda_j \phi(z_i) - \frac{\lambda_j}{2}})^2} \]

\[ C_{ij} = \sum_{y_i \neq y_j} \frac{e^{\Delta_{ij} + \lambda_j \phi(y_i) - \frac{\lambda_j}{2}}}{(1 + e^{\Delta_{ij} + \lambda_j \phi(y_i) - \frac{\lambda_j}{2}})^2} P(Y_i = y_i) \]

\[ \frac{\partial \log (\alpha^2)}{\partial \log (\tau^2)} = \sum_i \frac{\partial D_i x_i / \partial \log (\tau^2)}{D_i} + 1 - 2 \frac{\log (\sigma^2)}{1 + \log (\sigma^2)} \]

where

\[ \frac{\partial D_i x_i}{\partial \log (\tau^2)} = \frac{\sum_j w_{ij} \left( e^{\Delta_{ij} + \lambda_j \phi(z_i) - \frac{\lambda_j}{2}} \right) \left[ \frac{\partial D_i x_i}{\partial \log (\tau^2)} + \lambda_j \phi(z_i) \right]}{(1 + e^{\Delta_{ij} + \lambda_j \phi(z_i) - \frac{\lambda_j}{2}})^2} \]

\[ \frac{\partial \Delta_{ij}}{\partial \log (\tau^2)} = \frac{\partial \Delta_{ij}}{\partial \log (\tau^2)} \]

\[ \frac{\partial \Delta_{ij}}{\partial \log (\tau^2)} = \frac{-\sqrt{2} \Delta_{ij} \sum_j w_{ij} \left( e^{\Delta_{ij} + \lambda_j \phi(z_i) - \frac{\lambda_j}{2}} \right)}{(1 + e^{\Delta_{ij} + \lambda_j \phi(z_i) - \frac{\lambda_j}{2}})^2} \]

\[ \frac{\partial \Delta_{ij}}{\partial \log (\tau^2)} = \frac{-\sqrt{2} \log (\tau^2) \sum_i w_{ij} \Delta_{ij} \left( e^{\Delta_{ij} + \lambda_j \phi(z_i) - \frac{\lambda_j}{2}} \right)}{(1 + e^{\Delta_{ij} + \lambda_j \phi(z_i) - \frac{\lambda_j}{2}})^2} \]

\[ \frac{\partial \log (\alpha^2)}{\partial \log (\tau^2)} \frac{\partial \log (\alpha^2)}{\partial \log (\tau^2)} = \sum_i \frac{\partial D_i x_i}{\partial \log (\tau^2)} - \frac{\lambda_j}{\sigma^2} \] for \( m = 2, \ldots, r \).
\[ \frac{\partial \log f(\theta; y)}{\partial \theta_k} = \sum_{i=1}^{N} \sum_{j=1}^{r} \sum_{t=1}^{n} \left[ (y_{itj} - \frac{\Delta_{itj}^{1} + \gamma_{itj}^{1}}{1 + e^{\Delta_{itj}^{1} + \gamma_{itj}^{1}}}) \frac{\partial \Delta_{itj}^{1}}{\partial \theta_k} - \frac{\Delta_{itj}^{1}}{\sigma^2} \right] \]

where \( \frac{\partial \Delta_{itj}^{1}}{\partial \theta_k} = \frac{\partial \Delta_{ijt}^{1} \cdot \partial \Delta_{ijt}^{1}}{\partial \theta_k} \) for \( t \geq 2 \).

From convolution equation, we have \( \frac{\partial \Delta_{itj}^{1}}{\partial \theta_k} = B_{itj}/A_{itj} \).

\[ \frac{\partial \Delta_{itj}^{1}}{\partial \theta_k} \] takes two forms depending on \( t \). For \( t = 2 \),

\[ \frac{\partial \Delta_{itj}^{1}}{\partial \theta_k} = \frac{X_{itj}^{1} e^{x_{itj}^{1g}}}{(1 + e^{x_{itj}^{1g}})^2} - \sum_{y_{itj} = 0}^{1} \frac{e^{\Delta_{itj}^{1} + y_{itj}}}{1 + e^{\Delta_{itj}^{1} + y_{itj}}^2} \cdot \frac{\partial P(Y_{itj} = y_{itj})}{\partial \theta_k} / C_{itj} \]

where \( \frac{\partial P(Y_{itj} = 0)}{\partial \theta_k} = \frac{X_{itj}^{1} e^{x_{itj}^{1g}}}{(1 + e^{x_{itj}^{1g}})^2} \) and \( \frac{\partial P(Y_{itj} = 1)}{\partial \theta_k} = \frac{X_{itj}^{1} e^{x_{itj}^{1g}}}{(1 + e^{x_{itj}^{1g}})^2} \)

\[ \frac{\partial \log f(\theta; y)}{\partial \theta_k} = \sum_{i=1}^{N} \sum_{j=1}^{r} \left[ (y_{ij} - \frac{\Delta_{ij}^{1} + \gamma_{ij}^{1}}{1 + e^{\Delta_{ij}^{1} + \gamma_{ij}^{1}}}) \frac{\partial \Delta_{ij}^{1}}{\partial \theta_k} + (y_{ij} - \frac{\Delta_{ij}^{2} + \gamma_{ij}^{2}}{1 + e^{\Delta_{ij}^{2} + \gamma_{ij}^{2}}}) \frac{\partial \Delta_{ij}^{2}}{\partial \theta_k} - \frac{\Delta_{ij}^{1}}{\sigma^2} \right] \]

where \( \frac{\partial \Delta_{ij}^{1}}{\partial \theta_k} = \frac{\partial \Delta_{ij}^{1} \cdot \partial \Delta_{ij}^{1}}{\partial \theta_k} \) from convolution equation of \( \Delta_{ij}^{1} \), and

\[ \frac{\partial \Delta_{ij}^{1}}{\partial \theta_k} = \frac{\partial \Delta_{ij}^{1} \cdot \partial \Delta_{ij}^{1}}{\partial \theta_k} \]

\[ \frac{\partial \Delta_{ij}^{2}}{\partial \theta_k} = B_{ij}/A_{ij} \]

\[ \frac{\partial \Delta_{ij}^{2}}{\partial \theta_k} = \frac{X_{ij}^{1} e^{x_{ij}^{1g}}}{(1 + e^{x_{ij}^{1g}})^2} - \sum_{y_{ij} = 0}^{1} \frac{e^{\Delta_{ij}^{1} + y_{ij}}}{1 + e^{\Delta_{ij}^{1} + y_{ij}}^2} \cdot \frac{\partial P(Y_{ij} = y_{ij})}{\partial \theta_k} / C_{ij} \]

where \( \frac{\partial P(Y_{ij} = 0)}{\partial \theta_k} = \frac{X_{ij}^{1} e^{x_{ij}^{1g}}}{(1 + e^{x_{ij}^{1g}})^2} \) and \( \frac{\partial P(Y_{ij} = 1)}{\partial \theta_k} = \frac{X_{ij}^{1} e^{x_{ij}^{1g}}}{(1 + e^{x_{ij}^{1g}})^2} \)
From convolution equation, \( \frac{\partial \Delta_{ij}}{\partial \eta_{ij}} = B_{ij} \frac{\partial \Delta_{ij}}{\partial \eta_{ij}} + y_{it-1j}/A_{itj} \)

where, from marginal constraint equations,

for \( t = 2 \),

\[
\frac{\partial \Delta_{ij}}{\partial \eta_{ij}} = \frac{-\frac{\Delta_{ij} \eta_{ij} + \eta_{ij} \eta_{ij}}{(1+e^{\Delta_{ij} + \eta_{ij} \eta_{ij}})^2}}{(1+e^{\Delta_{ij} + \eta_{ij} \eta_{ij}})^2} \frac{X_{ij} \eta_{ij}}{1+e^{\Delta_{ij} + \eta_{ij} \eta_{ij}}}
\]

for \( t > 2 \),

\[
\frac{\partial \Delta_{ij}}{\partial \eta_{ij}} = -\left\{ \sum_{y_{it-1j}=0} y_{it-1j} \frac{e^{\Delta_{ij} + \eta_{ij} \eta_{ij} + x_{it-1j}}}{(1+e^{\Delta_{ij} + \eta_{ij} \eta_{ij} + x_{it-1j}})^2} P(Y_{it-1j}) + \sum_{y_{it-1j}=0} e^{\Delta_{ij} + \eta_{ij} \eta_{ij} + x_{it-1j}} \frac{\partial P(Y_{it-1j})}{\partial \eta_{ij}} \right\} / C_{itj}
\]

\[
= -\left\{ \sum_{y_{it-1j}} y_{it-1j} \frac{e^{\Delta_{ij} + \eta_{ij} \eta_{ij} + x_{it-1j}}}{(1+e^{\Delta_{ij} + \eta_{ij} \eta_{ij} + x_{it-1j}})^2} P(Y_{it-1j}) \right\} / C_{itj}, \text{ since } \frac{\partial P(Y_{it-1j})}{\partial \eta_{ij}} = 0 \text{ for } p=1.
\]

4.10 Appendix B: Details on sampling from posterior distribution of parameters in MTREM(2)

For simplicity of notation, for \( t \geq 3 \), define:

\[
\pi_{yt-1j, y_{it-2j}}^{(t)} = P(Y_{it-1j} = y_{it-1j}, Y_{it-2j} = y_{it-2j})
\]

Note that,

\[
\pi_{yt-1j, y_{it-1j}}^{(t+1)} = P(Y_{itj} = y_{itj}, Y_{it-1j} = y_{it-1j})
\]

\[
= \sum_{y_{it-2j}} P(Y_{itj} = y_{itj} | Y_{it-1j} = y_{it-1j}, Y_{it-2j} = y_{it-2j}) \pi_{yt-1j, y_{it-2j}}^{(t)}
\]

\[
\pi_{yt-1j, y_{itj}}^{(3)} = P(Y_{itj} | Y_{itj}, \theta) P(Y_{itj} | \theta)
\]

\[
= \left( \frac{e^{\Delta_{ij} + \eta_{ij} \eta_{ij} + x_{itj}}}{1+e^{\Delta_{ij} + \eta_{ij} \eta_{ij} + x_{itj}}} \right) \frac{e^{\eta_{ij} \eta_{ij} \theta}}{1+e^{\Delta_{ij} \eta_{ij} + \eta_{ij} \theta}}
\]
Calculating $\Delta_{12j}$, $\Delta^*_{12j}$ and derivatives:

- Marginal constraint for initial state intercept, i.e. for $\Delta_{12j}$:

$$f(\Delta_{12j}) = \frac{-X_{12j}}{1 + e^{X_{12j}}} + \sum_{y_{t-1}} \frac{e^{\Delta_{12j} + \gamma_{12j}y_{12j}}}{1 + e^{\Delta_{12j} + \gamma_{12j}y_{12j}}} e^{\gamma_{12j}y_{12j}} = 0$$

$$\frac{\partial f(\Delta_{12j})}{\partial \Delta_{12j}} = \sum_{y_{t-1}} \frac{e^{\Delta_{12j} + \gamma_{12j}y_{12j}}}{(1 + e^{\Delta_{12j} + \gamma_{12j}y_{12j}})^2} e^{\gamma_{12j}y_{12j}}$$

$\Delta_{12j}$ is a function of $\gamma_{12j}$, $\tilde{\beta}$ and $\beta^*$. 

- Convolution equations for initial state intercepts, i.e. for $\Delta^*_{11j}$ and $\Delta^*_{12j}$:

$$f(\Delta^*_{11j}) = \frac{-X_{11j}}{1 + e^{X_{11j}}} + \int \frac{e^{\Delta_{11j} + \gamma_{11j}y_{11j}}}{1 + e^{\Delta_{11j} + \gamma_{11j}y_{11j}}} \phi(z_1) dz_1 = 0$$

$$\frac{\partial f(\Delta^*_{11j})}{\partial \Delta_{11j}} = \int \frac{e^{\Delta_{11j} + \gamma_{11j}y_{11j}}}{(1 + e^{\Delta_{11j} + \gamma_{11j}y_{11j}})^2} \phi(z_1) dz_1$$

$\Delta^*_{11j}$ is a function of $\beta^*$, $\lambda_j$, $\sigma_1$.

$$f(\Delta^*_{12j}) = \frac{-X_{12j}}{1 + e^{X_{12j}}} + \int \frac{e^{\Delta_{12j} + \gamma_{12j}y_{12j}}}{1 + e^{\Delta_{12j} + \gamma_{12j}y_{12j}}} \phi(z_1) dz_1 = 0$$

$$\frac{\partial f(\Delta^*_{12j})}{\partial \Delta_{12j}} = \int \frac{e^{\Delta_{12j} + \gamma_{12j}y_{12j}}}{(1 + e^{\Delta_{12j} + \gamma_{12j}y_{12j}})^2} \phi(z_1) dz_1$$

$\Delta^*_{12j}$ is a function of $\tilde{\gamma}_{12j}$, $\tilde{\lambda}_j$, $\sigma_2$, $\Delta_{12j}$ (hence $\tilde{\beta}$ and $\beta^*$).

- Marginal constraint for $\Delta_{11j}$ when $t \geq 3$:

$$P(Y_{12j} = 1) = \sum_{Y_{t-1-j} = Y_{t-2-j}} P(Y_{12j} = 1|Y_{t-1-j} = Y_{t-2-j}, Y_{t-2-j} = Y_{t-2-j}) \pi_Y(t)$$

$$f(\Delta_{11j}) = \frac{-X_{11j}}{1 + e^{X_{11j}}} + \sum_{Y_{t-1-j} = Y_{t-2-j}} \frac{e^{\Delta_{11j} + \gamma_{11j}y_{11j}}}{1 + e^{\Delta_{11j} + \gamma_{11j}y_{11j}}} e^{\gamma_{11j}y_{11j}} = 0$$

$$\frac{\partial f(\Delta_{11j})}{\partial \Delta_{11j}} = \sum_{Y_{t-1-j} = Y_{t-2-j}} \frac{e^{\Delta_{11j} + \gamma_{11j}y_{11j}}}{(1 + e^{\Delta_{11j} + \gamma_{11j}y_{11j}})^2} e^{\gamma_{11j}y_{11j}}$$

$$\frac{\partial f(\Delta_{12j})}{\partial \Delta_{12j}} = \sum_{Y_{t-1-j} = Y_{t-2-j}} \frac{e^{\Delta_{12j} + \gamma_{12j}y_{12j}}}{(1 + e^{\Delta_{12j} + \gamma_{12j}y_{12j}})^2} e^{\gamma_{12j}y_{12j}}$$

$$\frac{\partial f(\Delta_{12j})}{\partial \Delta_{12j}} = \sum_{Y_{t-1-j} = Y_{t-2-j}} \frac{e^{\Delta_{12j} + \gamma_{12j}y_{12j}}}{(1 + e^{\Delta_{12j} + \gamma_{12j}y_{12j}})^2} e^{\gamma_{12j}y_{12j}}$$
\( \Delta_{tij} \) is a function of \( \beta, \gamma_{itj,1}, \gamma_{itj,2} \) and \( \pi_{it-1, it-2}^{(t)} \) (hence \( \beta^*, \tilde{\beta}, \tilde{\gamma}_{itj}, (\gamma_{itj,1}, \gamma_{itj,2}, 3 \leq k \leq t), (\Delta_{itj}, 2 \leq k \leq t - 1)) \).

• Convolution equation for \( \Delta_{tij}^* \) when \( t \geq 3 \):

\[
f(\Delta_{tij}^*) = -\frac{\Delta_{tij}^* + \gamma_{itj,1} \Delta_{it-1, it-2} + \gamma_{itj,2} \Delta_{it-2, it-2} - \Delta_{itj,1} \Delta_{it-2, it-2}}{1 + e^{\Delta_{tij}^* + \lambda_j \sigma_j \phi(z_i)dz_i}} + \int f(\frac{\Delta_{tij}^* + \lambda_j \sigma_j \phi(z_i)dz_i}{1 + e^{\Delta_{tij}^* + \lambda_j \sigma_j \phi(z_i)dz_i}}) dz_i = 0
\]

\[
\frac{\partial f(\Delta_{tij}^*)}{\partial \Delta_{tij}^*} = \int f(\frac{\Delta_{tij}^* + \lambda_j \sigma_j \phi(z_i)dz_i}{1 + e^{\Delta_{tij}^* + \lambda_j \sigma_j \phi(z_i)dz_i}}) dz_i
\]

\( \Delta_{tij}^* \) is a function of \( \lambda_j, \sigma_j, \gamma_{itj,1}, \gamma_{itj,2}, (\Delta_{itj}, 2 \leq k \leq t) \) (hence \( \beta, \beta^*, \tilde{\beta}, \tilde{\gamma}_{itj}, (\gamma_{itj,1}, \gamma_{itj,2}, 3 \leq k \leq t - 1)) \).

**Full conditional distributions:**

• \( f(\log(\sigma_j^2), \lambda^*| \theta_{-b}, \theta_{-\sigma}, \theta_{-\lambda}, Y) \) takes the same form as in \( p=1 \).

• \( f(\log(\sigma_j^2), \lambda| \theta_{-b}, \theta_{-\sigma}, \theta_{-\lambda}, Y) \propto \Pi(\log(\sigma_j^2)) \prod_{l=1}^{\Pi} \prod_{j=1}^{N} \left[ \frac{1}{\int \prod_{l=1}^{\Pi} \prod_{j=1}^{N} \left[ \frac{\Delta_{tij}^* + \gamma_{itj,1} \Delta_{it-1, it-2} + \gamma_{itj,2} \Delta_{it-2, it-2} - \lambda_j \sigma_j \phi(z_i)dz_i}{1 + e^{\Delta_{tij}^* + \lambda_j \sigma_j \phi(z_i)dz_i}} \right] \right]^{1 - \nu_{itj}} \frac{1}{\sigma_j^2} e^{-\frac{\lambda_j^2}{2\sigma_j^2}} db_{itj} \]

\[
\approx \frac{e^{|\log(\sigma_j^2)|}}{(1 + e^{\log(\sigma_j^2)})^2} \prod_{j=1}^{\Pi} \prod_{l=1}^{N} \left[ \frac{\Delta_{tij}^* + \gamma_{itj,1} \Delta_{it-1, it-2} + \gamma_{itj,2} \Delta_{it-2, it-2} - \lambda_j \sigma_j \phi(z_i)dz_i}{1 + e^{\Delta_{tij}^* + \lambda_j \sigma_j \phi(z_i)dz_i}} \right]^{1 - \nu_{itj}} \frac{1}{\sigma_j^2} e^{-\frac{\lambda_j^2}{2\sigma_j^2}} db_{itj} \]

\[
\equiv \frac{e^{|\log(\sigma_j^2)|}}{(1 + e^{\log(\sigma_j^2)})^2} \prod_{j=1}^{\Pi} \prod_{l=1}^{N} \left[ \frac{\Delta_{tij}^* + \gamma_{itj,1} \Delta_{it-1, it-2} + \gamma_{itj,2} \Delta_{it-2, it-2} - \lambda_j \sigma_j \phi(z_i)dz_i}{1 + e^{\Delta_{tij}^* + \lambda_j \sigma_j \phi(z_i)dz_i}} \right]^{1 - \nu_{itj}} \frac{1}{\sigma_j^2} e^{-\frac{\lambda_j^2}{2\sigma_j^2}} db_{itj} \]

where \( E_{itj} = \sum_{j=1}^{\Pi} \prod_{l=1}^{N} \left[ \frac{\Delta_{tij}^* + \gamma_{itj,1} \Delta_{it-1, it-2} + \gamma_{itj,2} \Delta_{it-2, it-2} - \lambda_j \sigma_j \phi(z_i)dz_i}{1 + e^{\Delta_{tij}^* + \lambda_j \sigma_j \phi(z_i)dz_i}} \right]^{1 - \nu_{itj}} \frac{1}{\sigma_j^2} e^{-\frac{\lambda_j^2}{2\sigma_j^2}} db_{itj} \)

• For \( t > 2 \):

\[
f(\log(\sigma_j^2), \lambda| \theta_{-b}, \theta_{-\sigma}, \theta_{-\lambda}, Y) \propto \Pi(\log(\sigma_j^2)) \prod_{l=1}^{\Pi} \prod_{j=1}^{N} \left[ \frac{1}{\int \prod_{l=1}^{\Pi} \prod_{j=1}^{N} \left[ \frac{\Delta_{tij}^* + \gamma_{itj,1} \Delta_{it-1, it-2} + \gamma_{itj,2} \Delta_{it-2, it-2} - \lambda_j \sigma_j \phi(z_i)dz_i}{1 + e^{\Delta_{tij}^* + \lambda_j \sigma_j \phi(z_i)dz_i}} \right]^{1 - \nu_{itj}} \frac{1}{\sigma_j^2} e^{-\frac{\lambda_j^2}{2\sigma_j^2}} db_{itj} \]

\[
\approx \prod_{j=1}^{\Pi} \frac{e^{|\log(\sigma_j^2)|}}{(1 + e^{\log(\sigma_j^2)})^2} \prod_{j=1}^{\Pi} \prod_{l=1}^{N} \left[ \frac{\Delta_{tij}^* + \gamma_{itj,1} \Delta_{it-1, it-2} + \gamma_{itj,2} \Delta_{it-2, it-2} - \lambda_j \sigma_j \phi(z_i)dz_i}{1 + e^{\Delta_{tij}^* + \lambda_j \sigma_j \phi(z_i)dz_i}} \right]^{1 - \nu_{itj}} \frac{1}{\sigma_j^2} e^{-\frac{\lambda_j^2}{2\sigma_j^2}} db_{itj} \]
\[ E_{it} = \sum_{j=1}^{r} u_{ij} \prod_{j=1}^{r} \left( \frac{A_{itj} + \sqrt{A_{itj} \text{log}(e)^{1/2}}}{1 + A_{itj} + \sqrt{A_{itj} \text{log}(e)^{1/2}}} \right) \]

where \( E_{it} \) is the expected value of the model.

- \( f(b_{it})|\theta_{-b}, \mathbf{Y} \) for \( t \geq 1 \) takes similar forms to the ones for \( p=1 \).

\[ f(\beta|\theta_{-\beta}, \mathbf{Y}) \propto \prod_{i,j} \left[ (\frac{A_{ij} + \lambda_{ij} y_{ij}}{1 + A_{ij} + \lambda_{ij} y_{ij}})^{1/2} \right]^{y_{ij}} \left( \frac{1}{1 + A_{ij} + \lambda_{ij} y_{ij}} \right)^{1/2} \left( \frac{A_{ij} + \lambda_{ij} y_{ij}}{1 + A_{ij} + \lambda_{ij} y_{ij}} \right)^{1/2} \left( \frac{1}{1 + A_{ij} + \lambda_{ij} y_{ij}} \right)^{1/2} \]

Derivatives of full conditional distributions:

- \( \frac{\partial \log f(\mathbf{Y}, \mathbf{x}, \mathbf{z}, \mathbf{w}, \mathbf{a}, \mathbf{b}, \theta)}{\partial \log (e)} \) takes the same form as \( p=1 \).

- \( \frac{\partial \log f(\mathbf{Y}, \mathbf{x}, \mathbf{z}, \mathbf{w}, \mathbf{a}, \mathbf{b}, \theta)}{\partial x_{i}} \) takes the same form as \( p=1 \).

- \( \frac{\partial \log f(\mathbf{Y}, \mathbf{x}, \mathbf{z}, \mathbf{w}, \mathbf{a}, \mathbf{b}, \theta)}{\partial z_{i}} \) are calculated as:

\[ \sum_{j} \left[ Y_{ij} - \frac{A_{ij} + \lambda_{ij} y_{ij}}{1 + A_{ij} + \lambda_{ij} y_{ij}} \right] \frac{A_{ij} + \lambda_{ij} y_{ij}}{1 + A_{ij} + \lambda_{ij} y_{ij}} \frac{y_{ij}}{\theta} + \left( Y_{ij} - \frac{A_{ij} + \lambda_{ij} y_{ij}}{1 + A_{ij} + \lambda_{ij} y_{ij}} \right) \frac{\partial A_{ij}}{\partial z_{i}} \]
\[\sum_{i} \sum_{j} \sum_{k} \left[ (Y_{ij} - \frac{e^{\Delta_{ij}^{(1)}} + 2Y_{ij}}{1 + e^{\Delta_{ij}^{(1)}} + 2Y_{ij}}) \frac{\partial \Delta_{ij}^{(1)}}{\partial \theta_k} \right] - \frac{\partial \gamma}{\partial \theta_k}\]

where \(\frac{\partial \Delta_{ij}^{(1)}}{\partial \theta_k} = \frac{X_{ij}X_{ij}^{*} \phi \left( \frac{X_{ij}X_{ij}^{*}}{1 + X_{ij}X_{ij}^{*}} \right)}{f \left( 1 + e^{\Delta_{ij}^{(1)}} + 2Y_{ij} \right) \phi \left( \frac{X_{ij}X_{ij}^{*}}{1 + X_{ij}X_{ij}^{*}} \right)}\)

For \(t \geq 3\),

\[\frac{\partial \Delta_{ij}^{(t)}}{\partial X_{ik}} = \frac{\Delta_{ij}^{(t)} - \Delta_{ij}^{(t-1)}}{X_{ik}^{2} + \Delta_{ij}^{(t-1)} + X_{ik}^{2} + \Delta_{ij}^{(t-1)}}\]

\[\frac{\partial \Delta_{ij}^{(t)}}{\partial Y_{ij}} = \frac{\Delta_{ij}^{(t)} - \Delta_{ij}^{(t-1)}}{X_{ik}^{2} + \Delta_{ij}^{(t-1)} + X_{ik}^{2} + \Delta_{ij}^{(t-1)}}\]

\[\frac{\partial \Delta_{ij}^{(t)}}{\partial \phi_{ij}} = \frac{\Delta_{ij}^{(t)} - \Delta_{ij}^{(t-1)}}{X_{ik}^{2} + \Delta_{ij}^{(t-1)} + X_{ik}^{2} + \Delta_{ij}^{(t-1)}}\]

\[\text{where } \frac{\partial \Delta_{ij}^{(t)}}{\partial \theta_k} = \frac{\partial \Delta_{ij}^{(t+1)}}{\partial \theta_k} \frac{\partial \Delta_{ij}^{(t)}}{\partial \theta_k}\]

\[\frac{\partial \Delta_{ij}^{(1)}}{\partial X_{ik}} = \frac{\Delta_{ij}^{(1)} - \Delta_{ij}^{(0)}}{X_{ik}^{2} + \Delta_{ij}^{(0)} + X_{ik}^{2} + \Delta_{ij}^{(0)}}\]

\[\frac{\partial \Delta_{ij}^{(1)}}{\partial Y_{ij}} = \frac{\Delta_{ij}^{(1)} - \Delta_{ij}^{(0)}}{X_{ik}^{2} + \Delta_{ij}^{(0)} + X_{ik}^{2} + \Delta_{ij}^{(0)}}\]

\[\frac{\partial \Delta_{ij}^{(1)}}{\partial \phi_{ij}} = \frac{\Delta_{ij}^{(1)} - \Delta_{ij}^{(0)}}{X_{ik}^{2} + \Delta_{ij}^{(0)} + X_{ik}^{2} + \Delta_{ij}^{(0)}}\]

\[\frac{\partial \Delta_{ij}^{(t+1)}}{\partial \theta_k} = \frac{\Delta_{ij}^{(t+1)} - \Delta_{ij}^{(t)}}{X_{ik}^{2} + \Delta_{ij}^{(t)} + X_{ik}^{2} + \Delta_{ij}^{(t)}}\]

\[\frac{\partial \Delta_{ij}^{(t+1)}}{\partial Y_{ij}} = \frac{\Delta_{ij}^{(t+1)} - \Delta_{ij}^{(t)}}{X_{ik}^{2} + \Delta_{ij}^{(t)} + X_{ik}^{2} + \Delta_{ij}^{(t)}}\]

\[\frac{\partial \Delta_{ij}^{(t+1)}}{\partial \phi_{ij}} = \frac{\Delta_{ij}^{(t+1)} - \Delta_{ij}^{(t)}}{X_{ik}^{2} + \Delta_{ij}^{(t)} + X_{ik}^{2} + \Delta_{ij}^{(t)}}\]
For $t \geq 3$,

$$\frac{\partial \Delta_{ij}^{(t)}}{\partial \beta_{k}} = \frac{\partial \Delta_{ij}^{(t)}}{\partial \beta_{k}} \frac{\partial \Delta_{ij}^{(t)}}{\partial \beta_{k}}$$

$$\frac{\partial \Delta_{ij}^{(t)}}{\partial \beta_{k}} = - \sum_{n, j, k, l} \left( \frac{\Delta_{ij}^{(t)} + \Delta_{kl}^{(t)} - \Delta_{il}^{(t)} - \Delta_{jk}^{(t)}}{1 + e^{-\Delta_{ij}^{(t)} - \Delta_{kl}^{(t)}}} \right) \frac{\partial \Delta_{ij}^{(t)}}{\partial \beta_{k}} \left( \frac{\partial \Delta_{ij}^{(t)}}{\partial \beta_{k}} \right)$$

$$\frac{\partial \log f(\theta, \beta, \mathbf{y})}{\partial \beta_{k}} = \sum_{i} \sum_{j} \left( Y_{ij} - \frac{\Delta_{ij}^{(t)} + \Delta_{ji}^{(t)} - \Delta_{il}^{(t)} - \Delta_{jk}^{(t)}}{1 + e^{-\Delta_{ij}^{(t)} - \Delta_{ji}^{(t)}}} \right) \frac{\partial \Delta_{ij}^{(t)}}{\partial \beta_{k}} = \frac{\mathbf{g}_{t}}{\sigma_{g_{t}}}$$

$$\frac{\partial \Delta_{ij}^{(t)}}{\partial \beta_{k}} = \frac{\partial \Delta_{ij}^{(t)}}{\partial \beta_{k}} \frac{\partial \Delta_{ij}^{(t)}}{\partial \beta_{k}}$$

$$\frac{\partial \Delta_{ij}^{(t)}}{\partial \beta_{k}} = \frac{X_{ij}^{(t)} - \sum_{k, l} \Delta_{ij}^{(t)} \Delta_{kl}^{(t)} \Delta_{il}^{(t)} \Delta_{jk}^{(t)}}{1 + e^{-\Delta_{ij}^{(t)} - \Delta_{kl}^{(t)}}} \frac{\partial \Delta_{ij}^{(t)}}{\partial \beta_{k}} \left( \frac{\partial \Delta_{ij}^{(t)}}{\partial \beta_{k}} \right)$$

$$\frac{\partial \log f(\theta, \beta, \mathbf{y})}{\partial \beta_{k}} = \sum_{i} \sum_{j} \left( Y_{ij} - \frac{\Delta_{ij}^{(t)} + \Delta_{ji}^{(t)} - \Delta_{il}^{(t)} - \Delta_{jk}^{(t)}}{1 + e^{-\Delta_{ij}^{(t)} - \Delta_{ji}^{(t)}}} \right) \frac{\partial \Delta_{ij}^{(t)}}{\partial \beta_{k}} = \frac{\mathbf{g}_{t}}{\sigma_{g_{t}}}$$

$$\frac{\partial \Delta_{ij}^{(t)}}{\partial \beta_{k}} = \frac{\partial \Delta_{ij}^{(t)}}{\partial \beta_{k}} \frac{\partial \Delta_{ij}^{(t)}}{\partial \beta_{k}}$$

$$\frac{\partial \Delta_{ij}^{(t)}}{\partial \beta_{k}} = \frac{X_{ij}^{(t)} - \sum_{k, l} \Delta_{ij}^{(t)} \Delta_{kl}^{(t)} \Delta_{il}^{(t)} \Delta_{jk}^{(t)}}{1 + e^{-\Delta_{ij}^{(t)} - \Delta_{kl}^{(t)}}} \frac{\partial \Delta_{ij}^{(t)}}{\partial \beta_{k}} \left( \frac{\partial \Delta_{ij}^{(t)}}{\partial \beta_{k}} \right)$$

$$\frac{\partial \log f(\theta, \beta, \mathbf{y})}{\partial \beta_{k}} = \sum_{i} \sum_{j} \left( Y_{ij} - \frac{\Delta_{ij}^{(t)} + \Delta_{ji}^{(t)} - \Delta_{il}^{(t)} - \Delta_{jk}^{(t)}}{1 + e^{-\Delta_{ij}^{(t)} - \Delta_{ji}^{(t)}}} \right) \frac{\partial \Delta_{ij}^{(t)}}{\partial \beta_{k}} = \frac{\mathbf{g}_{t}}{\sigma_{g_{t}}}$$

for $t \geq 4$.

* for $t = 3$: $\frac{\partial \log f(\theta, \beta, \mathbf{y})}{\partial \beta_{k}} = \sum_{i} \sum_{j} \left( y_{ij} - \frac{\Delta_{ij}^{(3)} + \Delta_{ji}^{(3)} - \Delta_{il}^{(3)} - \Delta_{jk}^{(3)}}{1 + e^{-\Delta_{ij}^{(3)} - \Delta_{ji}^{(3)}}} \right) \frac{\partial \Delta_{ij}^{(3)}}{\partial \beta_{k}} = \frac{\mathbf{g}_{3}}{\sigma_{g_{3}}}$
where

\[ \frac{\partial \Delta_{ij}^{(1)}}{\partial Y_{i+1,j}} = \frac{\partial \Delta_{ij}^{(1)}}{\partial Y_{i,j+1}} = \frac{\partial \Delta_{ij}^{(1)}}{\partial Y_{i,j+1}} = \frac{\partial \Delta_{ij}^{(1)}}{\partial Y_{i+1,j}} = \Sigma w_k \Delta_{ij}^{(1)} \nu_{ij} \frac{1}{\nu_{ij}} \]

\[ \frac{\partial \Delta_{ij}^{(1)}}{\partial Y_{i,j+1}} = \frac{\partial \Delta_{ij}^{(1)}}{\partial Y_{i+1,j}} = \frac{\partial \Delta_{ij}^{(1)}}{\partial Y_{i,j}} = \frac{\partial \Delta_{ij}^{(1)}}{\partial Y_{i+1,j}} = \sum_{\nu_{i-1,j}, \nu_{i,j-1}} \frac{\Delta_{ij}^{(1)} + \nu_{ij} \nu_{i-1,j} \nu_{i,j-1}}{\nu_{ij}} \]

\[ \frac{\partial \Delta_{ij}^{(1)}}{\partial Y_{i+1,j}} = \frac{\partial \Delta_{ij}^{(1)}}{\partial Y_{i,j+1}} = \frac{\partial \Delta_{ij}^{(1)}}{\partial Y_{i,j}} = \frac{\partial \Delta_{ij}^{(1)}}{\partial Y_{i+1,j}} = \sum_{\nu_{i-1,j}, \nu_{i,j-1}} \frac{\Delta_{ij}^{(1)} + \nu_{ij} \nu_{i-1,j} \nu_{i,j-1}}{\nu_{ij}} \]

\[ \frac{\partial \Delta_{ij}^{(1)}}{\partial Y_{i+1,j}} = \frac{\partial \Delta_{ij}^{(1)}}{\partial Y_{i,j+1}} = \frac{\partial \Delta_{ij}^{(1)}}{\partial Y_{i,j}} = \frac{\partial \Delta_{ij}^{(1)}}{\partial Y_{i+1,j}} = \sum_{\nu_{i-1,j}, \nu_{i,j-1}} \frac{\Delta_{ij}^{(1)} + \nu_{ij} \nu_{i-1,j} \nu_{i,j-1}}{\nu_{ij}} \]

**Derivatives of joint distributions:**

Unlike \( p=1 \), we need to calculate and update bivariate probabilities. These probabilities are used in the calculation of marginal constraints, and derivatives are required for the Hybrid step of some parameters.

Recall that, for \( t \geq 4 \),

\[ \pi_{y_{i+1},y_{i+1}}^{(3)} = P(Y_{i+1,j} | Y_{i,j}, \theta)P(Y_{i+1,j} | \theta) = \frac{e^{\Delta_{ij} + \gamma_{ij} Y_{i+1,j} \gamma_{ij}}}{1 + e^{\Delta_{ij} + \gamma_{ij}}} \]

\[ \frac{\partial \pi_{y_{i+1},y_{i+1}}^{(3)}}{\partial \theta} = \left( e^{\Delta_{ij} + \gamma_{ij} Y_{i+1,j} \gamma_{ij}} \right) \frac{\left[ \frac{\partial \Delta_{ij}}{\partial \theta} + \frac{\partial \gamma_{ij}}{\partial \theta} \right] Y_{i+1,j} \gamma_{ij}}{1 + e^{\Delta_{ij} + \gamma_{ij}}} \]

\[ \frac{\partial \pi_{y_{i+1},y_{i+1}}^{(3)}}{\partial \theta} = \left( e^{\Delta_{ij} + \gamma_{ij} Y_{i+1,j} \gamma_{ij}} \right) \frac{\left[ \frac{\partial \gamma_{ij}}{\partial \theta} Y_{i+1,j} \gamma_{ij} \right]}{1 + e^{\Delta_{ij} + \gamma_{ij}}} \]

\[ \frac{\partial \pi_{y_{i+1},y_{i+1}}^{(3)}}{\partial \theta} = \left( e^{\Delta_{ij} + \gamma_{ij} Y_{i+1,j} \gamma_{ij}} \right) \frac{\left[ \frac{\partial \Delta_{ij}}{\partial \theta} Y_{i+1,j} \gamma_{ij} \right]}{1 + e^{\Delta_{ij} + \gamma_{ij}}} \]

The derivative of \( \pi_{y_{i+1},y_{i+1}}^{(3)} \) with respect to any other parameter is zero. After calculating \( \pi_{y_{i+1},y_{i+1}}^{(3)} \) and \( \frac{\partial \pi_{y_{i+1},y_{i+1}}^{(3)}}{\partial \theta} \), we need to update \( \pi_{y_{i+1},y_{i+1}}^{(1)} \) for \( t \geq 4 \).
Recall that,

\[ p^{(t)}_{y_{t+1}, y_{t+2} | y_{t-1} = y_{t-1}, y_{t-2} = y_{t-2}} = P(Y_{t+1} = y_{t+1}, Y_{t+2} = y_{t+2}) \]

\[ = \sum_{y_{t-2}} \left( \frac{e^{\Delta t_{t-1} + \gamma_{t-1} t_{t-1} + \gamma_{t-2} t_{t-2} + \gamma_{t-2} t_{t-2}}}{1 + e^{\Delta t_{t-1} + \gamma_{t-1} t_{t-1} + \gamma_{t-2} t_{t-2} + \gamma_{t-2} t_{t-2}}} \right) \eta_{y_{t-1}, 2} - \eta_{y_{t-1}, 3} \]

\[ \sum_{y_{t-3}} \left( \frac{e^{\Delta t_{t-1} + \gamma_{t-1} t_{t-1} + \gamma_{t-2} t_{t-2} + \gamma_{t-3} t_{t-3}}}{1 + e^{\Delta t_{t-1} + \gamma_{t-1} t_{t-1} + \gamma_{t-2} t_{t-2} + \gamma_{t-3} t_{t-3}}} \right) \eta_{y_{t-1}, 2} - \eta_{y_{t-1}, 3} \]

\[ \frac{\partial p^{(t)}_{y_{t+1}, y_{t+2} | y_{t-1} = y_{t-1}, y_{t-2} = y_{t-2}}}{\partial a_{k}} = \sum_{y_{t-3}} \left( \frac{e^{\Delta t_{t-1} + \gamma_{t-1} t_{t-1} + \gamma_{t-2} t_{t-2} + \gamma_{t-3} t_{t-3}}}{1 + e^{\Delta t_{t-1} + \gamma_{t-1} t_{t-1} + \gamma_{t-2} t_{t-2} + \gamma_{t-3} t_{t-3}}} \right) \frac{\partial \psi^{(t-1)}_{y_{t-1}}}{\partial a_{k}} \eta_{y_{t-1}, 2} - \eta_{y_{t-1}, 3} \]

\[ \sum_{y_{t-3}} \left( \frac{e^{\Delta t_{t-1} + \gamma_{t-1} t_{t-1} + \gamma_{t-2} t_{t-2} + \gamma_{t-3} t_{t-3}}}{1 + e^{\Delta t_{t-1} + \gamma_{t-1} t_{t-1} + \gamma_{t-2} t_{t-2} + \gamma_{t-3} t_{t-3}}} \right) \frac{\partial \psi^{(t-1)}_{y_{t-1}}}{\partial a_{k}} \eta_{y_{t-1}, 2} - \eta_{y_{t-1}, 3} \]

\[ \frac{\partial p^{(t)}_{y_{t+1}, y_{t+2} | y_{t-1} = y_{t-1}, y_{t-2} = y_{t-2}}}{\partial b_{k}} = \sum_{y_{t-3}} \left( \frac{e^{\Delta t_{t-1} + \gamma_{t-1} t_{t-1} + \gamma_{t-2} t_{t-2} + \gamma_{t-3} t_{t-3}}}{1 + e^{\Delta t_{t-1} + \gamma_{t-1} t_{t-1} + \gamma_{t-2} t_{t-2} + \gamma_{t-3} t_{t-3}}} \right) \frac{\partial \psi^{(t-1)}_{y_{t-1}}}{\partial b_{k}} \eta_{y_{t-1}, 2} - \eta_{y_{t-1}, 3} \]

\[ \sum_{y_{t-3}} \left( \frac{e^{\Delta t_{t-1} + \gamma_{t-1} t_{t-1} + \gamma_{t-2} t_{t-2} + \gamma_{t-3} t_{t-3}}}{1 + e^{\Delta t_{t-1} + \gamma_{t-1} t_{t-1} + \gamma_{t-2} t_{t-2} + \gamma_{t-3} t_{t-3}}} \right) \frac{\partial \psi^{(t-1)}_{y_{t-1}}}{\partial b_{k}} \eta_{y_{t-1}, 2} - \eta_{y_{t-1}, 3} \]
The derivative of $\pi^{(t-1)}_{y_{1t-1}, y_{1t-2}}$ with respect to any other parameter is zero.

4.11 Appendix C: Details on missing data in MTREM(1)

In IYFP data, the covariates used are $X_{itj2} =$ gender, $X_{itj3} =$ negative life events1 (NLE1), $X_{itj4} =$ NLE2, $X_{itj5} =$ negative economical events (NEE), $X_{itj6} =$ Cutbacks1, $X_{itj7} =$ Cutbacks2, $X_{itj8} =$ Response1, $X_{itj9} =$ Response2, $X_{itj10} =$ Time1, $X_{itj11} =$Time2. We have missing observations only in the following covariates: $X_{itj3}, X_{itj4}, X_{itj5}, X_{itj6}, X_{itj7}$. All missing covariates and responses are observed after baseline.

Computational Algorithm for Data Augmentation:

Let $\theta$ be the all parameters in the "Y model", and $\psi$ be the parameters related to joint distribution of covariates.

1. Set $k=0$. Set starting values for $Y_{itj,mis}, X_{itj,mis}, \psi, \theta$.
2. Run "X model" for 1000 iterations, and randomly select 5 sets of $X$ matrices. We define "X model" as the marginal (over $Y$) model for the missing covariates: $X_{itj,mis}|X_{itj,obs}, \psi$. Specifically, repeat these two steps for 1000 iterations:
a. Impute $X_{it,j,miss}^{(k+1)}$ from the covariate distribution of $X$ that accounts for time dependence. (See below for details.)

b. Sample from the full conditional distribution of $\psi^{(k+1)}$ by Hybrid MC.

3. For $t > 1$, impute $Y_{it,j,miss} | X_{it,j}$ by the following steps:
   a. Set $l = 2$.
   b. Calculate $\Delta_{itj}$ (which uses $Y_{it-1,j}$) and $p_{itj} = P(Y_{it,j,miss}^{(k+1)} = 1 | Y_{it,j}; m \neq i; \theta)$. If missing cases are dropouts, then $p_{itj} = \frac{e^{\Delta_{itj} + \lambda_{itj} b_{itj}}}{1 + e^{\Delta_{itj} + \lambda_{itj} b_{itj}}}$. If they are intermittent missing, then $p_{itj}$ is
      \[
      \frac{e^{\Delta_{itj} + \lambda_{itj} b_{itj}}}{1 + e^{\Delta_{itj} + \lambda_{itj} b_{itj}}} \cdot \frac{1 - e^{-\lambda_{itj} b_{itj}}}{1 - e^{\Delta_{itj} + \lambda_{itj} b_{itj}}}.
      \]
   c. Generate $u$ from Uniform(0,1). If $u < p_{itj}$, then set $\tilde{Y}_{it,j} = 1$.
   d. Set $l = l + 1$. Go back to step 3b. Repeat until responses at all time points are imputed.

4. Sample from the full conditional distribution of $\theta^{(k+1)}$. This is the Gibbs sampling step of complete case analysis.

5. Set $k = k + 1$. Go back to step 3. Repeat until convergence is obtained.

With this algorithm, we are basically using the following partition of the likelihood: $P(Y | X, \theta) P(X | \psi)$.

**Joint distribution of covariates:**

- Exploiting time-varying structure of covariates:

At baseline, we don't have any cases with missing covariates. For $t > 1$, we extend Ibrahim et al.'s (2002) work to include the previous time covariate information. For simplicity of notation, we now suppress the indexes $i$ and $j$. Let $X_t = (X_{t,1,2}, X_{t,1,3}, X_{t,1,4}, X_{t,1,5}, X_{t,1,6}, X_{t,1,7}, X_{t,1,10}, X_{t,1,11})$ and $X_{t-1} = (X_{t-1,1,2}, X_{t-1,1,3}, X_{t-1,1,4}, X_{t-1,1,5}, X_{t-1,1,6}, X_{t-1,1,7})$.

Under a Markov assumption, the joint distribution of covariates can be factored as

\[ P(X_2, \ldots, X_n | \psi) = P(X_2 | X_1, \psi) P(X_3 | X_2, \psi) \ldots P(X_n | X_{n-1}, \psi) \]

For each time, we factor the conditional distribution as
\[ P(X_4|X_{i-1}, \psi) = P(x_{13}, x_{14}, x_{22}, x_{210}, x_{111}, X_{i-1}, \psi_3) P(x_{15}, x_{12}, x_{13}, x_{14}, x_{110}, x_{111}, X_{i-1}, \psi_3) \]

\[ \times P(x_{16}, x_{17}, x_{22}, x_{23}, x_{14}, x_{15}, x_{110}, x_{111}, X_{i-1}, \psi_3). \]

To sample from the joint distribution of \( X_{i1} \), we use following method: Since missing covariates are observed only on NLE (3 levels), NEE (2 levels), and cutbacks (3 levels), we have \( 3^2 \times 3 = 18 \) possible combinations of imputed values. We calculate the probability of all these 18 possible combinations of the missing \( X \) for each individual, sort and add these probabilities to obtain cumulative density function (CDF). Then, a variate \( u \) is uniformly sampled from \( \text{Unif}(0,1) \). If \( u \) is less than the smallest CDF, then combination that corresponds to this probability is chosen. If \( u \) is greater than first CDF, but less than second CDF, then we choose the combination that corresponds to second CDF, and so on. Then, set \( X^{(k+1)}_{ij, \text{mis}} = X^{(k+1)}_{ij, \text{mis}} \) for \( j = 2, \ldots, J \) since covariates should be same for all response items for a specific subject at a given time point.

If covariates in both previous and next time points are observed, we used information from both neighbours to impute the current covariate. That is, if \( X_{i1+11} \) is observed, we use

\[ P(X_{d1}|X_{im1}; m \neq l|\psi) \propto P(X_{i,l-1,1}, X_{d11}, X_{i,l+1,1}|\psi) = P(X_{d11}|X_{i,l-1,1}, \psi)P(X_{i,l+1,1}|X_{d11}, \psi). \]

For covariates with three 'ordered' levels (i.e. NLE and cutbacks), we assumed ordinal logistic regression with separate intercepts, but common slopes. For binary covariates, we assumed logistic regression. Specifically, the conditional distributions are assumed to be

\[ \logit P(x_{13} = 1|x_{12}, X_{i-1}, \psi_3) = \epsilon_{31} + \epsilon_{32} x_{12} + \epsilon_{33} x_{i-1,3} + \epsilon_{34} x_{i-1,4} + \epsilon_{35} x_{i-1,5} + \epsilon_{36} x_{i-1,6} + \epsilon_{37} x_{i-1,7} + \epsilon_{38} x_{i,10} + \epsilon_{39} x_{i,11} \equiv CUM_{14} \]

\[ \logit P(x_{14} = 1|x_{12}, X_{i-1}, \psi_3) = \epsilon_{3,10} + \epsilon_{32} x_{12} + \epsilon_{33} x_{i-1,3} + \epsilon_{34} x_{i-1,4} + \epsilon_{35} x_{i-1,5} + \epsilon_{36} x_{i-1,6} + \epsilon_{37} x_{i-1,7} + \epsilon_{38} x_{i,10} + \epsilon_{39} x_{i,11} \equiv CUM_{24} \]

where \( \psi_3 = (\epsilon_{31}, \epsilon_{32}, \epsilon_{33}, \epsilon_{34}, \epsilon_{35}, \epsilon_{36}, \epsilon_{37}, \epsilon_{38}, \epsilon_{39}, \epsilon_{3,10}) \).
logit\( P(x_{t,5} \mid x_{t,2}, x_{t,3}, x_{t,4}, X_{t-1}, \psi_5) = \beta_0 + \beta_2 x_{t,2} + \beta_3 x_{t-1,3} + \beta_4 x_{t-1,4} + \beta_5 x_{t,5} + \beta_6 x_{t-1,6} + \beta_7 x_{t-1,7} + \beta_8 x_{t,10} + \beta_9 x_{t,11} + \beta_{10} x_{t,4} + \beta_{11} x_{t,4} \equiv \text{CUM}3 \)

\[
\text{logit}\ P(x_{t,5} \mid x_{t,2}, x_{t,3}, x_{t,4}, X_{t-1}, \psi_6) = \beta_0 + \beta_2 x_{t,2} + \beta_3 x_{t-1,3} + \beta_4 x_{t-1,4} + \beta_5 x_{t-1,5} + \beta_6 x_{t-1,6} + \beta_7 x_{t-1,7} + \beta_8 x_{t,10} + \beta_9 x_{t,11} + \beta_{10} x_{t,4} + \beta_{11} x_{t,4} \equiv \text{CUM}4
\]

\[
\text{logit}\ P(x_{t,7} \mid x_{t,2}, x_{t,3}, x_{t,4}, x_{t,5}, X_{t-1}, \psi_0) = \beta_0 + \beta_2 x_{t,2} + \beta_3 x_{t-1,3} + \beta_4 x_{t-1,4} + \beta_5 x_{t-1,5} + \beta_6 x_{t-1,6} + \beta_7 x_{t-1,7} + \beta_8 x_{t,10} + \beta_9 x_{t,11} + \beta_{10} x_{t,4} + \beta_{11} x_{t,4} \equiv \text{CUM}5
\]

Priors for \( \psi_5, \psi_5, \psi_6 \) are assumed to be independent and identically distributed normal with 0 mean and large variance. The full conditional distributions take the form:

\[
f(\psi_5 | X) \propto e^{-\frac{1}{2} \sum_{i=0}^{2} \frac{\psi_i}{\psi_i}} \prod \prod \prod \left( \frac{\text{CUM}^4_{11} \psi_{t+1,2} \text{CUM}^4_{11} \psi_{t+1,4}}{1+e^{\text{CUM}^4_{11} \psi_{t+1,4}}} \right) \]

\[
f(\psi_5 | X) \propto e^{-\frac{1}{2} \sum_{i=0}^{2} \frac{\psi_i}{\psi_i}} \prod \prod \prod \left( \frac{\text{CUM}^4_{11} \psi_{t+1,5} \text{CUM}^4_{11} \psi_{t+1,6}}{1+e^{\text{CUM}^4_{11} \psi_{t+1,6}}} \right) \]

\[
f(\psi_6 | X) \propto e^{-\frac{1}{2} \sum_{i=0}^{2} \frac{\psi_i}{\psi_i}} \prod \prod \prod \left( \frac{\text{CUM}^4_{11} \psi_{t+1,6} \text{CUM}^4_{11} \psi_{t+1,7}}{1+e^{\text{CUM}^4_{11} \psi_{t+1,7}}} \right) \]

Order of conditioning:

Since covariates enter the model sequentially, the order of conditioning may change the inferences. Ibrahim et al. (2002) tested the sensitivity to ordering and concluded that posterior inferences of \( \psi \) are generally quite robust to changes. However, they still recommend trying different orders. The analysis in this paper used the following order: \([X_5, X_4, X_2, X_3, X_4, X_6, X_7, X_2, X_5, X_4, X_5] \) for convenience. There are five other possible orderings, one example being \([X_5, X_2, X_3, X_4, X_2, X_6, X_7, X_2, X_5, X_4, X_5] \). Because of time limitations, sensitivity analysis are left as a future work.
5 GENERAL CONCLUSIONS

Longitudinal data is collected to measure the change over time, and is common in most scientific fields. We proposed methods that improve the understanding of the complex structure of longitudinal data. Specifically, we proposed the linked brushing to explore the mean trend as a visualization technique. For modeling, we introduced marginalized transitional random effects model for multivariate binary longitudinal data. Although many datasets were introduced to illustrate the exploratory data analysis, our main focus was on two datasets: Iowa Youth and Families Project (IYFP) and The Panel Study of Income Dynamics (PSID). However, our main purpose was to use these datasets to illustrate our proposed methods, and answer questions specific to the subjects in these studies. Because of the limitations of these datasets, our intention was not to draw conclusions about general population or on subject matters.

5.1 Exploratory data analysis

When carefully crafted, exploratory data analysis can answer many questions and reveal many important features of data. The complex structure of longitudinal data with covariates begs for new visual methods that enable interactive exploration. Graphics can discover patterns missed by modeling, errors in data collection and reconstruction.

Longitudinal data comes in different forms. In Chapter 2, we introduced three different types of longitudinal data and supplied many real life examples for each. Some questions of interest and possible methods to answer them through visualization methods were discussed. We pointed out that default layouts in software or slicing the dimension in an obvious way may not be the best option, and might be misleading if not carefully considered. The most important part of exploration is to know what questions to ask, and to look at the data from different perspectives.
Our cataloging highlighted some limitations of existing software to explore longitudinal data. Such limitations include the lack of an environment that can easily integrate continuous data with the categorical one; and lack of time trace brushing by variable. Our suggestions for software in this field include enabling easy brushing of time traces, sorting by brushed percentages, 45% banking, and easy to construct smooth curves.

In Chapter 3, the linked brushing approach to explore the mean trend in both univariate and multivariate longitudinal data was discussed. Smooth curves or mean edges were constructed to display the mean trend over time. Interactive tools, such as, linked brushing and identification, combined with smooth curves prove to capture interesting patterns, and do not require any parametric assumptions. It is possible to reveal the unexpected, to explore the interaction between responses and covariates, and understand structure in multiple dimensions using these methods. The only limitation is that smoothing requires having "enough" time points, and calculating mean trends require having "enough" cases at each time point.

For the PSID data, we discovered an interaction effect between gender and education, which was missed by modeling. In the IYFP data, we focused on significant gender differences on distress. We explored this difference over 10 years in high dimension. Females are more likely to be depressed; once males admit feeling anxious, hostile or depressed, they are more likely to report hostility compared to depression. Our research also pointed out the age differences on distress; all distress measures; anxiety, hostility, and depression; decrease over time. Mood changes were more frequent among teenagers before 12th grade. Once they graduate from high school, flatter mean trends observed.

5.2 Models

While models for continuous responses are frequently proposed because of the flexibility of normal distribution assumptions, modeling longitudinal data containing binary response can be more challenging. In the literature, there is still little research on models for multivariate binary data.

In univariate longitudinal data, the population mean and/or temporal dependence might be of interest. When multiple responses are recorded, in addition to these two features, correlation between responses needs to be taken into account. We proposed a three-level model for multivariate longitudinal
binary data, MTREM(p), that accounts for all these three features of the data.

We proposed a Bayesian implementation of our model using MCMC methods to sample from the posterior distribution of parameters. "Fitting" the MTREM(p) is computationally intensive. The only limitation with MTREM model is that it requires observations to be taken at equally spaced time points. This is due to the use of AR models at the second level of model.

5.3 Relationship between exploratory data analysis and models

One of the main purposes of this thesis was to use visualization methods and models together to obtain a better understanding of complex structures. Although statisticians agree that both exploratory and confirmatory methods should be used hand in hand, there has not been significant amount of work in this area. Tukey's paper titled "We need both exploratory and confirmatory" is one of the few that attracts attention. The main motto of this thesis is driven from this paper, in which, Tukey mentions "Neither is sufficient alone. To try to replace either by the other is madness."

In this thesis, graphics are used before, during, and after model fitting with the following purposes:

- Using graphics before fitting model for:
  - Detecting errors in data recording (eg. Figure 2.22)
  - Detecting errors in data reconstruction (eg. Figure 2.14)
  - Exploring missing data structure (eg. Figures 2.24 and 2.25)
  - Understanding the complex structure of models (eg. Figure 4.1)

- Using graphics during fitting model for:
  - Model checking and making appropriate updates (see below for explanation)

- Using graphics after model fitting for:
  - Catching what model missed (eg. Figure 2.23)
The updates we made in models after looking at graphics should not to be confused with data snooping (Neter, et al. 1996, page 724-725). When graphics pointed out high correlations between and/or within parameters, some simulations were rerun with some adjustments: to avoid correlation between parameters covariates were centered, and to avoid within correlation only 1 in every 5th sample is collected. Moreover, posterior checks in graphics form highlighted that \( p=1 \) model was not a good model fit for our data, and extension to \( p=2 \) is necessary. That is, graphics suggested that regressing current response on a longer history data might improve model fit.

Comparison of findings: EDA and Models

We should first emphasize that direct comparison between two approaches is limited here due to two reasons. First of all, our models use dichotomized responses, whereas our EDA approaches use numerical ones. Second, use of AR models on the second level of our model forced us to use only the first four years of the IYFP data (1989-1992), while the flexibility of EDA approaches enabled us to use all available time points.

Both approaches were able to capture some of the interesting patterns in the data. The probability of targets feeling distressed slightly decreases over time in the first four years. After 7 th grade, there is a gender difference in distress levels; females tend to report more distress especially depression. There is a large subject to subject variation.

However, some benefits of EDA approaches lead us to recover undetected patterns by modeling. Since we were able to use data from all 8 waves, we were able to observe the dramatic drop in distress levels right after high school graduation. Bivariate response displays also revealed interesting patterns. For instance, for females, depression scores are likely to be higher than their hostility scores, and exactly opposite is observed for males. In the later years of high school, mean depression jumps relative to both anxiety and hostility. Although individual variation was detected by both approaches, detecting interesting individuals was easier with EDA. By linked brushing approach used in EDA, tracking the interesting individual traces were quite easy. However, due to computer disk space limitations, random effect estimates were not saved for all subjects in modeling, which limited the screening of individual traces.
The differences in findings could be a result of a couple of things. First, it could be the confounding effect of other covariates in the model. In our examples, EDA approaches focused mainly on gender differences, while models focused on several covariates that might be important in explaining the variation in distress. Second, as we have already mentioned, data used in two approaches differ in terms of response type, and wave length.

5.4 Future work and possible extensions

- Exploratory Data Analysis:
  - Smooth curves in R-GGobi: The method used to calculate smooth curves right now requires the computation of smooth curve points in R, and appending these points to the xml file to use in GGobi. Each time a different smoothing algorithm is requested, analyst has to recompute smoothing points, and update the xml file. Implementing different smooth algorithms through R-GGobi link is an area of future work. This feature will allow users to experiment and explore different smoothing methods without the need of updating data files.

  - Exploring mean trend for more challenging cases: We illustrated how to explore the mean trend for fully constrained and constrained longitudinal data by conditioning on a time-independent covariate. Applying this method to unconstrained data as well as to time-varying covariates are simple extensions. Finding an interesting example dataset to illustrate these methods under these conditions is still to be investigated. GGobi or R-GGobi can easily handle these analyses.

  - Software development: In the light of discussion in Chapter 2, an interactive visualization environment for longitudinal data would be useful. This might be attempted by R-GGobi. For instance, 45% banking is available in the “lattice” package of R. Moreover, linkage between mosaic plots, which can be constructed in R, and GGobi can be useful to analyze mixed data.

- Models:
Calculating computational times: Implementation of model in chapter 4 is computationally intensive and time consuming. Large number of subjects observed over time with multivariate response and covariates introduce thousands of parameters in generalized linear mixed models. Not only do we need to sample from the posterior distribution of these parameters, but also we need to frequently update the parameters via the induced constraints. The computational time is a function of number of subjects, time points, dimension of responses, and the lag used in the second level of model. Running code under different conditions to calculate this computational time has not been attempted yet. Further exploration of optimal algorithms for sampling from the posterior in these models is an area of future work.

R interface: The programs used to implement MTREM model are specific to our data, in terms of assumptions used in covariate imputing. They are implemented in Fortran77. An R interface that fits MTREM model as a general software would be useful, and will be attempted as a future work. However, such software would have the following drawbacks: 1) since assumptions for missing covariates can range widely, a general software cannot handle imputing covariates with the method we proposed. Therefore, they should be imputed outside this software. 2) Tuning parameters in Hybrid MC require careful attention, which might require a long test and trial period and a patient user.

Allowing random effects to be correlated over time: Current model assumes independent random effect parameters. Introducing temporal dependence in the random effects would be an interesting extension. Specifically, an AR(1) model \( h_{it} = \phi \cdot h_{i,t-1} \) or shrinking the dependence parameter towards AR(1) structure might be attempted.

A second model: We will propose a second marginalized model for bivariate longitudinal binary data. This model is an extension of Fitzmaurice and Laird (1993)'s model. They used a marginal logistic model to explain mean structure and conditional log odds-ratios to model the time dependence in longitudinal binary data. However, since the log odds-ratios considered are conditioned on all observed responses, and not just on the history (previous responses at a given time point), this model does not exploit the longitudinal nature of correlation. Unlike their model, our proposed model conditions the log-odds ratios only on the previous response. Moreover, we propose a computational algorithm that is expected to be
faster than theirs, especially for moderate and large measurement series.

- Bridge between the two:

  - Using graphics to communicate the complex model outputs: Exploratory data analysis would be useful in communicating the output from the model. Especially when statisticians are working with scientists from other disciplines, interpreting and explaining the model outputs with graphics can help in developing a common language.

5.5 References


APPENDIX A. MARKOV CHAIN MONTE CARLO METHODS

The purpose of MCMC methods is to construct a Markov chain that will converge to the stationary target distribution. We used Hybrid Monte Carlo methods to sample from the full conditional distributions within the Gibbs sampling, which is used to sample from the posterior distribution of \((b_{it}, \beta, \beta^*, \alpha_t, \lambda_t, \lambda_t^*, \sigma_t)\). Some computational issues are discussed below.

6.1 Hybrid MC

Hybrid Monte Carlo method is highly beneficial when the full conditionals are in a hard to simulate form. It can be applied as long as the derivatives of the logarithm of full conditionals can be obtained.

In physical applications, \(H(q,p)=E(q)+K(p)\) is the total energy as a sum of kinetic and potential energy. In statistical applications, \(q\) is the parameter vector, and \(p\) is just a computational device. \(E(q) \propto -\log(\text{posterior density})\) and \(K(p) = \frac{1}{2} \sum_{i=1}^{m} p_i^2\), where \(p_i \sim \mathcal{N}(0,1)\) and \(m = \text{dim}(p) = \text{dim}(q)\).

Our goal is to move from one point to another in the parameter space. Since we cannot move continuously, some discretization of the exact dynamics, such as leapfrog, is used. Suppose we are trying to move from \(\tau\) to \(\tau + \epsilon\). While \(\tau\) is the actual time in physical applications, it is artificial in statistical ones.

A single leapfrog iteration at time \(\tau + \epsilon\) starting from time \(\tau\) is:

\[
\dot{p}_{\tau + \frac{\epsilon}{2}}(\tau + \frac{\epsilon}{2}) = \dot{p}_{\tau}(\tau) - \frac{\epsilon}{2} \cdot \frac{\partial E}{\partial q_{\tau}}(\dot{q}(\tau))
\]

\[
\dot{q}_{\tau}(\tau + \epsilon) = \dot{q}_{\tau}(\tau) + \epsilon \cdot \dot{p}_{\tau}(\tau + \frac{\epsilon}{2})
\]
\( \dot{p}_i(\tau + \epsilon) = \dot{p}_i(\tau + \frac{\epsilon}{2}) - \frac{\epsilon}{2} \cdot \frac{\partial \mathcal{L}}{\partial q_i}(q(\tau + \epsilon)) \)

Note that the leapfrog iteration takes half step in \( p_i \) (moves from \( \dot{p}_i(\tau) \) to \( \dot{p}_i(\tau + \frac{\epsilon}{2}) \)), uses this information to take a full step in \( q_i \), and uses this to take another half step in \( p_i \). After the first iteration, the last half step of \( p_i \) is followed immediately by the other half step of \( p_i \) from the next iteration. Therefore, except in the first and last iterations, we can combine these half steps into a full step for \( p_i \), and save some computation time.

If the step size \( \epsilon \) is chosen to be small, the discretized dynamics approximates the exact dynamics better. \( \epsilon \) is recommended to be chosen such that we achieve a 90\% acceptance rate. The above leapfrog iteration is repeated \( K \) times, each with \( \epsilon \) size, before a Metropolis-Hastings accept-reject step is introduced (see algorithm below). Neal (1993) suggested the use of a reasonably large step number \( K \). However, one should note that within each iteration, the gradient \( \left( \frac{\partial \mathcal{L}}{\partial q_i} \right) \) has to be calculated, and large numbers of \( K \) requires more computation time.

We used Hybrid MC for sampling from the full conditionals of parameters in MTREM model. The algorithm is outlined below only for \( \beta \). The other algorithms are similar; except in steps 5 and 8, \( \Delta_{tj} \) and \( \Delta_{tj}' \) values are updated only if the parameter is a function of \( \Delta_{tj} \) and \( \Delta_{tj}' \).

Algorithm for Hybrid MC to sample from the full conditional of \( \beta \)

Given the current values of the other parameters, use the following algorithm to obtain the new state of \( \beta (^{n+1}) \). Let ACPR denote acceptance probability.

Step 1 Generate \( p_i \) \( i = 1, \ldots, \text{dim}(\beta) \) from standard normal distribution.

Step 2 Calculate \( H(q, p) \) by using the parameter values at time 0.

Step 3 Randomly choose a direction \( \lambda \), -1 or 1. Set the step size \( \epsilon = \lambda \cdot \epsilon_0 \), where, \( \epsilon_0 \) is chosen such that the average acceptance rate is around 90\%.

Step 4 Generate the step number, \( K_\beta \).

Step 5 Apply leapfrog iteration \( K_\beta \) times to obtain the candidate state \( (q^*, p^*) \). Update \( \Delta_{tj} \) and \( \Delta_{tj}' \) at each step.

Step 6 Calculate \( H(q^*, p^*) \) and acceptance probability \( \text{ACPR}_{\beta} = \min[\exp(H(q, p) - H(q^*, p^*)), 1] \).
Step 7 Generate $u_\beta$ from Uniform(0,1).

Step 8 If $u_\beta \leq ACP R_\beta$, then accept $(q^*, p^*)$ as the new state. Otherwise, keep the state $(q, p)$ and return back to $\Delta_{i+j}$ and $\Delta^*_{i+j}$ values at time 0.

6.2 Gibbs sampling

Gibbs sampling is a special case of Metropolis-Hastings, where the candidates are always accepted with probability 1. However, unlike Metropolis-Hastings and Hybrid MC, Gibbs sampling requires that the conditional distributions are proportional to a known parametric distribution. It updates the parameter vector one component at a time by sampling from conditional distributions with the goal of generating from the stationary joint posterior distribution.
Data from Iowa Youth and Families, R/S+/SAS script and Fortran77 code, that are used in the analyses of this thesis, are provided in the attached CD-ROM. The use of IYFP data is subject to the permission of Dr. Rand Conger of Institute for Social and Behavioral Research at ISU Research Park. You can obtain the software and languages from following links.

- Software and information about GGobi is available from:
  http://www.ggobi.org/

- R is freely available from:
  http://www.R-project.org

- Bayesian Output Analysis is freely available from:
  http://www.public-health.uiowa.edu/boa

- To use Splus on unix at Iowa State University (ISU), type the following:

  % add splus
  % Splus

- To use SAS on unix at ISU, type the following:

  % add sas
  % sas -work . &

- To use Fortran 77 on unix at ISU, type the following:

  % f77 -o filename filename.f
  % filename
8 ACKNOWLEDGEMENTS

I would like to thank my advisors, Drs. Dianne Cook and Michael Daniels, for their constant assistance and guidance throughout this project. I highly appreciate their patience and generosity with their time and experience, and I can only dream of being the kind of researcher, teacher and guide to my students as they have been to me. I would also like to thank Dr. Philip Dixon for supporting me through another project during the last period of this thesis, for serving in my committee, and for many insightful discussions. My thanks also go to Drs. Frederick Lorenz and Heike Hofmann for their helpful comments, ideas, and for never rejecting my request for help throughout this project. I would like to recognize my other committee member, Dr. Sigurdur Olafsson, for bringing a different perspective to the subject.

I would like to express my appreciation to Dr. Rand Conger for letting me use the data from Iowa Youth and Families Project. This appreciation extends to Dr. Becky Burzette, of Institute for Social and Behavioral Research at ISU Research Park, for graciously helping in data collection and clarifying some issues with the data.

During my stay in the United States, I have not only gained experience in Statistics, but I have also matured socially and culturally. I have learned so much about myself, my culture and other cultures. Most of my graduate studies were supported by a scholarship from the government of Turkey. I would like to thank my country, and my home university, Middle East Technical University of Turkey, for giving me the opportunity to broaden my horizons through their financial support.

I am grateful to have friends, too many to list here, who have been with me throughout this experience. I am especially indebted to one of my best friends, Olcay Zeynep Baran, for seeing me through both the good times and the tough times—by many emails and phone calls—and celebrating my success from the other side of the world for more than five years.
Lastly, I would like to thank my parents, Hatice and Sebahattin Ilk, for their devotion and support, and for making me who I am today. Special thanks goes to my favorite role model, my brother, Dr. Hakki Gokhan Ilk, for his guidance and cheer throughout my life. This thesis is dedicated to my parents, my brother and his newborn son, Yigit.