Title

Handling Incomplete High-Dimensional Multivariate Longitudinal Data with Mixed Data Types by Multiple Imputation Using a Longitudinal Factor Analysis Model

Author

Lu, Xiang

Publication Date

2016

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
Los Angeles

Handling Incomplete High-Dimensional Multivariate Longitudinal Data with Mixed Data Types by Multiple Imputation Using a Longitudinal Factor Analysis Model

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Biostatistics

by
Xiang Lu

2016

ABSTRACT OF THE DISSERTATION

Handling Incomplete High-Dimensional Multivariate Longitudinal Data with Mixed Data Types by Multiple Imputation Using a Longitudinal Factor Analysis Model

by

Xiang Lu

Doctor of Philosophy in Biostatistics

University of California, Los Angeles, 2016

Professor Thomas R. Belin, Chair

We developed an imputation model solving the missing-data problem in a high-dimensional longitudinal data set with mixed data types (continuous and ordinal) based on a factor-analysis and a linear mixed-effect model. Markov Chain Monte Carlo is used to fit the model, drawing parameters, latent variables and missing values iteratively. The imputation model is written in an R package.

We tested the newly developed imputation model using simulated data sets under 32 scenarios and 2 hypothetical missing-data mechanisms. Two competitive models PAN (Multiple

Imputation for Multivariate Panel or Clustered Data) and MICE (Multiple Imputation using Chained Equations) are also tested in the same way for comparison, to show the necessity of addressing the high-dimension and mixed continuous and ordinal data type issues.

Part of the effort we made is to accelerate the simulation using C++ (a low-level language) and the parallel computing by the Hoffman 2 Cluster. Compared to running the simulation evaluation in an R program on one single computer, the program we use for the simulation evaluation runs approximately 600 times faster.

We also tested the robustness of the newly developed imputation model in the cases of violation of assumptions. We found that assuming less than the true number of factors corresponds to invalid inferences, while assuming more than that corresponds to reasonable inferences. We also found that only omitting very strong underlying quadratic trends of the factor scores hurt the inferences based on the imputation. In the most unfavorable scenario we tested, when the underlying quadratic coefficient is as large as .8 of the linear coefficient, the actual coverage rates of 95% interval estimates start falling below 90%.

An application to a dentistry data is shown, in comparison to the PAN, NORM and a fore runner (Wang 2002) of the newly developed method.

The dissertation of Xiang Lu is approved.

Robert Erin Weiss

Mitchell David Wong

Weng Kee Wong

Thomas R. Belin, Committee Chair

University of California, Los Angeles
2016

To my family and friends.

Table of Contents

Vita

Xiang Lu studied computer science in Tsinghua University, Beijing China from 1996 to 2000. After working as a software engineer for few years, he became interested in statistics and came to Biostatistics UCLA for a master degree. Then he started his Ph.D. program there in 2008. His research interest was in Bayesian, missing data problem, latent variable models, and statistical computing.

Chapter 1
Missing Data in Longitudinal Studies

Health-science research often focuses on whether an intervention, such as a drug or another form of treatment, is effective. Cross-sectional data are not ideal for answering such questions. A better approach is to compare the same subject over two or more time points, which introduces longitudinal data. A comparison of the difference in outcomes before and after exposure to an intervention can provide a foundation for inferences about intervention effects. Additional measurements over time can enhance the ability to draw inferences and to provide better scientific knowledge.

High-dimensional data sets naturally arise in practice, partly as a consequence of the rapid development of information technology. Modern data sets could have dimensions hundreds of times larger than typical data sets from a few decades ago, for equal or lower cost. More importantly, when we recruit a group of patients for a study in health sciences, it is natural to want to measure as much as we can about them. Hence, statisticians see high-dimensional longitudinal data sets on a routine basis in research settings.

Suppose we have decided to collect high-dimensional longitudinal data in a study. Missingness can occur naturally, whether due to non-response to survey items, a missed appointment for scheduled follow-up, or an investigator's decision not to measure all variables at

all time points to reduce cost or response burden. Challenges can be expected in the analysis of incomplete longitudinal data since missing values cannot be assumed to arise completely at random when the reasons for missingness are outside the investigators' control. Except for the case when data are missing by design, a researcher needs to be aware of the possibility that naturally occurring missing-data mechanisms could compromise the validity of naive analyses based on estimation strategies such as complete-case analysis, available-case analysis, and imputation of the last-observation carried forward. For example, if some patients in the treatment group of a trial drop out due to non-effectiveness of a drug, this missing-data mechanism related to self-selection could generate false positive findings or could otherwise produce misleading information. Although the actual missing-data mechanism that arises in a real study may not be so extreme, it is worthwhile for researchers to avoid analysis errors by making use of available data that has the potential to predict the values that are missing for incomplete cases.

1.1 Motivating Example

In comparing the effects of maxillomandibular fixation (MMF) and rigid internal fixation (RIF), two treatments for mandible fractures (i.e., broken jaws), 336 patients were enrolled in a study, 142 of whom were diagnosed as having moderately severe mandible fractures and were randomized into receiving either MMF or RIF. Patients were followed up for 12 months prospectively. A variety of clinician and patient-reported measurements, which are listed in Table 1.1, were used to assess the effects of the treatments at discharge from the hospital and at follow-up visits at 10 days, 1 month, 6 months, and 12 months post-discharge. Exploratory analysis shows that the transformation $t = \log(d/12 + 1)$, where d refers to the number of days

since discharge, makes it more reasonable to assume a linear association between time and the main outcome variables.

Table 1.1 Outcome variables

PAIN	Patient self reported pain	integer 1 - 3, higher means more pain
PTHEALTH	Patient self reported health level	integer 1-5, higher means more healthy
GOHAI	GOHAI, a composite score for General Oral Health Assesment Index	integer 0 - 100, higher means better
MHI5	Mental Health Index, a composite score of five items	integer 0 - 100, higher means more issue
LNSST	Logarithm of the total number of social support	log of count number where the count number can be 100+ in the data
BSI_DEP	Brief Symptom Inventory for depression	continuous 0 - 4, same for all BSI's
BSI_OCD	Brief Symptom Inventory for obsessive-composite disorder	
BSI_ANX	Brief Symptom Inventory for anxiety	
BSI_HOS	Brief Symptom Inventory for hostility	
BSI_PHO	Brief Symptom Inventory for phobia	
BSI_ADD	Brief Symptom Inventory for attention deficit disorder	
N_COMP	Number of patient complaint	count number 0 up to 6 in the data
PTSDSUM	A summary of post trauma stress disorder	integer 0 – 100
SUM_LEW	A summary of weighted life events reflecting patient's potential stress	continuous 0 - 31 in the data
PTCMPL	Patient compliance	dichotomous 0=not complied, 1=complied

The General Oral Health Assessment Index (GOHAI), a quality-of-life measure, is one of two main outcomes of interest. GOHAI takes on integer values in the range from 0 to 100, but is

treated here for analysis purposes as a continuous variable. We used a linear model with different intercepts and slopes for the time effect in the MMF and RIF treatment groups, modeled with random intercept and slopes across time. Formally,

$$Y_{ij} = \beta_0 + \beta_{0i} + \beta_{0(RIF-MMF)} \cdot \mathrm{I}(i \text{ in RIF}) + (\beta_1 + \beta_{1i} + \beta_{1(RIF-MMF)} \cdot \mathrm{I}(i \text{ in RIF}))t_{ij} + \varepsilon_{ij}$$

where i indexes subjects, j indexes visits; Y_{ij} is GOHAI of subject i at visit j; β_0 and β_1 are intercept and slope of group MMF; $\beta_{0(RIF-MMF)}$ and $\beta_{1(RIF-MMF)}$ are the differences in the intercepts and slopes between the two treatment groups; β_{0i} and β_{1i} are random effects for the intercept and slope of subject i with an unstructured conditional variance covariance assumption; and ε_{ij} is assumed to be normally distributed error term with same variance in the two groups.

The self-reported pain level for patients (PAIN), originally recorded as having a value among the integers from 0 to 10, was the other main outcome variable. It is common to analyze such a variable as a continuous variable. However, invoking a normality assumption might be questionable. Considering the skewness in the PAIN data and our inability to find a power transformation to achieve symmetry, we collapsed the PAIN variable into 3 levels based on the modes we find from a histogram: low pain, which corresponds to values of 0, 1, or 2 on the original scale; medium pain, corresponding to values in the range from 3 to 7; and high pain, corresponding to values of 8, 9, or 10. We re-code these categories as 1 (low), 2 (medium) and 3 (high). In the analysis, we used a cumulative logistic model with different intercepts and the same slopes in the MMF and RIF treatment groups, with random intercept only. Formally

$$\text{logit}(\mathrm{P}(\mathrm{U}_{\mathrm{ij}} \geq \mathrm{c})) = \beta_{0c} + \beta_{0i} + \beta_{0(RIF-MMF)} \cdot \mathrm{I}(i \text{ in RIF}) + \beta_1 t_{ij}$$

where $c = 2, 3$ refers to the outcome category; i indexes subjects; j indexes visits; U_{ij} is the latent continuous pain level of subject i at visit j assumed in an ordinal logistic model, which is valued in $(-\infty, \infty)$; and other terms are similar as in the model for GOHAI.

Besides the outcome measurements listed in Table 1.1, an indicator for treatment group (MMF vs RIF), time from discharge, and several time-invariant covariates are used in the imputation model, including age group (18-34, 35-44,44+), race (Black, Hispanic, other), education (less than high school, high school and beyond), gender, marital status (married, single, widowed/divorced), previous trauma, alcohol use, drug use, maximum opening of the injury, and severity of mandible fracture (mild, moderate, severe). We consider analysis models involving GOHAI and PAIN that use treatment group and time as predictors. As noted by Collins, Schafer and Kam (2001), auxiliary variables not included in the analysis can still help add precision when producing imputations.

1.2 Theoretical Background in Handling Missing Data

Rubin (1978) developed the multiple imputation framework, a strategy for addressing missing data problems building on a Bayesian perspective in a way that is shown to have good statistical properties in frequency-based evaluations. In multiple imputation, we predict (impute) missing values given observed values multiple times. Each imputed data set is analyzed in the same way as if the data were completely observed. Average values of estimated parameters coupled with an estimate of the total variance that combines between-imputation variability with average within-imputation variability are used to form the inference based on the data with missing values.

One strategy for producing valid point estimates and dispersion estimatesis to produce multiple imputations based on a joint model for the data. For general-purpose procedures, a common starting point is to assume that the missing data are missing at random, or MAR (Rubin 1976), an assumption that will not always be defensible. However, a model involving more variables generally leads to better prediction since ignoring important predictors may give rise to predictable biases (Rubin 1996; Collins, Schafer and Kam 2001). In addition to the potential for greater precision with more predictors, having more predictors may transform what would have been not missing at random (NMAR), into an MAR scenario (Schafer 1997). Hence we prefer to use an imputation model involving more variables, leading to the necessity of modeling a high-dimensional mixture of both continuous and categorical variables.

1.3 Multiple Imputation Based on Longitudinal Factor Analysis and Probit Model

Factor analysis is one way of handling high-dimensional data to provide useful information for scientific research. It is the basis for many scoring systems in health sciences, e.g. the General Oral Health Assessment Index, or GOHAI (Atchison and Dolan 1990), various subscales of the RAND 36-item short-form quality-of-life questionnaire, or SF-36 (Ware and Sherbourne 1992) and the Children's Depression Inventory, or CDI (Saylor, Finch, Baskin, Saylor, Darnell and Furey 1984). Using algorithms such as principal-component analysis or maximum likelihood, latent variables can be identified to summarize certain properties of subjects into factor scores. Including factor scores in a model, either as an outcome or as a covariate, is a common helpful practice.

In this project, we are interested in recovering information from incomplete longitudinal high-dimensional data by introducing latent variables in a factor-analysis model that encompasses a large number of measured quantities. The possibility of a longitudinal trend in the factor scores is modeled by a random-effect linear model.

Another challenge in this project is to incorporate an imputation strategy for categorical variables. In some settings, categorical variables have been treated the same way as continuous variables in a regression model, where the imputed values are rounded to the nearest possible value. This strategy allows the categorical values to contribute their numerical values to the regression, which does pick up some association among variables. However, extreme categories sometimes signal a wider range of variability than can be well modeled with normal distribution. This drawback can contribute to inadequate variability in the imputed values with considerable over-representation of the middle-level categories, as observed by Belin, Hu, Young and Grusky (1999). Hence, we propose imputing ordinal categorical variables based on a probit model that is integrated into a factor-analysis framework that incorporates both the observed continuous outcome variables and latent continuous variables underlying ordinal categorical measures.

Chapter 2
Literature Review

In this chapter, we describe the basic theory for handling missing data and we summarize earlier attempts to model continuous and ordinal variables together. The proposed approach in this dissertation for handling missing data can be viewed as drawing on elements from the literature and putting them together in a novel way to go beyond the capabilities of existing methods.

2.1 Modeling Assumptions for Incomplete Multivariate Data

In general, a data set can be presented in an $n \times p$ matrix form, with each of the n rows representing an individual case and each of the p columns representing a variable. Missing values can occur anywhere in a data set. A schematic representation of a data set with missing values is shown in Figure 2.1, where missing values are denoted by question marks.

	Variable							
Unit	V1	V2	...	...	...	...	...	Vp
1					?			
2	?							
3								
.								
.		?						
.								
.							?	
n			?					

Figure 2.1 An incomplete multivariate data set

Missing data may occur in many different fashions. A helpful classification of missing-data mechanisms is given by Rubin (1976). Let Y represent the underlying complete data that **would** have been observed if all values were not missing, which can be partitioned into the actually observed part, Y_{obs}, and the actually missing part, Y_{mis}. Let M be an $n \times p$ matrix of indicator of variables, where each element equals to 1 if the corresponding value of Y is missing, 0 otherwise.

	Variable							
Unit	V1	V2	...	...	...	...	...	Vp
1					1			
2	1							
3								
.								
.		1						
.								
.							1	
n			1					

Figure 2.2 A missing indicator matrix, which is a dichotomous matrix valued in 0 and 1 (0's are not shown).

The joint distribution of complete data Y with missingness indicators M, given parameters (θ, ξ) can be written as

$$\mathrm{P}(Y, M \mid \theta, \xi) = \mathrm{P}(Y \mid \theta)\mathrm{P}(M \mid Y, \xi),$$

where θ is the parameter of the distribution of data Y and ξ is the parameter of missing mechanism. The marginal distribution of the observed data (Y_{obs}, M) can be obtained by integrating the above joint distribution over the missing values:

$$\mathrm{P}(Y_{obs}, M \mid \theta, \xi) = \int \mathrm{P}(Y_{obs}, Y_{mis} \mid \theta)\mathrm{P}(M \mid Y_{obs}, Y_{mis}, \xi) dY_{mis}.$$

Missing data are said to be missing completely at random (**MCAR**) if the missing data mechanism is a distribution that does not depend on the data at all, i.e.

$$\mathrm{P}(M \mid Y_{obs}, Y_{mis}, \xi) = \mathrm{P}(M \mid \xi).$$

On the other hand, we say data are missing at random (**MAR**) if the missing data mechanism is a distribution that does not depend on the missing values given the observed values, i.e.

$$\mathrm{P}(M \mid Y_{obs}, Y_{mis}, \xi) = \mathrm{P}(M \mid Y_{obs}, \xi).$$

In other words, MAR corresponds to the missing indicator being conditionally independent from the missing values given observed data. Note that MAR is related to the set of variables of the data set in consideration and can be affected by the inclusion of auxiliary variables (Schafer 1997; David, Little, Samuhel and Triest 1986). It is possible that the MAR assumption could be valid in a data set containing, say, 10 variables, but that it might not hold any more when we drop a number of variables from the data set. In practice, it is not always safe to assume MAR, but because it is a good starting point for many analyses, we will make use of an MAR framework in developing a longitudinal factor model.

Under MAR, the likelihood of the observed data (Y_{obs}, M) can be simplified as

$$\begin{aligned} \mathrm{P}(Y_{obs}, M \mid \theta, \xi) &= \int \mathrm{P}(Y_{obs}, Y_{mis} \mid \theta)\mathrm{P}(M \mid Y_{obs}, Y_{mis}, \xi) dY_{mis} \\ &= \mathrm{P}(M \mid Y_{obs}, \xi) \int \mathrm{P}(Y_{obs}, Y_{mis} \mid \theta) dY_{mis} \\ &= \mathrm{P}(M \mid Y_{obs}, \xi)\mathrm{P}(Y_{obs} \mid \theta) \end{aligned}$$

If we further assume that the parameter θ of the data model and the parameter ξ of the missing data mechanism are distinct, i.e. they are not the same parameter and do not share any components, then inference for the model parameter θ can be based only on the observed-data likelihood $\mathrm{P}(Y_{obs} \mid \theta)$. In this case, the missing-data mechanism is said to be ignorable. Ignorability includes two assumptions, missing at random and distinct parameters. Since we can safely assume distinct parameters in most cases, ignorability can be thought of as roughly equivalent to missing at random.

By using the definition of conditional probability, MAR is equivalent to the identity

$$\frac{\mathrm{P}(M, Y_{obs}, Y_{mis})}{\mathrm{P}(Y_{obs}, Y_{mis})} = \frac{\mathrm{P}(M, Y_{obs})}{\mathrm{P}(Y_{obs})},$$

which is simplified into

$$\mathrm{P}(Y_{mis} \mid M, Y_{obs}) = \mathrm{P}(Y_{mis} \mid Y_{obs})$$

as we expect from conditional independence of M and Y_{mis}. Hence, assuming MAR, we can generate multiple complete data sets in which Y_{mis} is identically distributed as $\mathrm{P}(Y_{mis} \mid M, Y_{obs})$ by drawing from the distribution $\mathrm{P}(Y_{mis} \mid Y_{obs})$.

When the missing-data mechanism is not MAR (and hence not MCAR, since MCAR is a special case of MAR), the distribution of missing-data indicators depends on unobserved values even conditioning on the observed values. Sometimes we call this scenario not missing at random (NMAR), in which case the missing-data mechanism is nonignorable, i.e. likelihood-

based or Bayesian inferences for parameters of the data require modeling of the missing-data mechanism.

2.2 Early Efforts for Handling Missing Data

One way to analyze incomplete data is through complete-case analysis, where cases with any missing values are simply deleted. One problem of this method is the inefficiency due to data being dropped. Furthermore, the accuracy of this way is often questionable. Complete-case analysis works well in scenarios that are "close to" MCAR in the sense that the bias generated is negligible. However, MCAR rarely happens in reality unless missing data are missing by design; when human decisions contribute to missingness, one can expect systematic patterns to emerge related to missingness.

A minor variation on complete-case analysis is available-case analysis, in which all available data are used to estimate given parameters. Little and Rubin (2002) show that available-case analysis can suffer from problems with incompatibility, as when the values used to estimate variances are not the same as the values used to estimate covariances, since the cases where variables are jointly observed might be a subset of the cases where separate variables are observed.

Early efforts aiming at handling missing data also include single imputation, e.g. mean substitution in which the missing values are replaced by the mean (or conditional mean) of the observed values. Mean substitution ensures every variable is imputed within its plausible range. However, it can generate biases for not considering the missing-data mechanism; it also tends to

underestimate the uncertainty due to missing data, since the uncertainty due to missingness is not counted.

To make statistically sound analysis of data with missing values, Rubin (1978, 1987) developed the multiple imputation method, which is sketched in the next section.

2.3 Multiple Imputation

Imputation involves filling in values for missing items, so that a complete data set is obtained on which one can carry out a standard complete-data analysis. However, the uncertainty associated with the missing values is not reflected in the imputation if we obtain only one such complete data set. This lack of uncertainty in the imputed data set usually leads to over-estimation of precision of statistical analysis, implying smaller error components and narrower confidence intervals than appropriate, where the actual level (false positive rate) of statistical tests will not in general be equivalent to the nominal level.

Rubin (1978, 1987) developed the multiple imputation (MI) technique to reflect the uncertainty in imputed values while retaining the advantage of allowing existing complete-data analysis methods to be applied to data sets with missing values. In MI, a data set with missing values is imputed to form $m \geq 2$ complete data sets. A generally valid procedure for producing imputations is to replace the missing values by an independent sample from their posterior predictive distribution. Variation among the m imputations reflects the uncertainty of the missing values that is predicted from the observed data. The resulting m versions of the complete data sets are then analyzed by existing statistical methods. After performing the identical analysis on each of those data sets, the results are combined using the rules developed by Rubin and

Schenker (1986) to produce an overall statistical inference incorporating missing-data uncertainty.

Specifically, suppose $\widehat{Q_j}$ is an estimate of a scalar parameter of interest Q obtained from imputed data set j where $j = 1 \ldots m$, and $\widehat{U_j}$ is the estimated variance of $\widehat{Q_j}$. Then the overall estimate of Q is

$$\overline{Q} = \sum_{j=1}^{m} \frac{\widehat{Q_j}}{m},$$

and the overall estimated variance is

$$T = \overline{U} + (1 + \frac{1}{m})B,$$

where

$$\overline{U} = \sum_{j=1}^{m} \frac{\widehat{U_j}}{m}$$

is the average within-imputation variance, and

$$B = \frac{1}{m-1} \sum_{j=1}^{m} (\widehat{Q_j} - \overline{Q})^2$$

is the between-imputation variance. The distribution of $(\overline{Q} - Q) \cdot T^{-1/2}$ has an approximate t-distribution with degrees of freedom

$$v = (m-1)(1 + \frac{1}{r})^2,$$

Where *r* represents the relative increase in variance due to non-response

$$r = (1 + m^{-1})\frac{B}{\bar{U}}.$$

Inference for a multi-dimensional parameter can be handled by multivariate generalizations of the above rules (Li, Raghunathan and Rubin 1991; Li, Meng, Raghunathan and Rubin 1991).

Rubin (1987) shows that only a small number of imputations are needed to achieve a relatively efficient estimate in most cases. The relative efficiency reflecting the ratio of the estimated precision based on *m* imputations to that based on an infinite number of imputations can be approximated by the quantity $(1 + \lambda/m)^{-1}$, where λ is the fraction of missing information, estimated by

$$\hat{\lambda} = \frac{r + 2/(v+3)}{r+1}.$$

For example, for a data set with 30% missing information ($\lambda = .3$), an estimate based on *m*=5 imputations will have relative efficiency 94.4%.

Three general strategies can be applied to produce multiple imputations. In a joint modeling strategy, one develops a joint model for multivariate data and bases the draws for missing values on the implied conditional distributions. Such an approach is often implemented through the use of a Markov chain Monte Carlo (MCMC) procedure (Cassella, Robert and Wells 2004).

An alternative approach is the sequential regression strategy, e.g. ICE/MICE (Van Buuren and Oudshoorn 1999) and IVEWare (Raghunathan, Lepkowski, Van Hoewyk and Solenberger 2001), which is motivated by an analogy with MCMC methods. This idea involves specifying a set of conditional distributions of outcome variables with missing values given all other variables. Theoretically, the collection of conditional distributions might not compatible, in the sense that the sequence would not converge to any joint distribution (Gelman and Speed 1993; Hobert and Casella 1998; Liu, Gelman, Hill, Su and Kropko 2013). However, this method embraces an approximation at the modeling stage for the sake of flexibility and simplicity and has been seen to work reasonably well in applications. (e.g., Raghunathan et al. 2001)

Another category of imputation method is the implicit-model strategy, e.g. hot-deck imputation (Rubin 1976) and predictive-mean mean matching (Heitjan and Little 1991, Schenker and Taylor 1996). The idea is to borrow imputed values that are observed on other cases sharing some relevant properties, like coming from the same batch or having a similar predicted mean given multiple covariates. Siddique and Belin (2008) described a distance-based donor selection approach with an approximate Bayesian bootstrap where donors are selected with probability inversely proportional to their distance from the donee. A related SAS macro MIDAS has been developed by Siddique and Harel (2009) for multiple imputation using distance-aided selection of donors.

2.4 Existing Methods Related to Multiple Imputation for High-Dimensional Data with Mixed Data Types

The goal of multiple imputation is to predict missing values given observed values and missing indicators. Hence any statistical method that supports prediction of outcome variables can potentially be the basis of a strategy for imputation.

2.4.1 Multivariate Normal Imputation

One general strategy of dealing with missing values is to produce imputations under a completely specified joint model. A multivariate normal model, which implies that conditional distributions are governed by a linear regression relationship, is widely used.

The approach has been implemented in Schafer's NORM program (Schafer 1997), which is also available as an R package. NORM assumes that the variable $Y_i \overset{iid}{\sim} N(\underset{p\times 1}{\mu}, \underset{p\times p}{\Sigma})$, but each Y_i may have some components missing. To facilitate multiple imputation using NORM, a preprocessing step of finding the maximum likelihood estimate (MLE) of (μ,Σ) using E-M algorithm can be helpful to orient the needed statistical computing steps. Multiple imputations can then be produced by first drawing Markov Chain Monte Carlo (MCMC) samples using a data augmentation (Tanner and Wong 1987) procedure starting from the MLE and then selecting *m* values, typically after long lags in the iterative simulation steps to avoid autocorrelation, to serve as imputations. These draws are then plugged into the original incomplete data set to form m complete data sets. A viable strategy is to generate a long chain and to allow the m draws to be equally spaced with large gaps so they can be expected to have minimal auto-regression.

Standard practice includes discarding the initial part of chain as "burn-in" iterations to allow the Markov-Chain sequence to approach convergence in distribution.

2.4.2 Imputation Based on Mixed-Effect Linear Model

For use with longitudinal or panel-study data, Schafer (1997) proposed the PAN model for imputing high dimensional data with missing values using a mixed-effect linear model. Suppose we have r outcome variables $Y_1 \ldots Y_r$, all of which may contain missing values. Also suppose we have I units (e.g. subject, clusters) indexed by $i = 1, 2, \ldots I$, each of which contributes n_i repeated measure of the outcome variables $Y_1 \ldots Y_r$. Consider design variables $X_1 \ldots X_p$ and $Z_1 \ldots Z_q$ for $q \le p$, which correspond to the fixed effects and random effects. The model can be written as

$$\underset{n_i \times r}{y_i} = \underset{n_i \times p}{x_i}\ \underset{p \times r}{\beta} + \underset{n_i \times q}{z_i}\ \underset{q \times r}{b_i} + \varepsilon_i$$

where y_i, x_i and z_i are sub-matrices of Y, X and Z corresponding to subject i, ε_i represents multivariate normal errors, β is the matrix of fixed coefficients and b_i is the matrix of random coefficients with variance-covariance matrix possibly assumed to be $\underset{q \times q}{\text{diag}}(\underset{q \times q}{R_1} \ldots \underset{q \times q}{R_r})$, $\mathrm{I}_r \otimes R_{q \times q}$ or $R_{qr \times qr}$ depending on the generality we allow in the model.

With a prior distribution containing very little information, which is often used in the absence of more precise information about parameters, this model is capable of imputing high-dimensional missing data. Except for the design variables X and Z, this model allows missing

data to appear in any variable in a study. Accordingly, it can also address the problem of missing covariates of an analysis model by treating them as outcomes in the imputation model.

2.4.3 Imputation Based on Factor Analysis Model

Song and Belin (2004) developed a multiple imputation model based on common factor analysis to overcome the problem of over-parameterization with multivariate data. Rubin (1996) suggests that, when carrying out multiple imputation, one should consider as many variables as possible to account for patterns of association in the data; hence, the number of parameters in the imputation model could be very large. But given that the number of subjects (sample size) in any study will be limited, imputation-model parameters could be poorly estimated or inestimable.

Another motivation for factor analysis arises when several variables are highly collinear. In this situation, the sample covariance matrix can become poorly conditioned or hard to invert, which results in difficulty in estimation. Invoking a ridge prior (Schafer 1997) is one possible solution, while factor analysis is a competing approach.

In the factor analysis model, the parameters include the factor loading matrix and variance of errors. Compared to multivariate normal imputation, which estimates the covariance of every possible pair of variables, the factor analysis model economizes on the number of parameters by making use of the factor loading matrix to simplify the overall covariance matrix.

2.4.4 Imputation based on General Location Model

In situations featuring a mixture of categorical and continuous data, one joint-modeling idea is the general location model, originally developed by Olkin and Tate (1961). Categorical variables $W_1 \ldots W_q$ are treated as one single categorical variable W with $M = \prod_{k=1}^{q} I_k$ categories, where I_k is the number of categories of variable W_k. Given $W = m$ where , m is in $1 \ldots M$, the multivariate continuous variable Y has distribution $N(\mu_m, \Sigma)$, where vector μ_m is mean for subjects of m^{th} category, while Σ is a covariance matrix shared among all categories of subjects.

Little and Schluchter (1985) refined this model by adopting a log-linear model for the categorical portion of the data and a regression relationship between continuous variables and categorical variables. Specifically, categorical variables $W_1 \ldots W_q$ are assumed to have cell probabilities consistent with a log-linear model having main effects and interactions associated with the categorical variables. Given a value of W (a q-vector), we can write $Y \sim N(\Gamma W, \Sigma)$ where Y is p-dimensional continuous variable, Γ is $p \times q$ coefficient matrix, where Σ is still a common covariance matrix of errors. Γ could further be restricted by $\Gamma = BA$ where A is a known matrix of design and B is the set of unknown parameters.

Belin, Hu, Young and Grusky (1999) evaluated the performance of imputation using an adaption of Little and Schluchter's refined model developed by Schafer (1997). The method was applied to a mental health study assuming ignorable missingness. The comparison between predictions of missing values and actual follow-up measurements showed statistically significant differences, especially for binary proportions and ordinal variables. The main issue for the binary variables came from the common Σ assumption of the general location model, as the implied

linear discriminant for predicting binary variables gave rise to predictions with unrealistically small amounts of binomial uncertainty. Also regarding ordinal variables treated as continuous when the underlying distribution is actually skewed, imposing a bell-curve assumption produced too many imputations close to the mean of the distribution, which can bias subsequent estimates of quantities that depend on such values.

2.4.5 Multiple Imputation by Chained Equations

With high-dimensional data, another possible strategy is to produce imputations one variable at a time using a series of overlapping regression relationships, such as implemented in the MICE (Multiple Imputation by Chained Equations) package developed by Van Buuren and Oudshoorn (1999). The idea can be described as follows.

Let data be $Y_{n\times k} = (Y_1, Y_2, \ldots, Y_k)$, a set of k random variables. Each variable may be partially observed, say, $Y_j = (Y_j^{obs}, Y_j^{mis})$ where $j = 1 \ldots k$. Let $Y^{obs} = (Y_1^{obs}, \ldots, Y_k^{obs})$ and $Y^{mis} = (Y_1^{mis}, \ldots, Y_k^{mis})$. The imputation problem is to draw a sample of Y, or a sample from $P(Y^{mis} \mid Y^{obs})$ and combined with Y^{obs}.

Denote $Y_{-j} = (Y_1, \ldots, Y_{j-1}, Y_{j+1}, \ldots Y_k)$, i.e. all variables except Y_j. Assuming a series of conditional distributions

$$\begin{matrix} P(Y_1 \mid Y_{-1}, \theta_1) \\ \cdots \\ P(Y_k \mid Y_{-k}, \theta_k) \end{matrix},$$

we could then mimic the process of obtaining a sample of $P(Y^{mis} \mid Y^{obs})$ using a Gibbs sampler:

$$\begin{aligned}
&\theta_1^{(t)} \sim P(\theta_1 \mid Y_1^{obs}, Y_2^{(t-1)}, \ldots, Y_k^{(t-1)}) \\
&Y_1^{mis(t)} \sim P(Y_1^{mis} \mid Y_1^{obs}, Y_2^{(t-1)}, \ldots, Y_k^{(t-1)}, \theta^{(t)}) \\
&\ldots \\
&\theta_k^{(t)} \sim P(\theta_k \mid Y_k^{obs}, Y_1^{(t)}, \ldots, Y_{k-1}^{(t)}) \\
&Y_k^{mis(t)} \sim P(Y_k^{mis} \mid Y_k^{obs}, Y_1^{(t)}, \ldots, Y_{k-1}^{(t)}, \theta_k^{(t)})
\end{aligned} \quad .$$

The name "chained equation" comes from the feature of this algorithm that a k-dimensional problem is split into k one-dimensional problems, which provides a considerable amount of flexibility. It allows us to specify various conditional distributions (e.g. multinomial or Poisson) for each random variable, resulting in a more flexible and possibly better imputation method than adapting draws from joint models that do not approximate actual data distributions very well. Constraints involving imputed values are also relatively easy to maintain.

2.4.6 MCMC Strategies for Ordinal Repeated Measures and for a combination of Ordinal and Continuous Repeated Measures

Early development of MCMC strategies for multivariate data focused on normal-theory models (Gelfand and Smith 1990, Schafer 1997), where conditional distributions could be represented using familiar regression relationships. Modeling of ordinal data involves additional technical challenges, with some early methods relying on non-conjugate prior-to-posterior updating in a Bayesian paradigm (Chib and Greenberg 1998).

Zhang, Boscardin and Belin (2006) outlined an MCMC strategy using parameter-expanded data augmentation with a Metropolis-Hastings step to embed the analysis of dependent

ordinal repeated measures in a multivariate-normal latent variable framework. Zhang, Boscardin and Belin (2008) extended the idea to accommodate both ordinal and continuous repeated measures.

Define

$$T_i = \begin{bmatrix} v_{i1} \\ \vdots \\ v_{ip_1} \\ c_{i1} \\ \vdots \\ c_{ip_2} \end{bmatrix} \begin{matrix} \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} \text{continuous} \\ \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} \text{ordinal} \end{matrix} = \begin{bmatrix} V_i \\ C_i \end{bmatrix},$$

where

$$\begin{bmatrix} v_{i1} \\ \vdots \\ v_{ip_1} \end{bmatrix} = V_i \sim N(X_i^{(1)}\beta, \Sigma_{vv}).$$

One can assume the categorical outcome c_{ij} be determined by the latent normal variable z_{ij} through the relationship $c_{ij} = l$ iff $z_{ij} \in (\gamma_{jl-1}, \gamma_{jl}]$, where $\gamma_{j0} = -\infty$, $\gamma_{jJ_j} = \infty$, and c_{ij} is an ordinal variable with possible values $0 \cdots (J_j - 1)$. Here $X_i^{(1)}$ is the covariates of subject *i* associated with continuous variables. Similarly $X_i^{(2)}$, introduced below, is associated with ordinal variables.

One can also assume the conditional distribution $z_{ij} \sim N(X_i^{(2)}\beta, R_{zz})$, wehre R_{zz} is a correlation matrix. Here z_{ij} is assumed to have unit variance to preserve the identifiability of the model.

Hence we have

$$\begin{bmatrix} V_i \\ Z_i \end{bmatrix} \sim N(\begin{bmatrix} X_i^{(1)} \\ X_i^{(2)} \end{bmatrix} \beta, \begin{bmatrix} \Sigma_{vv} & \Sigma_{vz} \\ \Sigma_{zv} & R_{zz} \end{bmatrix} \equiv \Lambda),$$

where Λ is a covariance matrix with constraints related to the structure of R_{zz}.

Compared to a familiar multivariate model, which assumes

$$\underset{1\times k}{(y_{i1}\cdots y_{ik})} = \underset{1\times p}{(x_{i1}\cdots x_{ip})}\ \underset{p\times k}{\beta} + \varepsilon_i \ ,$$

this model has the form

$$\begin{bmatrix} v_{i1} \\ \vdots \\ v_{ip_1} \end{bmatrix} = \underset{p_1\times k}{X_i^{(1)}}\ \underset{k\times 1}{\beta} + \text{error}_1$$

$$\begin{bmatrix} z_{i1} \\ \vdots \\ z_{ip_2} \end{bmatrix} = \underset{p_2\times k}{X_i^{(2)}}\ \underset{k\times 1}{\beta} + \text{error}_2$$

where p_1 is the number of repetitions of continuous outcomes and p_2 is the number of repetitions of ordinal outcomes. It can be more general to assume separate linear parameters β_1 and β_2 instead of β, depending on the relationship between the continuous and the categorical outcome variables.

The main challenge in sampling parameters from this model is drawing the covariance matrix, which contains a sub-matrix that has to be a correlation matrix in order to identify the ordinal portion of the model. Zhang, Boscardin and Belin (2006) suggested using a parameter extended Wishart (PXW) proposal distribution and a parameter extended Metropolis-Hastings step for drawing Λ.

A matrix can be drawn from a parameter-extended Wishart distribution by drawing a matrix from a Wishart distribution and dropping the variance part in order to satisfy the correlation-matrix constraint. Specifically, to draw proposal values of Λ, one can draw $\Sigma \sim Wishart(m_0, \Omega)$ and let

$$\Sigma = \begin{bmatrix} D_{vv}^{1/2} R_{vv} D_{vv}^{1/2} & D_{vv}^{1/2} R_{vz} D_{zz}^{1/2} \\ D_{zz}^{1/2} R_{vz}^T D_{vv}^{1/2} & D_{zz}^{1/2} R_{zz} D_{zz}^{1/2} \end{bmatrix}.$$

Only the part

$$\Lambda = \begin{bmatrix} D_{vv}^{1/2} R_{vv} D_{vv}^{1/2} & D_{vv}^{1/2} R_{vz} \\ R_{vz}^T D_{vv}^{1/2} & R_{zz} \end{bmatrix}$$

is used in the iterative simulation procedure. Note that D_{zz} is an artificial construct introduced to help in the estimation procedure. Letting

$$D = \begin{bmatrix} D_{vv} & \\ & D_{zz} \end{bmatrix}$$

and

$$R = \begin{bmatrix} R_{vv} & R_{vz} \\ R_{zv} & R_{zz} \end{bmatrix},$$

a mapping $g(\Sigma) = (R, D)$ on the Wishart random variable is used for variable transformation. The density function of (R, D) is $\nabla g^{-1}(R, D) \cdot W(g^{-1}(R, D))$ where $W(\cdot)$ is the Wishart density function. This density function shows up in the Metropolis-Hastings step of the iterative simulation procedure.

Zhang, Boscardin and Belin (2008) assumed complete data, with the objective being to find posterior distribution of $p(\beta,\Lambda,\gamma,Z|V,C)$ where Z is the latent variable, V represents continuous variables, C represents ordinal variables and γ represents the cut-points associated with the ordinal variables, while β, Λ represents the same as above. We then can represent the posterior distribution as

$$p(\beta,\Lambda,\gamma,Z|V,C)$$

$$\propto p(\beta,\Lambda,\gamma)p(Z,V|\beta,\Lambda,\gamma)p(C|\beta,\gamma,\Lambda,Z,V).$$

Making use of conditional independence properties in the model, we can simplify to

$$\propto p(\beta,\Lambda,\gamma)p(Z,V|\beta,\Lambda)p(C|\gamma,Z)$$

$$= p(\beta)p(R,D)p(\gamma)\prod_{i=1}^{N} N(Y_{i|}|X_i\beta,\Lambda)\cdot I_i \ ,$$

where

$$Y_i = \begin{bmatrix} V_i \\ Z_i \end{bmatrix}$$

$$X_i = \begin{bmatrix} X_i^{(1)} \\ X_i^{(2)} \end{bmatrix}$$

and I_i is an indicator for C_i being compatible with Z_i and γ_i.

Assuming prior distributions

$$\begin{aligned} &\beta \sim N_k(b_0,B_0) \\ &p(\gamma_{jl}) \propto 1 \\ &\Lambda \sim PXW(m_0,\Omega) \end{aligned} \ ,$$

an MCMC algorithm with one Metropolis-Hastings step is derived from the posterior distribution:

— $p(\beta \mid \Lambda, \gamma, Z, V, C) \propto p(\beta) \prod_{i=1}^{N} I_i \cdot N(Y_i \mid X_i \beta, \Lambda)$. Hence

$$\beta \mid \Lambda, \gamma, Z, V, C \sim N(\widehat{\beta}, V_\beta),$$

where

$$V_\beta = (\sum_{i=1}^{N} X_i^T \Lambda^{-1} X_i + B_0^{-1})^{-1}$$

and

$$\widehat{\beta} = V_\beta (\sum_{i=1}^{N} X_i^T \Lambda^{-1} X_i + B_0^{-1} b_0),$$

according to a standard Bayesian result.

— $p(Z_i \mid \Lambda, \gamma, \beta, V, C) \propto I_i \cdot N(Y_i \mid X_i^{(2)} \beta, \Lambda_{22})$, i.e. the multivariate normal distribution constrained on the possible intervals. Specifically, if we let $j' \equiv p_1 + j$, each component of Z given other components of Z and other variables in the model has density function $p(z_{ij} \mid \beta, \Lambda, \gamma, V_i, C_i, z_{ik}, k \neq j) \propto I_{ij} \cdot N(z_{ij} \mid \widetilde{\mu_{ij}}, \widetilde{\lambda_{ij}})$ where $\widetilde{\mu_{ij}}$ and $\widetilde{\lambda_{ij}}$ are standard results of conditional distribution of one component given the rest in a multivariate normal distribution. The latent variables are drawn component by component for better efficiency (constraining a series of univariate normal variables, each to an interval vs. constraining a multivariate normal variable to a high-dimensional rectangle).

— $p(\gamma \mid \Lambda, \gamma, \beta, V, C)$ is uniformly sampled from the interval that is compatible with other cut points and ordinal data C.

—The Metropolis-Hastings step for $\Lambda = (R, D)$ involves the following: generate (R^*, D^*) by sampling a $D^{*1/2} R^* D^{*1/2} = W^* \sim Wishart(m, W^{(i)})$, and then take

$$(R^{(i+1)}, D^{(i+1)}) = \begin{cases} (R^*, D^*) & \text{with probability } \alpha \\ (R^{(i)}, D^{(i)}) & \text{otherwise} \end{cases}$$

where $\alpha = \min\{(p(R^*, D^*) q(W^{(i)} | W^*)) / p(R^{(i)}, D^{(i)} | Y) q(W^* | W^{(i)})), 1\}$, and where $q(W^*, W^{(i)})$ is the proposal density function of $PXW(m, W^{(i)})$ described earlier.

2.4.7 A Bayesian Implementation of Factor Analysis

Quinn (2004) used a Bayesian method to fit a factor analysis model to a mixture of ordinal and continuous variables. The author assumes that the ordinal variables follow probit models, while the latent variables underlying the ordinal variables and the continuous outcomes are collectively governed by a factor analysis model. Specifically, let subject $i = 1...N$ and outcome variable $j = 1...J$. Suppose $\underset{N \times J}{Y}$ is the outcome variable matrix. Let the matrix $\underset{N \times J}{Y^*}$ be composed of elements y_{ij}^* being the latent variable associated with y_{ij} if outcome j is ordinal, while y_{ij}^* being equal to y_{ij} otherwise. Suppose variable j is ordinal valued in $1...C_j$. Define $x_{ij} = c$ if $x_{ij}^* \in (\gamma_{j(c-1)}, \gamma_{j(c)})$, where $(\gamma_{j(0)}, \gamma_{j(1)}, ..., \gamma_{j(C_j)})$, or $(-\infty, 0, ..., +\infty)$ are cut-points, where $\gamma_{j(2)}, ..., \gamma_{j(C_j - 1)}$ are free parameters.

According to the factor analysis model, let J-vector $y_i^* = \Lambda \phi_i + \varepsilon_i$. By transposing and stacking the y_i^*' vectors for all subjects, we obtain the matrix form of the factor analysis model

$$\underset{N\times J}{Y^*} = \underset{N\times K}{\Phi}\ \underset{K\times J}{\Lambda'} + E$$

where

$$\Lambda = \begin{bmatrix} \lambda_{11} & \lambda_{12} & 0 & & \cdots & 0 \\ \lambda_{21} & \lambda_{22} & \lambda_{23} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \vdots & & & 0 \\ \lambda_{K1} & \cdots & & & & \lambda_{KK} \\ \vdots & & & & & \vdots \\ \lambda_{J1} & \cdots & & & & \lambda_{JK} \end{bmatrix}$$

with constraints $\lambda_{12} > 0$ ... $\lambda_{K,K} > 0$ for identifiability. Here $K-1$ is the number of factors because of the intercept term in the model. Accordingly, the matrix Φ contains a column of 1's.

Assume the factor scores for subject i is $\phi_{i(2:K)} \sim N(0, I_{K-1})$, with error terms represented as $\text{Var}\ \varepsilon_i \sim \Psi_{J\times J}$, where $\Psi_{J\times J}$ is a diagonal matrix with constraint $\psi_{jj} = 1$ if variable j is ordinal.

Prior distributions for free parameters are taken to be independent with:

$$\psi_{jj} \sim \text{IG}(\frac{a_{0j}}{2}, \frac{b_{0j}}{2}) \ ,$$

$\lambda_{jk} \sim N(l_{0jk}, L_{0jk})$ truncated from 0 below if $\lambda_{jk} > 0$ is assumed,

and cut-points are assumed to have a flat priors.

The joint likelihood can be written as

$$\begin{aligned} & p(Y, Y^*, \gamma, \Lambda, \Phi, \Psi) \\ \propto\ & p(Y \mid Y^*, \gamma) \cdot p(Y^* \mid \Lambda, \Phi, \Psi) \cdot p(\Lambda) p(\Phi) p(\Psi) \\ \propto\ & \prod_{i=1}^{N}\prod_{j=1}^{J} (\text{I(j is continuous)} + \text{I(j is ordinal)} \sum_{c=1}^{C_j} \text{I}(y_{ij} = c)\text{I}(y_{ij}^* \in (\gamma_{j(c-1)}, \gamma_{j(c)}))) \cdot \prod_{i=1}^{N} (p_{N(y_i^* \mid \Lambda\phi_i, \Psi)}) \cdot p(\Lambda) p(\Phi) p(\Psi) \end{aligned}$$

An MCMC procedure can be developed using the following conditional distributions to guide steps of the algorithm:

$$\phi_{i(2:K)} \mid -\sim N((I_{K-1} + \Lambda'_{\cdot(2:K)} \Psi^{-1} \Lambda_{\cdot(2:K)})^{-1} \Lambda'_{\cdot(2:K)} (y_i^* - \Lambda_{\cdot 1}) \ , (I_{K-1} + \Lambda'_{\cdot(2:K)} \Psi^{-1} \Lambda_{\cdot(2:K)})^{-1})$$

$\psi_{jj} \mid - \sim \text{IG}(\frac{a_{0j}+N}{2}, \frac{b_{0j}+(Y^*_{\bullet j}-\Phi(\Lambda_{j\bullet})')'(Y^*_{\bullet j}-\Phi(\Lambda_{j\bullet})')}{2})$ for ordinal j

$y^*_{ij} \mid - \sim N(\Lambda_{j\bullet}\phi_i, 1)$ truncated to $(\gamma_{j(x_{ij}-1)}, \gamma_{j(x_{ij})})$ for ordinal variable j

$\gamma \mid \Phi\ \Lambda$ (blocking Y^*): a Metropolis-Hastings step for the $C_j - 2$ free components of γ_j, i.e. $\gamma_{j(c)}$ where $c = 2...C_j$. The proposal $\gamma^{(can)}_{j(c)}$ is drawn from $N(\gamma_{j(c)}, t_j^2)$ truncated to $(\gamma^{(can)}_{j(c-1)}, \gamma_{j(c+1)})$ with acceptance probability

$$\alpha_j = \prod_{i=1}^{N} \frac{\Phi(\gamma^{(can)}_{j(y_{ij})} - \Lambda_{j\bullet}\phi_i) - \Phi(\gamma^{(can)}_{j(y_{ij}-1)} - \Lambda_{j\bullet}\phi_i)}{\Phi(\gamma_{j(y_{ij})} - \Lambda_{j\bullet}\phi_i) - \Phi(\gamma_{j(y_{ij}-1)} - \Lambda_{j\bullet}\phi_i)}$$
$$\prod_{c=2}^{C_j-1} \frac{\Phi((\gamma_{j(c+1)} - \gamma_{j(c)})/t_j) - \Phi((\gamma^{(can)}_{j(c-1)} - \gamma_{j(c)})/t_j)}{\Phi((\gamma^{(can)}_{j(c+1)} - \gamma^{(can)}_{j(c)})/t_j) - \Phi((\gamma_{j(c-1)} - \gamma^{(can)}_{j(c)})/t_j)}$$

where the former part comes from the normal assumption of Y^* (integrated out), while the latter part comes from the proposal distribution $\gamma^{(can)}_{j(c)} \sim N(\gamma_{j(c)}, t_j^2)$ constrained on the interval $(\gamma^{(can)}_{j(c-1)}, \gamma_{j(c+1)})$.

$\Lambda \mid -$: Each row of Λ is drawn separately. Let $\lambda_{j\circ}$ be the column vector of free elements of the j th row of Λ, $\lambda_{j\bullet}$ be the column vector of fixed elements of the j th row of Λ. Similarly let $\Phi_{j\circ}$ and $\Phi_{j\bullet}$ be the corresponding parts of Φ. Since

$$Y^*_{\bullet j} \mid \Phi\Lambda\Psi\gamma \sim N((\Phi_{j\circ}\Phi_{j\bullet})\begin{pmatrix}\lambda_{j\circ}\\ \lambda_{j\bullet}\end{pmatrix}, \psi^2_{jj} I_N)$$

is truncated onto some support only related to γ, we draw from the conditional distribution

$$\lambda_{j\circ} \mid Y^*\Phi\Psi \sim N((L_{0j\circ} + \psi_{jj}^{-1}\Phi'_{j\circ}\Phi_{j\circ})^{-1}(l_{0j\circ} + \psi_{jj}^{-1}\Phi'_{j\circ}(Y^*_{\bullet j} - \Phi_{j\bullet}\lambda_{j\bullet})),$$
$$(L_{0j\circ} + \psi_{jj}^{-1}\Phi'_{j\circ}\Phi_{j\circ})^{-1})$$

.

Building on these ideas, in the rest of this dissertation, we develop a longitudinal factor analysis model to accommodate incomplete high-dimensional longitudinal data, and we explore the statistical properties of the procedure.

2.4.7 Longitudinal Factor Models

Wang (2002) developed a model imputing missing data in a multivariate longitudinal data set. This model included only continuous outcome variables; any categorical variable is imputed using an approximation based on continuous variables (for example, rounding the imputed values to the nearest possible values). One objective of our project is to incorporate a probit model for ordinal variables, in order to make the model more flexible while maintaining the validity of the imputed values. The strategy developed here builds on the framework outlined in Wang (2002).

Lin (2012) developed a related model analyzing multivariate longitudinal data measured at irregular time points. Based on the general linear mixed-effect model (GLMM), this model included outcome variables distributed as normal (for continuous variables), Poisson (for count variables) and Bernuolli (for dichotomous variables), where dimension reduction is achieved by a factor analysis model.

Chapter 3
The Longitudinal Factor Imputation Model

The Longitudinal Factor (LF) imputation model consists of two stages explaining the variability of the data within-time-point and between-time-point. Specifically, we use a linear mixed-effect model specification to characterize longitudinal patterns of associations, and we use a factor-analysis model specification to characterize cross-sectional associations among variables measured at the same time. The model is complicated enough that we need to use an iterative-simulation MCMC method to fit the model.

3.1 Structure of the LF Imputation Model

Stage 1 of the LF Imputation Model characterizes patterns of cross-sectional association within time points. The variables in the imputation model can be either covariates or outcomes in subsequent analysis models. Specifically, we can represent the variables measured on subject *i* as:

$$Y_i = \begin{bmatrix} Y_{i11} & \dots & Y_{i1p_1} & Y_{i1p_1+1} & \dots & Y_{i1p_1+p_2} \\ & \dots & & \dots & & \\ Y_{iT1} & \dots & Y_{iTp_1} & Y_{iTp_1+1} & \dots & Y_{iTp_1+p_2} \end{bmatrix},$$

where Y_i is a $T \times (p_1 + p_2)$ matrix, where element Y_{itj} stands for the j^{th} variable at the t^{th} time point on subject *i*. Without loss of generality, we let indices $1 \dots p_1$ refer to continuous variables,

while indices $p_1 \ldots (p_1 + p_2)$ refer categorical variables with possible values $1 \ldots C_j$ (i.e., the j[th] variable has C_j categories).

To facilitate our formulation, let $U_{itj} = Y_{itj}$ for $1 \le j \le p_1$ and U_{itj} be a probit model latent variable for $p_1 + 1 \le j \le p_1 + p_2$. So we have

$$U_i = \begin{bmatrix} U_{i11} & \ldots & U_{i1p_1} & U_{i1p_1+1} & \ldots & U_{i1p_1+p_2} \\ & \ldots & & & \ldots & \\ U_{iT1} & \ldots & U_{iTp_1} & U_{iTp_1+1} & \ldots & U_{iTp_1+p_2} \end{bmatrix}$$

which is a $T \times (p_1 + p_2)$ matrix of continuous variables. Denote the cut-points involved in the probit model for variable j, where $p_1 + 1 \le j \le p_1 + p_2$, by

$$-\infty = \xi_{j(0)} \le \xi_{j(1)} \ldots \xi_{j(C_j - 1)} \le \xi_{j(C_j)} = +\infty .$$

Hence we have

$$Y_{itj} = a \text{ if and only if } \xi_{j(a-1)} \le U_{itj} \le \xi_{j(a)}$$

where $1 \le a \le C_j$. Accordingly, the identity $U_{itj} \in (\xi_{j(Y_{itj}-1)}, \xi_{j(Y_{itj})})$ always holds. Since shifting all of the cut-points by some quantity is equivalent to shifting the probit latent variable by the same quantity, we add the constraint $\xi_{j(1)} = 0$ to ensure identifiability of model parameters.

Letting U_{it}' be the t^{th} row of U_i, we conceptualize the factor analysis model as:

$$\underset{(p_1+p_2)\times 1}{U_{it}} = \begin{bmatrix} U_{it1} \\ \vdots \\ U_{it(p_1+p_2)} \end{bmatrix} = \underset{(p_1+p_2)\times(K+1)}{\Lambda} \underset{(K+1)\times 1}{\begin{bmatrix} 1 \\ f_{it} \end{bmatrix}} + \varepsilon_{it} = \Lambda \begin{bmatrix} 1 \\ f_{it1} \\ \vdots \\ f_{itK} \end{bmatrix} + \varepsilon_{it}, \qquad (3.1)$$

where $\begin{bmatrix} f_{it1} \\ \vdots \\ f_{itK} \end{bmatrix}$ is the factor score of subject *i* at time *t*, and Λ is the loading matrix. We may transpose both sides of Equation 3.1into rows, then stack those rows of the *n* subjects and *T* time points into a matrix form

$$\underset{nT\times(p_1+p_2)}{U} = \underset{nT\times(K+1)}{(1_{nT}, F)}\Lambda^{\mathrm{T}} + \mathrm{E}$$

which is more convenient to use in the computing process. We let the loading matrix Λ be common across all time points, while the temporal trend is explained by the trajectory model (Stage 2) of the factor scores over time.

The elements of the error term ε_{it} are assumed to be independently normally distributed, with variances ψ_j^2 / w_t for all *j* and *t*. This framework anticipates carrying at a weighted least square linear regression for the j^{th} variable where $j = 1...p_1$, and where the probit latent variables (indices $p_1 + 1 \ldots p_1 + p_2$) have unit variance. The weight parameters w_t for $t = 1...T$ model the relative dispersion of the outcome variables at the t^{th} time point, where it is assumed that $0 < w_t < 1$ and $w_1 + \cdots + w_T = 1$ for $t = 1...T$.

These weight parameters are designed to deal with the possibility of hetroscedasticity in the distributions governing a given quantity at different time points. Such a scenario could arise when the resources available to generate data vary over time. For example, outcome measurements might be based on an average of subjective assessments by multiple raters, and the number of raters could be different at distinct time points.

In modeling the dispersion parameters of the p_1 continuous variables across the T time points, the most general assumption could be assuming $p_1 \cdot T$ variance terms, which could result

in too many parameters in the model, making it hard to estimate. A model without weight parameters would assume each continuous variable has a constant variance over time, corresponding to p_1 variance parameters. Adopting the weighted regression model, we include $p_1 + T$ parameters to accommodate the possibility of hetroscedasticity, thereby going in between these alternatives. A feature of this approach is that we assume that the dispersion of all continuous variables change in the same way across time points, as in the scenario where the number of raters change over time but is the same for all outcomes measured at a given time. In the application considered here, patients' measurements are made at discharge from the hospitalization following treatment for a broken jaw as well as 10 days 1 month, 6 months, 12 months afterwards, which gives rise to a concern that dispersion of outcome measure could increase over time. The weighted regression model offers a framework for researchers to check for certain departures from a homoscedasticity assumption.

As in a general weighted least square linear regression, time points with higher weights correspond to more influence on the model parameters, and lower weights correspond to less influence. The weights are drawn from their posterior distribution, which is determined mostly by the conditional variances and missingness of the outcome variables in each time point. Formally,

$$\varepsilon_{it} \sim N_{p_1+p_2}\left(0, \begin{bmatrix} \frac{1}{w_t}\begin{bmatrix} \psi_1^2 & & \\ & \ddots & \\ & & \psi_{p_1}^2 \end{bmatrix} & \\ & I_{p_2} \end{bmatrix}\right) \tag{3.2}$$

The constant element 1 attached to the vector of the factor scores in formula (3.1) simplifies the way to allow or not to allow the intercepts in the outcome model. For example,

fixing an element in the first column of Λ to 0 means not to allow an intercept; while dropping this constraint implies that an intercept is included.

The intercepts associated with the latent probit variables correspond to the conditional probability of the lowest level of the ordinal variables, a quantity that is needed to produce imputations. On the other hand, we chose not to model the intercepts of the continuous variables, since compared to not modeling them, modeling the intercepts yields similar imputation results but much greater auto-correlation in the Markov-chain Monte Carlo estimation procedure based on our simulation evaluations. As the mission of the LF model is to impute missing values based on observed data rather than to estimate parameters, we shift the continuous variables by a constant (say, their sample means) in the beginning and add back the constant after imputation to avoid modeling the intercepts of continuous variables.

We therefore fix elements in the first column of Λ to 0, i.e., let $\Lambda_{j1} = 0$ for $1 \le j \le p_1$ as a constraint as proposed by Geweke and Zhou (1996), we also impose constrains on a triangular array of elements in the upper-right corner of Λ, i.e. $\Lambda_{jk} \ge 0$ for $k \ge j+2$, to avoid undesired rotations of Λ,.

In Stage 2 of the LF Imputation Model, we assume that factor scores follow a linear mixed-effect model

$$\underset{K\times 1}{f_{it}} = \underset{K\times a}{\alpha}\ \underset{a\times 1}{X_{it}} + \underset{K\times b}{\gamma_i}\ \underset{b\times 1}{Z_{it}} + \delta_{it} \tag{3.3}$$

where X is a set of variables with complete data. Theoretically, any variable that is collected completely can be put in X. However, more parameters are introduced in the model if we include a variable in X compared to adding a variable into the outcome Y. In choosing where to put a

variable, we need to consider how it makes sense to relate a variable to factor scores and their trajectories.

We assume the random effects to be distributed as

$$(\gamma_i')^{\mathrm{V}} \sim N_{Kb}(0, \begin{bmatrix} R_1 & & \\ & \ddots & \\ & & R_K \end{bmatrix})$$

where each of R_k, $k = 1...K$ is a $b \times b$ positive-definite matrix. Here the "V" operator means the vectorization of a matrix by concatenating its columns.

The error part of the factor score, δ_{it}, is assumed to have unit variance for the purpose of identifiability of the model. In additional to the block-diagonal error structure for random effects, we use a one-step auto-regressive (AR1) error structure to accommodate longitudinal correlation among factor scores. Given X_{it}, Z_{it} and the model parameters, each of the *K T*-length vectors of factor score is assumed to be independent from each other. Specifically, we assume

$$\begin{pmatrix} \delta_{i1k} \\ \vdots \\ \vdots \\ \delta_{iTk} \end{pmatrix} \sim N_T(0, \mathrm{AR1}_T(\rho))$$

for $k = 1... K$, where the matrix

$$\mathrm{AR1}_T(\rho) = \begin{bmatrix} 1 & \rho & \cdots & \rho^{T-1} \\ \rho & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \rho \\ \rho^{T-1} & \cdots & \rho & 1 \end{bmatrix}.$$

To complete the LF Imputation Model specification, the prior distributions of the parameters can be characterized as follows:

ξ ~ uniform over its support

$\Lambda \sim$ element-wise $N(\mu_{0\Lambda}, \sigma^2_{0\Lambda})$ subject to factor-model identifiability constraints

$\psi_j^2 \sim \text{Inv-Gamma}(\alpha = \frac{n_0}{2}, \beta = \frac{s^2_{jj0}}{2})$

$w_t \sim N(\mu_{0w_t}, \sigma^2_{0w_t})$ constrained such that on $\sum_{t=1}^{T} w_t = 1$ and $w_t > 0$ for all t

$\alpha \sim$ element-wise $N(\mu_{0\alpha}, \sigma^2_{0\alpha})$

$R_k \sim \text{Inv-Wishart}(\text{scale} = \nu_0 S_0, \text{df} = \nu_0)$

$\rho \sim N(\mu_{0\rho}, \sigma^2_{0\rho})$ constrained on the possible support $(-1,1)$.

3.2 Fitting the LF Imputation Model

With the development of the LF Imputation Model in a hierarchical modeling framework, it is natural to use Markov-chain Monte Carlo (MCMC) statistical computing techniques to fit the model and to draw a Bayesian inference. We start the development in this section by outlining the conditional distributions that provide the foundation for an MCMC fitting procedure. We use a Metropolis-within-Gibbs sampling framework (Gelman, Carlin, Stern, Dunson, Vehtari and Rubin 2014) to draw a sample from the posterior distribution of the parameters, latent variables, and missing data. Certain steps are implemented using the Metropolis-Hastings algorithm due to the conditional distributions being not well-known distributions that can be sampled directly.

We draw the missing outcome variables and the probit-model latent variables based on

$$\underset{(p_1+p_2)\times 1}{U_{it}} \mid \Lambda\ f\ \psi\ w\ \xi\ Y\ \sim N_{p_1+p_2}\left(\Lambda\begin{pmatrix}1\\ f_{it}\end{pmatrix}, \begin{bmatrix}\frac{1}{w_t}\begin{bmatrix}\psi_1^2 & & \\ & \ddots & \\ & & \psi_{p_1}^2\end{bmatrix} & \\ & I_{p_2}\end{bmatrix}\right),$$

then truncating U_{itj} on the interval $(\xi_{j\,Y_{itj}-1}, \xi_{j\,Y_{itj}})$ for non-missing ordinal outcome variables, which correspond to indices $p_1+1 \le j \le p_1+p_2$. For U_{itj}'s corresponding to missing continuous outcome variables and missing ordinal outcome variables, no truncation is needed. Since the variance-covariance matrix is assumed to be diagonal, we draw normally distributed components individually. The covariances among the outcomes variables are explained by factor scores and the factor loading matrix as described below.

We draw the conditional variances of the outcome variables as

$$\psi_j^2 \mid U, w, \Lambda, f \sim \text{Inv-Gamma}(\ \alpha = \frac{1}{2}(n_0 + nT)\ ,\ \beta = \frac{1}{2}(s_{jj0}^2 + s_{jj}^2)\ )$$

for $j = 1 \ldots p_1$, where

$$s_{jj}^2 = \underset{nT\times 1}{(U_{\cdot j} - \begin{pmatrix}1 & F\end{pmatrix}\Lambda_{j\cdot}{}^{\mathrm{T}})^{\mathrm{T}}}(I_n \otimes \begin{bmatrix}w_1 & & \\ & \ddots & \\ & & w_T\end{bmatrix})(U_{\cdot j} - \begin{pmatrix}1 & F\end{pmatrix}\Lambda_{j\cdot}{}^{\mathrm{T}}).$$

Here $\underset{nT\times(p_1+p_2)}{U}$ and $\underset{nT\times K}{F}$ are long-form matrices of the outcome variables (including the latent variables underlying binary or ordinal outcomes) and factor scores.

We draw the weights $(w_1, ..., w_T) \mid U, f, \Lambda$ using a Metropolis-Hastings step. We start from equal weights

$$(\frac{1}{T}, ..., \frac{1}{T}),$$

so that all of the elements sum up to 1. The proposal draw is the sum of the current weights and a multivariate jumping step vector. The result is truncated such that every element is within the interval (0,1). We use the step vector

$$D-(\frac{1}{T},...,\frac{1}{T}),$$

where D~Dirichlet(τ_w, τ_w , ... τ_w). Theoretically, any multivariate distribution with elements summing up to 0 is suitable to be chosen for the step vectors, so that the proposals weights sum to 1 by constriction. The magnitude of the step vector is determined by the tuning parameter τ_w with higher values corresponding to smaller steps. We calculate the estimated scaled error part of the continuous outcome variables by

$$\underset{p_1\times 1}{h_{it}}=\begin{bmatrix}\psi_1^{-1} & & \\ & \ddots & \\ & & \psi_{p_1}^{-1}\end{bmatrix}((\underset{p_1\times 1}{U_{it}})_{1:p_1}-\Lambda_{1:p_1\cdot}\begin{pmatrix}1\\ f_{it}\end{pmatrix}).$$

We use the probability $\min(\alpha_w,1)$ to determine whether $(w_1^{(\text{can})},...,w_T^{(\text{can})})$ is accepted, where

$$\begin{aligned}\alpha_w=\exp(\ &-\frac{\sum_{t=1}^{T}(w_t^{(\text{can})}-\mu_{0w_t})^2-(w_t-\mu_{0w_t})^2}{2\sigma_{0w}^2}\\ &+\frac{n}{2}\sum_{t=1}^{T}\log\frac{w_t^{(\text{can})}}{w_t}-\frac{1}{2}\sum_{t=1}^{T}(w_t^{(\text{can})}-w_t)\sum_{i=1}^{n}h_{it}^{\mathrm{T}}h_{it}\\ &+\sum_{t=1}^{T}(\tau_w-1)\log\frac{w_t-w_t^{(\text{can})}+\frac{1}{T}}{w_t^{(\text{can})}-w_t+\frac{1}{T}}\ )\end{aligned}.$$

We draw the cut-points $\xi\mid Y,\Lambda,f$ by a Metropolis-Hasting step. The proposal draws of the cut-points for the j^{th} variable are generated iteratively for $c=2...(C_j-1)$ as

$\xi_{j(c)}^{(\text{can})} \sim N(\xi_{j(c)}, \tau_j^2)$ truncated on $(\xi_{j(c-1)}^{(can)}, \xi_{j(c+1)})$. Here the τ_j^2 where $j = (p_1+1)...(p_1+p_2)$ are tuning parameters for the j^{th} variable. Then we use the probability $\min(\alpha_j, 1)$ to determine whether $\xi_{j(2)}^{(can)} ... \xi_{j(C_j-1)}^{(can)}$ are accepted, where

$$\alpha_j = \prod_{i\,t} \frac{\Phi(\gamma_{j(Y_{itj})}^{(can)} - \Lambda_{j\cdot}\begin{bmatrix}1\\ f_{it}\end{bmatrix}) - \Phi(\gamma_{j(Y_{itj}-1)}^{(can)} - \Lambda_{j\cdot}\begin{bmatrix}1\\ f_{it}\end{bmatrix})}{\Phi(\gamma_{j(Y_{itj})} - \Lambda_{j\cdot}\begin{bmatrix}1\\ f_{it}\end{bmatrix}) - \Phi(\gamma_{j(Y_{itj}-1)} - \Lambda_{j\cdot}\begin{bmatrix}1\\ f_{it}\end{bmatrix})} \cdot \prod_{c=2}^{C_j-1} \frac{\Phi((\xi_{j(c+1)} - \xi_{j(c)})/\tau_j) - \Phi((\xi_{j(c-1)}^{(can)} - \xi_{j(c)})/\tau_j)}{\Phi((\xi_{j(c+1)}^{(can)} - \xi_{j(c)}^{(can)})/\tau_j) - \Phi((\xi_{j(c-1)} - \xi_{j(c)}^{(can)})/\tau_j)} .$$

The factor scores of the i^{th} subject $f_i \mid U_i, w, \psi, \Lambda, \rho$ are drawn based on the Bayesian solution of a linear regression for the subject, where the f_i is treated in the same way as a vector of linear coefficients. We denote

$$f_i = \underset{KT\times 1}{\begin{bmatrix} f_{i1} \\ \vdots \\ f_{iT} \end{bmatrix}}$$

and

$$\underset{(p_1+p_2)T\times(p_1+p_2)T}{\Psi_w} = \begin{bmatrix} \frac{1}{w_1}\begin{bmatrix}\psi_1^2 & & \\ & \ddots & \\ & & \psi_{p_1}^2\end{bmatrix} & & & & \\ & I_{p_2} & & & \\ & & \ddots & & \\ & & & \frac{1}{w_T}\begin{bmatrix}\psi_1^2 & & \\ & \ddots & \\ & & \psi_{p_1}^2\end{bmatrix} & \\ & & & & I_{p_2} \end{bmatrix}.$$

We draw $f_i \mid U_i, w, \psi, \Lambda, \rho \sim N(\mu_i, V_i)$, where

$$V_i = ((I_T \otimes \Lambda_{\cdot(-1)}^{\mathrm{T}})\Psi_w^{-1} I_T \otimes \Lambda_{\cdot(-1)} + (\mathrm{AR1}_T(\rho) \otimes I_K)^{-1})^{-1}$$

and

$$\mu_i = V_i \cdot (I_T \otimes \Lambda_{\cdot(-1)}^{\mathrm{T}} \Psi_w^{-1} (\begin{bmatrix} U_{i1} \\ \vdots \\ U_{iT} \end{bmatrix} - 1_T \otimes \Lambda_{\cdot 1}) + (\mathrm{AR1}_T(\rho) \otimes I_K)^{-1} (I_T \otimes \alpha \begin{bmatrix} X_{i1} \\ \vdots \\ X_{iT} \end{bmatrix} + I_T \otimes \gamma_i \begin{bmatrix} Z_{i1} \\ \vdots \\ Z_{iT} \end{bmatrix}))$$

The loading matrix $\Lambda \mid f, w, \psi$ is drawn by row since the variance-covariance matrix of the error term of U is assumed to be diagonal. The regression equation of the j^{th} outcome can be written as

$$U_{itj} = (1, f_{it}^{\mathrm{T}})_{P_j} (\Lambda_{j P_j})^{\mathrm{T}} + \varepsilon_{itj},$$

where P_j is a subset of $\{1, \cdots, (K+1)\}$ indicating which elements of $\Lambda_{j\cdot}$ are free elements instead of being assumed to fixed to 0. Only the corresponding part of the factor score vector is involved in the regression equation. The values of P_j can be written exhaustively as

$$P_1 = \{2\}$$

$$P_2 = \{2, 3\}$$

$$P_{K-1} = \{2, 3, \cdots, K\}$$

$$P_K = \cdots = P_{p_1} = \{2, 3, \cdots, K\}$$

$$P_{p_1} = \cdots = P_{p_1 + p_2} = \{1, 2, 3, \cdots, K\},$$

corresponding to the constraints of Λ we described in the section outlining the model.

Letting

$$\underset{nT\times nT}{\Psi_{w(j)}} = \begin{cases} \psi_j \, I_n \otimes \begin{bmatrix} 1/w_1 & & \\ & \ddots & \\ & & 1/w_T \end{bmatrix} & \text{if } 1 \le j \le p_1 \\ I_n \otimes \begin{bmatrix} 1/w_1 & & \\ & \ddots & \\ & & 1/w_T \end{bmatrix} & \text{if } p_1 + 1 \le j \le p_1 + p_2 \end{cases},$$

which is the conditional variance-covariance matrix of $U_{\cdot j}$, a Bayesian linear regression formula is applied to achieve the conditional distribution of $\Lambda_{j P_j} \sim N(\mu_{\Lambda_{j\cdot}}, V_{\Lambda_{j\cdot}})$, where

$$V_{\Lambda_{j\cdot}} = ((1_{nT}, F)_{\cdot P_j}^{\mathrm{T}} \Psi_{w(j)}^{-1} (1_{nT}, F)_{\cdot P_j} + V_{0\Lambda_{j\cdot}}^{-1})^{-1}$$

and

$$\mu_{\Lambda_{j\cdot}} = V_{\Lambda_{j\cdot}} ((1_{nT}, F)_{\cdot P_j}^{\mathrm{T}} \Psi_{w(j)}^{-1} U_{\cdot j} + V_{0\Lambda_{j\cdot}}^{-1} \mu_{0\Lambda_{j\cdot}}) .$$

This normal draw is then truncated invoking non-negativity constraints on the diagonal elements. Based on experience, there is a small chance for the non-negativity constraint not to be satisfied even if we have a large number of trials. In this case, one can flip the sign of the corresponding element of the factor scores for all subjects at all time points, which does not change the meaning of the model while making the non-negativity constraint more straightforward to satisfy.

We draw the temporal correlation coefficients of the factors scores $\rho \mid f, \alpha, \gamma$ by a Metropolis-Hastings step. The proposal is drawn from $\rho_{(\mathrm{can})} \sim N(\rho, \tau_\rho^2)$ truncated on the interval $(-1, 1)$, where τ_ρ^2 is a tuning parameter and ρ is the current value. We define the estimated error terms of the factor scores for subject *i* at time *t* as $e_{it} = f_{it} - \alpha X_{it} - \gamma_i Z_{it}$, and we then let $(g_{i1} \ldots g_{iK}) = (e_{i1} \ldots e_{iT})^{\mathrm{T}}$, where g_{ik} is a *T*-vector of the estimated error terms of the k^{th} element of

the factor scores across all time points belonging to subject *i*. Then we use the probability $\min(\alpha_\rho, 1)$ to determine whether $\rho_{(\text{can})}$ is accepted, where

$$\alpha_\rho = \exp(-\frac{(\rho_{(\text{can})} - \mu_{0\rho})^2 - (\rho - \mu_{0\rho})^2}{2\sigma_{0\rho}^2}$$
$$-\frac{nk}{2}\log\frac{\det \text{AR1}_T(\rho_{(\text{can})})}{\det \text{AR1}_T(\rho)}$$
$$-\frac{1}{2}\text{tr}((\text{AR1}_T(\rho_{(\text{can})}) - \text{AR1}_T(\rho))\sum_{ik} g_{ik} g_{ik}^{\text{T}}))$$

We draw the fixed effect α and the random effects γ_i by a similar rearrangement of the elements of the factor scores such that one component of the factor scores across all *T* time points of a subject is treated as a *T*-vector in a regression equation. Specifically, letting

$$(\phi_{i1}...\phi_{iK}) = (f_{i1}...f_{iT})^{\text{T}},$$

Equation 3.3 (on page 36) implies

$$\underset{T\times 1}{\phi_{ik}} = \underset{T\times a}{\begin{bmatrix} X_{i1}^{\text{T}} \\ \vdots \\ X_{iT}^{\text{T}} \end{bmatrix}} \alpha_{k\cdot}^{\text{T}} + \underset{T\times b}{\begin{bmatrix} Z_{i1}^{\text{T}} \\ \vdots \\ Z_{iT}^{\text{T}} \end{bmatrix}} (\gamma_i)_{k\cdot}^{\text{T}} + g_{ik} \quad ,$$

where g_{ik} is defined in the ρ step. For a given $k = 1...K$, the *n* subjects contribute a sample for deriving the conditional distribution of α, γ_i and R_k. Therefore, we stack the factor scores and covariates of all subjects into long-form matrices of Φ, *X* and *Z*, while denoting Φ_i, X_i and Z_i as the rows belonging to the i^{th} subject in these matrices. We hence draw

$$\alpha_{k\cdot}^{\text{T}} \mid f, R_k, \rho \sim N(\mu_{\alpha_{k\cdot}}, V_{\alpha_{k\cdot}})$$

where

$$V_{\alpha_{k\cdot}} = (\sum_{i=1}^{n} X_i^T (AR1_T(\rho) + Z_i^{\text{T}} R_k Z_i)^{-1} X_i + V_{0\alpha}^{-1})^{-1},$$

and

$$\mu_{\alpha_{k\cdot}} = V_{\alpha_{k\cdot}} \left(\sum_{i=1}^{n} X_i^{\mathrm{T}} (\mathrm{AR1}_T(\rho) + Z_i^{\mathrm{T}} R_k Z_i)^{-1} (\Phi_i)_{\cdot k} + V_{0\alpha}^{-1} \mu_{0\alpha} \right).$$

We draw

$$(\gamma_i)_{k\cdot}^{\mathrm{T}} \mid f, R_k, \alpha, \rho \sim N(\mu_{\gamma_{ki}}, V_{\gamma_{ki}})$$

where

$$V_{\gamma_{ki}} = (Z_i^{\mathrm{T}} \ \mathrm{AR1}_T(\rho)^{-1} \ Z_i + R_k^{-1})^{-1}$$

and

$$\mu_{\gamma_{ki}} = V_{\gamma_{ki}} \ Z_i^{\mathrm{T}} \ \mathrm{AR1}_T(\rho)^{-1} ((\Phi_i)_{\cdot k} - X_i \alpha_{k\cdot}{}^{\mathrm{T}}).$$

We finally draw the variance-covariance matrices of the random effects

$$R_k \mid \gamma \sim \text{Inv-Wishart}_{\mathrm{b}}(\text{scale} = \nu_0 S_0 + \sum_{i=1}^{n} (\gamma_i)_{k\cdot}^{\mathrm{T}} (\gamma_i)_{k\cdot} \ , \text{df} = \nu_0 + n).$$

The MCMC procedure iterates the above steps a large number of times until we see evidence of convergence.

3.3 Starting Values

In general, the number of iterations leading to convergence of an MCMC procedure depends on starting values for model parameters. In line with principles developed in Gelman and Rubin (1992) and Schafer (1997), we chose starting values based on the following considerations:

For parameters of associations in the imputation model, we choose starting values that suggest limited or no associations and that overstate variability. This is built on a perspective that

the greatest risk to Bayesian inference emerges from assumptions that are artificially precise and favor inferences showing substantial associations.

We anticipate an investigator would run less complicated models using available software before applying the LF Imputation Model. For example, a set of univariate one-way ANOVA's or univariate linear mixed models would give quantities such as the total sum of squares, which leads to a reasonable upper bound for conditional variances.

Specifically, the covariance, the correlation, the association and the loading parameters are taken to start from their null values. That is, the loading matrix Λ starts from a 0 matrix; the weights $(w_1,...w_T)$ start from equal weights $1/T$; the fixed effects α start from a 0 matrix, and the auto-regressive coefficient ρ starts from 0. The variances of error terms start from values 1 for ψ_j^2 and values $\mathrm{I}_T/2$ for R_k, which based on preliminary analyses can be expected to be overestimates. This ensures a reasonable contribution in the model-fitting across all outcomes variables in the first few iterations.

The cut-points start from crude guesses based on the marginal probabilities of categories. For example, ξ_j, which is a (C_j+1)-vector, starts from a transformation of cut-points resulted from an ordinal logistic model of the corresponding ordinal outcome variable without any predictors. Compared to the cut-points ζ_j in an ordinal logistic regression, which can be arrayed as

$$-\infty=\zeta_{j(0)}<\zeta_{j(1)}<\ldots<\zeta_{j(C_j-1)}<\zeta_{j(C_j)}=+\infty,$$

the cut-points in the LF Imputation Model can be arrayed as

$$-\infty=\xi_{j(0)}<0=\xi_{j(1)}<\xi_{j(2)}<\ldots<\xi_{j(C_j-1)}<\xi_{j(C_j)}=+\infty.$$

We let $\xi_{j(c)}$ start from $.588\,(\zeta_{j(c)} - \zeta_{j(1)})$ for $c = 2 \ldots C_j - 1$, where the proportion .588 comes from the shape difference between a normal distribution and a logistic distribution.

This set of starting values worked reasonably well in the simulations described in Chapter 5, in the sense that all Markov chains we drew showed satisfactory convergence within roughly 6000 iterations.

Chapter 4
Statistical Computing Techniques

A substantial part of this project has been to develop statistical software implementing the LF Imputation Model and to test its performance in different hypothetical scenarios. This chapter documents the choices we made in implementing the LF Imputation Model, in order to facilitate future efforts in the continuation of this project or other projects involving similar intensity of statistical computing.

To place the implementation choices in context, we describe briefly the evaluation strategy outlined in Chapter 5, given the imperative to carry out simulation studies to assess both the accuracy of the program and the statistical properties of the procedure, both of which induce certain demands on the computational task at hand. We defer many of the relevant details of the evaluation strategy to the next chapter, aiming here only to motivate the choices in the architecture of a statistical-computing strategy.

As mentioned in Chapter 1, the research question is to evaluate the performance of the LF Imputation Model across different settings characterized by the dimension of outcome measures, the number of subjects, the dimension of internal factor scores, and the number of time points. We call each combination of the above factors a **scenario**, e.g. the combination of 40 outcome variables, 5 internal factor scores, 50 subjects and 5 time points is one scenario of interest. In each scenario, we incorporated some item missingness through a missing-completely-at-random (MCAR) mechanism and we also consider 2 hypothetical missingness mechanisms with missingness rates of either 20% or 100% on the outcome variable we analyzed, where 100%

missingness implies that the variable was not measured at the given time point and imputations are based entirely on patterns of associations with other variables measured at the same time point or with the same variable at different time points. Both of the missingness mechanisms belong to missing-at-random (MAR). We call a scenario with one of the above two missingness mechanism a **case**. The performance of the LF Imputation Model is tested on 64 cases (32 scenarios × 2 missingness mechanisms). For each case, the test results are based on 204 data sets with shared underlying model parameters but with different errors and random effects. The quantities used to evaluate of the LF Imputation Model include the actual coverage rates of model parameters, the average bias in estimates, and the average confidence interval lengths. For example, when intervals from 190 of the 204 generated data sets cover the parameter of interest in a certain case, we calculate $190/204 = 93.1\%$ as the actual coverage rate in this case.

For a certain scenario, all of the 204 complete data sets share the same factor structure, leaving the normal errors and random effects to vary. Different factor structures are nested in the 32 scenarios, i.e. we tested the LF Imputation Model under 32 different scenarios and 32 different factor structures which are randomly chosen, assuming factor structures do not interact with the scenario settings.

As described earlier, we test the LF Imputation Model in 64 cases. We generate one data set for each case and apply the proposed imputation method and analysis procedures. We repeat this approach 204 times using different random seeds to produce different residual errors and random effects. We call each repetition among the sequence of 204 data sets a **replication**, which consists of applying the proposed procedure once for each case. In this dissertation, if a specific case is not mentioned, "a replication" could also be interpreted to mean applying the LF Imputation Model to all cases across all cases within a scenario.

Fitting the LF Imputation Model in each of 204 replications across 64 cases is a computationally demanding task. To tackle the issues surrounding the amount of computation, we wrote our MCMC procedure in a low-level language (C++) instead of R, and parallelized the 204 replications using the Hoffman 2 cluster. Compared to R, programs written in low-level languages can often perform the same amount of computation by a factor of 10−100 times faster. On the other hand, the efficiency gain from parallelization depends on the current workload of the cluster. Based on our experience, the cluster sometimes allowed all 204 jobs to be running at a time, but at times the cluster allowed as few as 10 jobs at a time. Typically we experienced about 60 jobs running at a time. Our efforts led to a gain in computation efficiency by a factor of approximately 600 times compared to a single computer running the same program written in R.

Each replication, which corresponds to 64 data sets, typically took about 2 hours when 8000 MCMC iterations were drawn for each chain. Because each scenario differs in complexity, we count the total time taken across all data sets in a replication. Without any parallelization, about 408 hours (204 replications × 2 hours) would have been needed. With the parallelization of about 60 jobs running simultaneously, it took about 6−10 hours to finish the whole simulation.

The content in the following sections can be viewed as a brief introduction to some options of the front end of efforts to implement a project that requires intense statistical computation. This material might motivate further reading, including documents available on the internet that are accessible through web searching.

4.1 Making an R Package for the LF Imputation Model

Since we want any program used to implement te LF Imputation Model be easy to translate for future applications, the way it is organized needs to be familiar to others. The open source R package is a standard way to implement statistical programs and to make them available to other researchers.

An R package consists of a set of R functions collected in a folder, with possible data and other files wrapped together. The folder containing the R functions is attached to the search path when the package is loaded. For example, we can run the following commands in an R session:

```
> search()  # look at the search path
[1] ".GlobalEnv"        "package:stats"     "package:graphics"
[4] "package:grDevices" "package:utils"     "package:datasets"
[7] "package:methods"   "Autoloads"         "package:base"

> library(norm)   # load the package norm
Warning message: ...

> search()  # look at the search path again, now "norm" is in the search path
[1] ".GlobalEnv"        "package:norm"      "package:stats"
[4] "package:graphics"  "package:grDevices" "package:utils"
[7] "package:datasets"  "package:methods"   "Autoloads"
[10] "package:base"

> ls(2)     # look into the loaded package norm, all functions of norm
package are located in this folder
[1] "da.norm"        "em.norm"        "getparam.norm"  "imp.norm"
[5] "loglik.norm"    "logpost.norm"   "makeparam.norm" "mda.norm"
[9] "mi.inference"   "ninvwish"       "prelim.norm"    "rngseed"
```

As we see from the above outputs, loading a package results in adding a folder into the search path, which itself is a list of folders. From that point forward, R can find the functions, e.g. "da.norm" by looking through the search path one by one, until "da.norm" is found in a folder in this list. It is possible that two functions in different packages happen to share the same name.

When both of the packages are loaded, the function loaded earlier is "masked". To refer to the "masked" function, one would need to use the expression "package::function" instead of the function name only. However, although this issue could arise in future uses of the package developed here, it did not compromise our present work.

Other actions before loading and after unloading of a package, like opening a data file or creating and deleting a temporary file, can also be specified.

In order to make a package, one can run the function package.skeleton() to generate the source code of an empty package in the current working folder before adding the actual program. The current working folder can be shown using the command

```
> getwd()
[1] "C:/Users/Xiang/Documents"
```

The source code folder contains sub-folders "R", "src" and "man" in which we put R programs, source files of low-level language programs if any, and help files respectively. Once these contents are in place, it is possible to use the command

```
> install.packages("package_name", type = "source", repos = NULL)
```

Incorporating compiled low-level language programs if needed to create and to copy the package folder into the library installation folder. The package installation folder can be seen by using the command

```
> .libPaths()
[1] "C:/Program Files/R/R-2.15.0/library"
```

The above folder information can be found in the "R Environment Variables" section of related R documents. By clarifying the organization of packages and letting every user follow the same location conventions, members of the R community can exchange programs which work together with a minimal chance of any conflict arising between shared programs.

4.2 Using a Cluster for Parallel Computing

To evaluate the LF Imputation Model in a parallel computing environment, we divide a large computational task into smaller independent pieces (jobs) which can be run in any order and hence can be parallelized. One can then submit these jobs to a parallel-computing cluster and wait until they are all done.

On the service side, a cluster manages a set of computers which keep running jobs. Drawing on a built-in queuing system, jobs from the queue are passed to idle computers until either all computers are in use or the queue is empty. Since the jobs are run by a number of computers in parallel, the running time (wall-clock time) is reduced. Another reason why a cluster of computers with a queuing system is appealing is that the system can be shared by multiple users, which lowers the total cost of computing for a group of users.

To carry out parallel computing using a cluster, a user first cuts his or her computation into small jobs. The way this is done might differ for each different computational problem. The computation involved in the current project ran an R program calling C++ functions 204 times with different seeds submitted to a random-number generator used as input. Letting each job running the R program one time, with different seeds generating random numbers, a sequence of 204 jobs working together comprised the computation needed in this project.

We use a script “R.q” to build a command file (.cmd) for the R program. A complete list of available scientific computing software and corresponding instructions is available in the document "Hoffman 2 software" available at

http://www.hoffman2.idre.ucla.edu/software/ .

The “R.q” script asks the user a number of questions interactively, including what R program to run, what arguments to pass, how much time to allow at most, how much memory is needed at most, and what notification, if any, is needed after finishing. The answers to these questions are collectively written into a command file containing all information needed by a computer to run the given job. This command file is then submitted by the user with the command “qsub”.

Instead of generating and submitting command files for each of the 204 replications, we automated the process by saving one single .cmd file and writing an R function qsub() to repeatedly edit a line

```
Rscript --default-packages=pkg -e 'cmd1' -e 'cmd2' ...
```

in the command file before submitting. Here, the arguments “pkg”, “cmd1”, “cmd2”, and so on contain the arguments passed to the R function qsub(). The option “-e” means to execute the subsequent command, while the option “--default-packages=pkg” means to load the contents of “pkg” before executing. The arguments “cmd1”, “cmd2”, and so on are executable R programs that can be run one by one in R. By putting different Rscript lines into a set of command files and then submitting them, we can parallelize R programs using a cluster.

There are other ways to structure parallel computing through packages, such as with the packages MPI and OpenMP, in which users start multiple processes or threads within a program. Users have more responsibility to achieve parallelization with these packages, which provide options to deal with the complexities involved in communications within the cluster and barriers among active processes, thereby providing users with finer control of parallelization. Our project did not require parallelization to this level.

4.3 Using the R−C/C++ Interface

R is a free statistical software package based on the S-Plus language, which is an interpreted language in the sense that each statement in the program is processed right before execution. The interpretation results in worse performance in terms of computational efficiency than is attainable on a computer in exchange for greater simplicity and convenience. Programmers appreciate the flexibility and compatibility of the S-Plus language, for routine tasks but favor low-level languages for their performance in situations requiring high computational intensity.

R provides a framework for calling a program written in low-level language from within R to perform computationally-intensive tasks. Structured this way, the low-level language components have full access to any type of input data objects, and have the ability generating any type of output data objects. Although the low-level language instructions have the ability to access all objects in the R environment, cross-language reference is not typically recommended as a programming strategy to avoid possible chaotic interactions. Similar situation is sometimes seen when programs induce excessive communications using global variables. The motivation for dividing tasks here, namely an efficiency tradeoff, justifies the minimally elevated risk of excessive communication between program components.

After extracting data from R objects, a low-level language program can carry out the actual computation. Then the results are "copied" and "pasted" to the resulting data objects in the shared memory, and passed back to R.

Below is an illustrative example of a C++ program to calculate the sum of all of the elements of a matrix. This example incorporates all the operations mentioned above, clarifying what is needed in a program written in low-level language and integrated with R. The program

developed for this project to implement the LF Imputation Model is analogous, although longer in each step is more involved.

```
// Include header files to use the Scythe Statistical Library
#include <scythestat/matrix.h>
...
// Include header files to access R objects
#include <R.h>
...
// use "scythe" (Scythe Library) namespace as default
using namespace scythe;


// Specify C style arguments passing in functions, required by R
extern "C"
{
      SEXP matsum(SEXP mat)
            // SEXP stands for "simple expression", which can
            // be any type of object in R
      {
            // Find nrow, ncol and elements of the matrix
            SEXP dim = getAttrib( mat, R_DimSymbol ) ;
            int nrow = INTEGER(dim)[0];
            int ncol = INTEGER(dim)[1];
            double *vec = REAL(mat);

            // Construct a matrix (Scythe Library matrix object)
            Matrix<> A (nrow, ncol, vec);

            // Calculate the sum of all numbers in the matrix
            double s = sum(A);

            // Prepare return value, a vector with length=1
            SEXP Rval;
            PROTECT( Rval = allocVector(REALSXP,1) );
            REAL(Rval)[0]=s;
            UNPROTECT(1);

            return Rval;
      }
}
```

Within R, one can use the following function to encompass this C++ function.

```
matsum<-function(x)
{
      x1<-x
      storage.mode(x1)<-"double"
      .Call("matsum",PACKAGE="blfmi",mat=x1)
}
```

This R function first makes sure that the input matrix consists of double-precision floating point numbers, changing inputted values to double-precision if necessary. Then the matrix is passed to the C++ function. After the operations in the C++ function are done, the R function returns whatever the C++ program returns to it. The reason why we change the storage type is that R allows any storage type so long as the implied computations make sense, but the C++ program needs the data type be unambiguous and specified explicitly. This flexible feature of R partially accounts for why R programs run slower than programs written in low-level languages.

4.4 The Scythe Statistical Library

Another tool we found useful in this project is the Scythe Statistical Library (Pemstein, Quinn and Martin 2011), an open source C++ library for statistical computing. This library includes a suite of matrix functions, pseudo-random number generators, and numerical optimization routines. Based on the author's test, the Scythe carried out a bootstrap computation faster than R by a factor of about 10 times (3.6 seconds versus 40 seconds).

To use the Scythe Statistical Library, one needs to download the library's source code, including a set of header files, into a folder. When compiling or building an R package, one uses

the -I option to add that folder into the "include path" so that the #include part of a C++ program can find those files. Adapting to the local programming environment, we put the line

PKG_CPPFLAGS = -I ~/include

in the file ~/.R/Makevars. As a result, the C++ compiler would look into the ~/include folder for the header files in building an R package.

As a brief explanation of the example program, the class “Matrix” and the function “sum” shown in the C++ program are defined in the Scythe Statistical Library. The "namespace" line makes it straight ward to refer to them; otherwise, we would need to introduce class and function names by their package name every time, like “scythe.Matrix” and “scythe.sum”, to allow them to be recognized by the compiler. The ensemble of R as a programming interface, C++ as a low-level language, the Scythe Statistical Library as an efficient toolkit for familiar statistically-relevant computational tasks, and the Hoffman 2 parallel computing environment thus provided a powerful engine for carrying at multiple imputation for high-dimensional multivariate incomplete data sets with mixed data types and longitudinal structure.

Chapter 5
Simulation Studies

The goal of multiple imputation (MI) is to recover information in an incomplete data set by drawing on available information, usually based on associations between variables allowing one to model the relationship between the missing data and the observed data, but also based on information in prior specifications. By fitting a Bayesian model encompassing the missing data, one can obtain the posterior predictive distribution of the missing data given observed data. One can then draw a number of samples from this distribution to serve as multiple imputations, from which incomplete-data inferences can be made through a combined statistical analysis. In considering the properties of statistical procedures, difficulties can arise from flaws in parametric models, small sample size, complications in the data structure, and other features of the problem. In this section, we consider whether such difficulties lead to deficiencies in the newly developed LF Imputation Model. The sample size, the number of variables, the number of time points, and the dimension of the factor scores are expected to be relevant in this sense.

Besides the LF Imputation Model, we also tested the PAN and MICE imputation methods in the same way to compare their performance.

5.1 Simulation Scenarios and Replications

We evaluate the LF Imputation Model through the performance of two statistical analyses applied to multiply imputed data sets. We consider 32 different scenarios, constructed as

factorial combinations of the number of time points (4 or 5), the number of outcome variables (10, 20, 40, or 60, with half of the variables continuous and the other half ordinal), the dimension of the underlying factor scores (3 or 5), and the number of subjects (50 or 200). Two hypothetical missing-data mechanisms, both of which belong to the class of MAR mechanisms, are also considered.

In each **scenario**, which is defined as a combination of the number of time points, the number of outcome variables, the dimension of factor scores, and the number of subjects, we first draw the underlying parameters from their specified distributions, including the fixed-effect coefficients, loading matrix, and variances of the observation-level errors or random coefficients.

The underlying fixed effects are drawn from independent uniform distributions on the interval $(-1,1)$. The variance-covariance matrices of random effects are drawn from scaled Wishart distributions independently with scale parameter

$$\begin{bmatrix} .16 & -.1 \\ -.1 & .16 \end{bmatrix}$$

and 20 degrees of freedom for all factor scores. The autoregressive parameter ρ is drawn from a normal distribution with mean .2 and standard deviation .05. The underlying loading matrix of the factor analysis model comes from drawing standard normal distributions for each element independently. Covariate matrices are taken to have elements that are independent with standard normal distributions. The variances of error terms of the continuous outcomes in the factor analysis model are drawn from a normal distribution with mean .15 and standard deviation .03, making the variances of error terms roughly comparable to the variances of the random effects. The proportions for the 3-category ordinal variables are drawn from a Dirichlet distribution with shape parameters (10, 10, 10), with cut-points that are set correspondingly. Taken together,

considering the combination of the variances of the random effects and the variances of the error terms, the magnitudes of the fixed effects range from small to large effect sizes. Properties of the procedure, such as coverage of interval estimates, are provided across a range of effect-size scenarios.

After the parameters of the model are determined, we take 4000 independent draws from the error distribution of the data, which includes the error terms in the factor analysis model and random coefficients, making use of variance parameters previously drawn. By putting the common model parameters and the independent draws of error terms together, we thus obtain 4000 independent data sets sharing one single model parameter for a scenario. In generating data, we keep the factor loading matrix unchanged across all time points as is assumed in the LF Imputation Model.

Here, one scenario corresponds to 4000 independent data sets sharing a common model parameter. We sometimes refer to them as **replications** since analyses on these data sets can be viewed as repeated efforts seeking insight regarding the same underlying scenarios. When interval estimates are generated, some replications give intervals that cover the underlying true parameter values while others do not, allowing us to assess the actual coverage rate. Besides the actual coverage rate, two other quantities we use for the purpose of evaluating the respective imputation procedures are the average lengths of interval estimates and the average biases of point estimates compared to the underlying true values.

It is reasonable to worry that data sets sharing one particular set of model parameters would not be sufficient to expose potential problems of a statistical procedure, because those model parameters might combine to mask problems that would be revealed with different parameter values. The performance results based on 32 scenarios, each of which corresponds to a

different draw of model parameters, helps address this concern. We also changed the random seed for generating the model parameters to verify that good performance results are repeatable instead of just representing "lucky" results in terms of problems being masked by possible interaction between underlying parameters and scenario settings.

In a given scenario, we apply each of the two statistical analyses, one focused on continuous variables and the other focused on categorical variables, on the 4000 complete data sets to approximate the performance measures one would obtain with an infinite number of replications. We generate missing data based on the first 204 complete data sets, targeting a margin of error of $1.96 \times \sqrt{.95 \times (1 - .95) / 204} \approx 3\%$ for estimates of the 95% coverage rate. Incomplete-data analyses applied to these 204 replications generate the observed coverage rates, the average lengths of intervals, and the average biases in parameters. Since the incomplete-data analyses we use here are based on standard complete-data analyses following different approaches to handling missing data (including available-case analysis as well as multiple imputation based on either the LF Imputation Model, the PAN imputation model supplemented with rounding for categorical values, or the MICE imputation procedure) the performance evaluation thus provides a foundation to contrast these alternative ways of handling missing data.

Since two missing-data mechanisms (described in following section) and 32 scenarios are considered, the above performance-evaluation procedure is repeated across the $2 \times 32 = 64$ cases or combinations.

5.2 Missing-Data Mechanisms in the Simulation

Generally any non-MCAR missing-data mechanism gives rise to the concern that statistical analyses are apt to be invalidated by some association between missingness and the underlying missing values on the variable of interest. We focus here on MAR mechanisms, which are likely to give rise to adverse impact on statistical analyses if the missing-data problem is not properly addressed and where it is possible to recover information using available data.

The two hypothetical missing-data mechanisms we consider here, both of which are MAR, will be described in following paragraphs. Thanks to the functionality of the R software package, other investigators can have access to the R package developed for this project and can modify the simulation program to evaluate the performance of the LF Imputation Model under an alternative missing-data mechanism. Importantly, the results presented here should not be understood as covering all MAR mechanisms; rather, these mechanisms considered here can be viewed as offering insight relevant plausible MAR mechanisms although perhaps not all MAR mechanisms.

In missing-data mechanism 1 (M1), letting Y_j be the j th outcome variable, the first K/2-1 continuous variables $Y_1 \cdots Y_{K/2-1}$ and ordinal variables $Y_{K/2+1} \cdots Y_{K-1}$ are subject to missing-completely-at-random missingness for each item following a Bernuolli coin flip with 20% probability of missingness. For the missingness of $Y_{K/2}$ and Y_K we use a logistic-regression model conditioned on the other continuous variables, and similarly, for the last ordinal variable, we follow the framework introduced by Wang (2002) and use a logistic-regression model conditioned on the other ordinal variables. Specifically, let

$$P(y_{1it} = \text{missing}) = .2\,, \quad \ldots \quad P(y_{K/2-1\,it} = \text{missing}) = .2\,,$$

$$P(y_{K/2+1\,it} = \text{missing}) = .2\,, \quad \ldots \quad P(y_{K-1\,it} = \text{missing}) = .2\,,$$

while

$$P(y_{K/2\,it} = \text{missing}) = r_1 \cdot n \cdot T \cdot \frac{\exp(a_1 y_{1it} + \cdots a_{K/2-1} y_{K/2-1\,it}) / (1 + \exp(a_1 y_{1\,it} + \cdots a_{K/2-1} y_{K/2-1\,it}))}{\sum_{i,t} \exp(a_1 y_{1it} + \cdots a_{K/2-1} y_{K/2-1\,it}) / (1 + \exp(a_1 y_{1\,it} + \cdots a_{K/2-1} y_{K/2-1\,it}))} \quad (5.1)$$

$$P(y_{K\,it} = \text{missing}) = r_2 \cdot n \cdot T \cdot \frac{\exp(b_1 y_{K/2+1\,it} + \cdots b_{K/2-1} y_{K-1\,it}) / (1 + \exp(b_1 y_{K/2+1\,it} + \cdots b_{K/2-1} y_{K-1\,it}))}{\sum_{i,t} \exp(b_1 y_{K/2+1\,it} + \cdots b_{K/2-1} y_{K-1\,it}) / (1 + \exp(b_1 y_{K/2+1\,it} + \cdots b_{K/2-1} y_{K-1\,it}))} \quad (5.2)$$

For the values of *a*'s and *b*'s, which determine the association between underlying values of variables and missingness rates, we first used absolute values of draws from a standard normal distribution to make them all positive invoking the framework used by Song and Belin (2004). However, in preliminary assessments, we found that such an approach does not generate substantial bias for available-case analysis in all scenarios. The apparent explanation for why available-case analysis does not break down in some scenarios has to do with our use of both positive and negative factor loadings in the factor-loading matrix incorporated in the LF Imputation Model, unlike the non-negative loadings (hence non-negative correlation coefficients among outcome variables) used Song and Belin (2004) in their factor-model framework. Hence, correlation coefficients among outcome variables can be either positive or negative and can cancel each other's effect in the logistic when the a's and b's are all positive. Although we do see good statistical properties of the LF Imputation Model when drawing *a*'s and *b*'s as absolute values of standard normal draws, the fact that available-case analysis doing a fair job in some scenarios raises a concern about judging the extent to which the LF imputation Model is really advantageous compared to simple-minded approaches that are not expected to perform well in general. For the purpose of demonstration, we use an artificial choice of *a*'s and *b*'s, which could

be taken as a "worst case scenario" for available-case analysis where all scenarios are affected by the missingness but where it is possible using MI analyses to obtain valid analysis results.

Specifically, we draw a's and b's from a standard normal distribution, then decide their signs based on the sign of corresponding correlation coefficient. This results in a setting where all terms in the logistic regression model contribute to missingness in the same direction, where higher values of covariates predict missingness to be more likely. This eliminates the possibility that the contribution from one term offsets the contribution from another.

In equations (5.1) and (5.2), we incorporate a scaling operation dividing the expit of a subject by the sum of the expits of all subjects into the calculation of missingness probabilities, in order to make the missingness rates equal to the target rates r_1 or r_2.

In the first missing-data mechanism (M1), we set r_1 and r_2 both equal to .2.

In the context of a basic mixed-model repeated-measures analysis in longitudinal research, Wang (2002) developed a related method and was able to show that inferences about cross-sectional means could be recovered accurately even if data were 100% missing. Here the inference relied on the trajectory assumption of the longitudinal data. The missing-data mechanism M2 used here is designed to investigate whether the LF Imputation Model can similarly recover information about a quantity of interest connected with a variable that is 100% missing.

In the second missing-data mechanism (M2), the approach remains the same except that we set r_1 and r_2 equal to .1, and then make the remaining variable of interest to be completely missing at the particular time point we analyze. Such a context can be thought of as representing a situation particular variable is not measured at a given time point.

5.3 Statistical Analyses Involved in Evaluating the Imputation Methods

At the analysis stage, we first consider the mean of the second-to-the-last observation of the K/2-th variable, which has a continuous distribution, to be the target quantity of interest.

The second analysis we consider is similar to the former except that it estimates the proportions of the K-th variable, a 3-category ordinal variable with p_1 and p_3 defining the distribution (the middle category having probability $1 - p_1 - p_3$). The reason why we use p_1 , p_3 instead of p_1 , p_2 is that the extreme categories are affected more by missingness than the middle category in the MAR missing-data mechanism we described in the previous section.

5.4 Simulation Results

After carrying out an analysis of the underlying complete data for reference, we assess the performance of available-case analysis, where we make use of all item data without further consideration of missing data, as well as the following three multiple-imputation approaches: the Longitudinal-Factor Imputation (LF), the PAN method based on the R program implementing a multivariate linear model longitudinal panel data, and the Multiple Imputation by Chained Equations method (MICE). Since the way MICE is implemented in R does not make it straightforward to handle the more standard "long-form" representation of multivariate longitudinal data with one row per time point, we use a "wide-form" representation of the data such that a row reflects a subject, which can be handled by MICE.

Across a total of 32 scenarios, available-case analysis is seriously biased. This is not surprising given the "worst case scenario" missingness mechanism we have deliberately chosen. The performance of the LF Imputation Model is valid across all the scenarios we tested in terms of reproducing a valid coverage rate with reasonable lengths of interval estimates. The performance of the MICE imputation method was not as good as expected, although in hindsight, this finding seems reasonable, knowing that MICE is not designed to apply on longitudinal data. In comparison, PAN performs generally well on the continuous outcomes when the program does not crash, although the PAN program often crashed when confronted with a large number of variables. Meanwhile, for the ordinal outcome variables, PAN imputation followed by rounding to the nearest integer generates invalid inferences in some scenarios in our test, since PAN is not designed based on a model of ordinal data.

As described above, our simulation shows the LF Imputation Model producing uniformly valid results across all scenarios tested. For brevity, since qualitative conclusions do not vary substantially across scenarios, we present a subset of the analysis results from our simulation evaluations, focusing on scenarios where other missing-data approaches yield invalid inferences.

5.4.1 Findings Based on Available-Case Analysis

Key simulation results are shown in Table 5.1 and 5.2. As shown in the columns corresponding to Available-Case Analysis, the sample mean of the continuous variable under analysis is biased toward lower values, consistent with the missingness mechanism inducing more missing data for higher values. The same reasoning explains a downward bias of p_3 and an upward bias of p_1 in the available-case analysis.

5.4.2 Findings Based on the LF Imputation Model

The LF Imputation Model works well for both continuous and ordinal variables in all of the scenarios we tested. With reference to analysis results based on the underlying complete data, the LF Imputation Model achieves valid coverage rates, negligible biases, and reasonable changes in interval lengths. The performance of the LF Imputation Model is either comparable or better than that of the other missing-data methods including Available-Case, PAN and MICE. This is not surprising, given that the data were generated to be consistent with an LF Imputation Model. Nevertheless, it is reassuring to know that the method is able to produce sensible inferences under the data model used to develop the method.

5.4.3 Findings Based on the PAN Imputation Model

Unlike available-case analysis and MICE, PAN has the same capacity in M2 to estimate the mean of a variable that is 100% as LF does, drawing in the same way on longitudinal patterns of associations. However, the PAN program crashes in the scenarios with both higher number of variables and lower number of subjects due to its inability to invert covariance matrices necessary to fit the PAN model. This generally happens in data sets with a large number of outcomes but low actual dimension, a combination that is a recipe for substantial collinearity. Across all of the replications within a scenario, we observed no exception in terms of whether PAN crashes or not, i.e., either all 204 replications crashed after some number of MCMC draws, or all 204 replications generated meaningful results within the 8000 draws we tried.

Table 5.1 shows the results of analyses of a continuous variable in the scenarios with 50 subjects, 5 time points, 5 factor scores, and the number of variables varying among 10, 20, 40 or 60. PAN provides valid inferences for the continuous variable in the scenarios with 10 or 20 variables, but crashes in scenarios with 40 or 60 variables.

We did not observe such crashes in the scenarios with 200 subjects, which is a sufficient number allowing PAN to produce valid estimates up to 60 variables. Since all inferences with 200 subjects are similar to those in Table 5.1 except for there have been no crashes of the PAN program, we did not show those results.

In summary, in the scenarios we considered, PAN yields valid statistical inferences for the mean of a 20% and 100% missing continuous variable so long as the program does not crash, with interval lengths that are slightly higher than but comparable to those produced in the same scenarios by the LF Imputation Model. A higher number of outcome variables and lower sample sizes were seen to be associated with crashes of the PAN procedure.

5.4.4 Findings Associated with Rounding for Ordinal Variables After Continuous Imputation

Table 5.2 shows the analysis of the proportions for an ordinal variable across scenarios with 200 subjects, 5 time points, 3 factors, and the number of variables varying among 10, 20, 40 and 60. With this number of subjects, PAN did not crash. However, the actual coverage rate of p_3 in the scenario with 40 variables was only 66%, substantially lower than the target rate of 95%.

Since PAN treats ordinal variables as continuous, it yields provisional imputed values that have decimal expansions (e.g., 1.35 or 2.56). Rounding to the nearest possible value after imputation is one way to deal with this problem. But such an approach might result in deficiencies in the imputation, given the ad-hoc nature of the procedure which is not backed by a well-motivated statistical model. We hypothesized that the categorical feature of the data gave rise to this problem of PAN imputation.

To verify our hypothesis, we replaced a continuous variable in the data with the underlying latent variable associated with an ordinal variable under analysis. We applied the corresponding missing-data indicator to the latent variable. Hence we were working with two versions of the same variable, one categorical and the other continuous, with exactly the same missing-data indicator and appearing in the same data set. We multiply imputed such a data set with PAN, where the same predicting information is used to produce the imputed values of these two versions of the variable. The analysis of the multiply imputed categorical variable yields a coverage rate of 58%, slightly worse than the previous finding of 66%. However, the analysis of the multiply imputed continuous version of the variable gave a valid coverage rate (98%). So PAN appears to impute the underlying latent variable appropriately while giving rise to deficiencies in the conversion of the underlying continuous measure into a categorical quantity.

This problem is related to non-linearity in the relationship between the numerical values of the categorical variable and the underlying latent variable. In other words, imputing the ordinal variable by its numerical value leads to a violation of the PAN model assumptions (linearity, and possibly conditional normality). As we know, a univariate transformation may either help or hurt linearity. A set of cut-points, can be used to categorize a continuous variable into a categorical variable, based on a discrete version of a transformation. In some situations,

these cut-points transform a linear variable (in relation to other variables) into a highly non-linear quantity. Putting the numerical values of a categorical variable into a linear regression or a multivariate normal model may produce misleading results in this case.

Another example similarly demonstrates how the strategy of imputing the values of a categorical variable by a multivariate normal model followed by rounding to the nearest integer value can lead to incorrect prediction, reinforcing the message that non-linearity induces a breakdown in this strategy. We use the NORM package, which like PAN relies on multivariate-normal assumptions that imply linear-regression relationships governing conditional distributions. We generated 420 independent standard normal random numbers and call them variable x_1; we use 280 1's, 60 2's and 80 3's as variable x_2. Here x_2 can be regarded as the categorized version of x_1 by some corresponding cut-points. The numerical values of x_2 are clearly non-linear with respect to x_1, in that, x_2 does not change in relation to x_1 over the greater part of the space, rather just increasing in relation to x_1 at the cut-points. We let the values of x_2 be missing with probability

$$\frac{\exp(1.4x_1 - 1.9)}{1 + \exp(1.4x_1 - 1.9)},$$

where the constants 1.4 and 1.9 yield a missingness rate about 20% for x_2. We let all values of x_1 be observed, and we assume that the missing-data mechanism is MAR. Then we use the package NORM, which assumes a multivariate normal distribution (bivariate here) to carry out multiple imputation followed by rounding to the nearest integer, as a foundation for a multiple-imputation analysis of proportions. As shown in Figure 5.4, the predictions based on the linear regression line, when using half integers (1.5 and 2.5 here) as cut-points, are incorrectly categorized to the middle (2nd) category in many imputations. We found the actual coverage rate of the interval

estimates for the proportion associated with the third category (the true rate being $80/420 \doteq .19$), to be only about 13%.

The bias related to the rounding-to-the-nearest integer after a normal-model imputation has long been noted. Horton, Lipsitz and Parzen (2003) predicted the amount of the bias related to rounding normal imputations to handle missing binary values. Song, Harrison, Hanson and Hall (2009) extended this idea to ordinal data. Bernaards, Belin and Schafer (2007) introduced an adaptive cut-off method to improve the imputation, followed by Demirtas (2010), who used a distance-based rounding approach for ordinal variables. A discrete model can also be involved in imputation of categorical variables. For example, the LF Imputation Model, which assumes a probit model and the MICE package (Van Buuren and Groothuis-Oudshoorn 2011) give users the capability of building such customized models. Vermunt (2008) and Gebregziabher and Desantis (2010) used latent-class-variable imputation, which is another way to deal with this problem.

5.4.5 Findings Associated with the MICE Imputation Model

The analysis based on the MICE imputation does not show as much flexibility to adapt to the scenarios we tested as LF and PAN do. This is due to the design of MICE not being tailored to the panel data. The R package of MICE does not support "long-form" panel data, nor does it make it straightforward to specify panel structure with the "wide-form" data.

In order to maintain the independence among the data rows, a user has to transpose a long-form data set into wide-form, which complicates the task of representing the panel structure covariance within a subject due to the repeated measurements. The corresponding imputation model hence assumes the existence of all possible covariance structures, allowing for distinct

correlations between any two variables and between any two time points. This may lead to a model with too many parameters to estimate. At the same time, the default version of the MICE model does not make any trajectory assumption. Thus, the multiple imputation of a multivariate longitudinal data set given by the current implementation of MICE picks up less information from the observed data than what is possible. Ideally, reasonable smoothness and exchangeability assumptions should be built into a model's implied conditional distributions; providing both flexibility and precision is a central challenge for any imputation procedure.

Table 5.1 Performance of alternative procedures for mean of variable K/2 among K=10,20,40, or 60 variables, scenarios with 50 subjects, 5 time points and 5 factors

Complete Data Analysis					*Available Case Analysis*			*Longitudinal Factor Model (LF)*			*Linear Model for Panel Data (PAN)*			*Chained Equation (MICE)*		
					K	**Bias**	**Len**	**Cvrg**	**Bias**	**Len**	**Cvrg**	**Bias**	**Len**	**Cvrg**	**Bias**	**Len**	**Cvrg**
				M1	10	-.49	2.3	.87	.04	1.9	.96	-.04	2.0	.96	-.59	2.9	.89
					20	-.40	1.6	.84	-.04	1.4	.96	-.04	1.5	.96	-.17	2.0	.95
K	**Bias**	**Len**	**Cvrg**		40	-.36	1.4	.76	-.004	1.1	.97	-	-	-	-.10	1.1	.89
10	.05	1.9	.95		60	-1.4	2.7	.47	.02	2.7	.93	-	-	-	-.05	1.3	.88
20	-.03	1.4	.94														
40	.007	1.1	.93	**M2**	10	-	-	-	.04	2.1	.95	.05	2.2	.97	-	-	-
60	.03	2.7	.93		20	-	-	-	-.05	1.6	.97	-.04	1.7	.98	-	-	-
					40	-	-	-	-.006	1.3	.97	-	-	-	-	-	-
					60	-	-	-	.01	2.8	.93	-	-	-	-	-	-

Note: M1 refers to missingness based on a logistic regression with coefficients unfavorable to available case analysis.
M2 refers to missingness based on M1 plus observations of outcome variables in our analysis missing 100%.
Bias refers to average bias, across 204 replications.
Len refers to average width of 95% interval across 204 replications.
Cvrg refers to the actual coverage rate of 95% interval across 204 replications (implying margin of error of ±3%)
"-" refers to findings that program crashed or did not produce usable result.

Table 5.2 Performance of alternative procedures for p_1 and p_3 of ordinal variable K among K=10,20,40, or 60 variables scenario with 200 subjects, 5 time points and 3 factors

Complete Data Analysis					*Available Case Analysis*			*Longitudinal Factor Model (LF)*			*Linear Model with Paneled Data (PAN)*			*MICE*		
				K	**Bias**	**Length**	**Cvrg**	**Bias**	**Length**	**Cvrg**	**Bias**	**Length**	**Cvrg**	**Bias**	**Length**	**Cvrg**
				M1	.08	.16	.39	.005	.13	.99	.002	.12	.99	.003	.14	.97
				10	-.14	.15	.01	-.008	.15	.99	-.03	.15	.91	-.008	.14	.97
					.04	.17	.71	.009	.15	.99	-.0001	.14	.99	.003	.13	.98
				20	-.13	.15	.22	-.009	.13	.98	-.03	.15	.97	-.007	.14	.98
K	**Bias**	**Length**	**Cvrg**		.08	.11	.62	.005	.13	1	.008	.14	.97	.04	.15	.92
				40	-.08	.18	.48	-.01	.15	.98	-.05	.17	.85	-.02	.16	.95
10	-.001	.12	.97		.1	.15	.30	.002	.12	.98	-.004	.12	.98	.06	.16	.71
	-.001	.14	.98	60	-.14	.14	.05	-.003	.14	.97	-.04	.15	.89	-.03	.15	.93
20	-.001	.13	.91													
	.005	.14	.98													
40	.003	.13	.97					-.002	.17	1	-.01	.16	.99			
	-.002	.13	.99	**M2**	10	-	-	-.004	.20	1	-.05	.18	.93	-	-	-
60	-.003	.12	.99					-.003	.23	1	-.03	.18	.99			
	.0006	.14	.97		20	-	-	.02	.19	1	-.05	.18	.90	-	-	-
								.001	.20	1	-.03	.12	.96			
				40	-	-	-	-.01	.20	1	-.08	.19	.66	-	-	-
								-.003	.16	1	-.007	.17	1			
				60	-	-	-	-.002	.17	1	-.05	.18	.93	-	-	-

Note: M1 refers to the missing mechanism based on a logistic regression with coefficients unfavorable to available case analysis, M2 refers to the missing mechanism based on M1 plus observations of outcome variables in our analysis missing 100%, see Section 5.2.

Bias refers to average bias, across 204 replications.

Length refers to average width of 95% interval, across 204 replications.

Cvrg refers to the actual coverage rate of 95% interval, across 204 replications (implying margin of error of ±3%).

"-" refers to findings that program crashed or did not produce usable result.

Table 5.3 Analysis under different combinations of the assumed number of factors and the true number of factors

Assumed No. of factors	True No. of factors=3					True No. of factors=5				
	No. of scenarios failed* for μ	Relative % length of CI** of μ	No. of scenarios failed*** for p_1 and p_3	Relative % length of CI of p_1 and p_3		No. of scenarios failed for μ	Relative % length of CI of μ	No. of scenarios failed for p_1 and p_3	Relative % length of CI of p_1 and p_3	
1	12	+14 (-6,+46)	6	+9 (+1,+24)	+4 (-2,+15)					
2	3	+3 (-4,+23)	0	+3 (-0,+18)	+2 (-2,+8)					
3	0	–	0	–	–	3	+6 (-1,+21)	1	+4 (-3,+12)	+2 (-6,+12)
4	0	-0 (-2,+3)	0	+0 (-4,+6)	+0 (-3,+3)	1	+3 (-3,+14)	0	+1 (-6,+6)	+1 (-3,+7)
5	0	+0 (-2,+4)	0	-0 (-5,+2)	+0 (-3,+4)	0	–	0	–	–
6	0	+0 (-2,+6)	0	-0 (-3,+4)	+0 (-4,+7)	0	+1 (-4,+2)	0	+0 (-6,+6)	-0 (-5,+3)
7						0	+1 (-1,+3)	0	+0 (-7,+5)	-0 (-3,+4)
8						0	+1 (-1,+6)	0	+0 (-4,+4)	-0 (-3,+3)

* Fail is defined as a coverage rate < 90% in the analysis of μ in either missing mechanism.

** Average length of CI of 16 scenarios with both missing mechanism compared to that assuming the right number of factors.

*** Fail is defined as any coverage rate < 90% in the analysis of p_1 and p_3 in either missing mechanism.

5.5 Robustness to Violations of Assumptions

5.5.1 The Impact of a Difference Between the Assumed and True Number of Factors

In an application of the LF Imputation Model, the number of factors is generally not known. To deal with this issue, we explore what happens when the assumed number of factors is incorrect for the number of factors deliberately used in generating the data. Statistical analyses based on imputation from alternative approaches are listed side by side to see the implications of an incorrectly assumed number of factors.

Specifically, we considered examples ranging from having 2 fewer factors up to 3 more factors than the correct underlying number of factors being used in the generating the data. Overall, we found that incorrectly assuming more than the actual number of factors yielded acceptable imputations while incorrectly assuming fewer than the actual number gives rise to invalid inferences in many of our test scenarios. This insight could prove helpful in fixing the number of factors in a real application of the LF Imputation Model. For example, we can try different number of factors from low to high until we see a number from which qualitatively similar statistical inferences are achieved. Any number at that level or above it might then be a reasonable assumption for the assumed number of factors.

Table 5.3 shows the results of analyses of both continuous and ordinal variables under both correct and incorrect assumptions about the number of factors. We considered the true underlying number of factors being 3 and 5, separated in the left and right sections in the table. Each row within a section is generated based on 16 scenarios

(sample size either 50 or 200; number of variables either 10, 20, 40 or 60; number of time points either 4 or 5) and missing-data mechanism either M1 or M2. We counted the number of failed scenarios, i.e., scenarios in which invalid inferences are made in either M1 or M2. Here, an invalid inference is indicated when the actual coverage of any parameter in the analysis falls below 90%. We also compared the lengths of 95% interval estimates based on an incorrectly assumed number of factors to those based on a correctly assumed number of factors. The average, minimum, and maximum change in the length of the interval estimates over all our test scenarios are also reported.

Our tests show that assuming fewer than actual number of factors leads to invalid inferences in some scenarios. The further the departure from the true underlying number of factors, the worse the coverage tends to be, and the longer the typical lengths of 95% intervals at the same time.

Due to the conservativeness of the intervals we use in estimating proportions, fewer failures were observed in the ordinal variable analyses than in the continuous variable analysis.

Compared to the scenarios with 3 true factors, the scenarios with 5 true factors tended to be more robust in terms of the performance of the LF Imputation Model when we assume fewer than the correct number of factors by the same amount. For example, "under-assuming" the number of factors by 1 roughly corresponds to misrepresenting 1/3 of the factor model parameters when the true number of factors is 3, a more severe offset than misrepresenting 1/5 of the factor-model parameters when the true number of factors is 5.

5.5.2 The Impact of a Possible Underlying Quadratic Trend

The LF Imputation Model assumes factor scores to have a linear trajectory. In order to study the impact of a possible underlying quadratic trend in the imputed data, we set the parameters governing the linear contribution of visit time to factor scores to .5, and systematically vary the parameter governing the quadratic contribution of visit time to factor scores from -.4 to .4 in increments of .1. Since the linear coefficients are fixed to .5, the resulting data sets can be compared regarding the extent to which the model assumptions are violated. For example, when the quadratic coefficient is .4, the equation for the *k*-th factor scores can be written as

$$f_{itk} = \beta_{0k} + .5(t - \bar{t}) + .4(t - \bar{t})^2 + \beta_{3k} x_{i3} + \beta_{4k} x_{i4} + \varepsilon_{itk}$$

where t=1, 2, 3, 4, 5, and $\bar{t}$=3. Thus, the quadratic model is capable of generating values that depart substantially from a linear trend, with the ratio of the quadratic coefficient to the linear coefficient ranging up to .8 in absolute value. The question then becomes whether the linear model can adapt in some way to accommodate variation from the quadratic model.

By looking through out the analysis results of the 32 test scenarios, we generally find that scenarios with larger sample size are more vulnerable to violation of the linearity assumption. This might have been expected, as larger sample sizes yield more precise estimates and shorter interval lengths, so that a deviation from the assumed trend is easier to detect. Also, data associated with the missing-data mechanism M2, which corresponds to a missingness rate of 100% for variables at the visit time we analyze, is more vulnerable to violation of the linearity assumption. This makes sense, as the

subsequent analysis rely more heavily on imputed values that are based on invalid assumptions.

Figure 5.3A plots the average endpoint values from 95% interval estimates of the mean of the 5th variable, each based on data generated in the scenario with 5 time points, 10 outcome variables, 200 subjects, 5 factor scores with different quadratic coefficients, and with missing-data mechanism M2. The actual coverage rate remains above 90% until the quadratic coefficients reach -.4 and .4, at which point the coverage rate drops to 90% for -.4 and 88% for .4. Figure 5.3B plots the average endpoint values from 95% interval estimates of proportion p_1 associated with the 10th variable. The coverage rate remains above 90% for quadratic coefficients throughout the range of -.4 to .4. The analyses of the proportion p_3 yielded similar result as those for p_1.

We initially thought the results for the largest values of the quadratic coefficients might give rise to much worse coverage of parameter values. Here, the variance-covariance structure seems to enable the model to adapt somewhat to the quadratic pattern, thereby mitigating the bias that might otherwise have been expected.

Table 5.4 Average, minimum and maximum of values of $\left(\frac{\text{multiple-chain average interval length}}{\text{single-chain average interval length}} - 1\right) \times 100\%$ across 32 simulation scenarios

	Continuous	Ordinal
M1	0 (-1%,+1%)	0 (-2%,+3%)
M2	0 (-1%,+3%)	0 (-4%,+4%)

Figure 5.1 Convergence of the correlation parameter in LF model
true value is .2, start values are -.5, -.3, -.1, .1, .3, .5

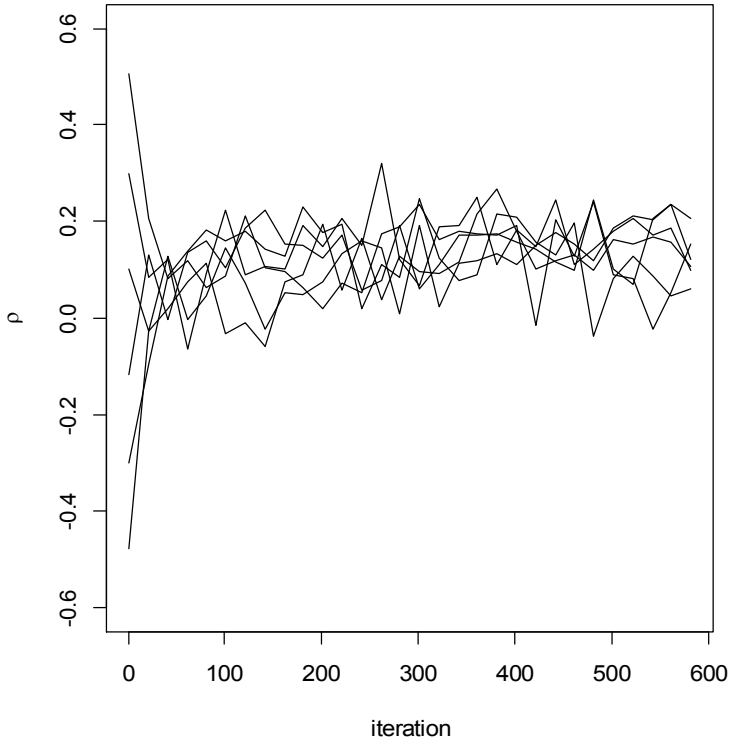
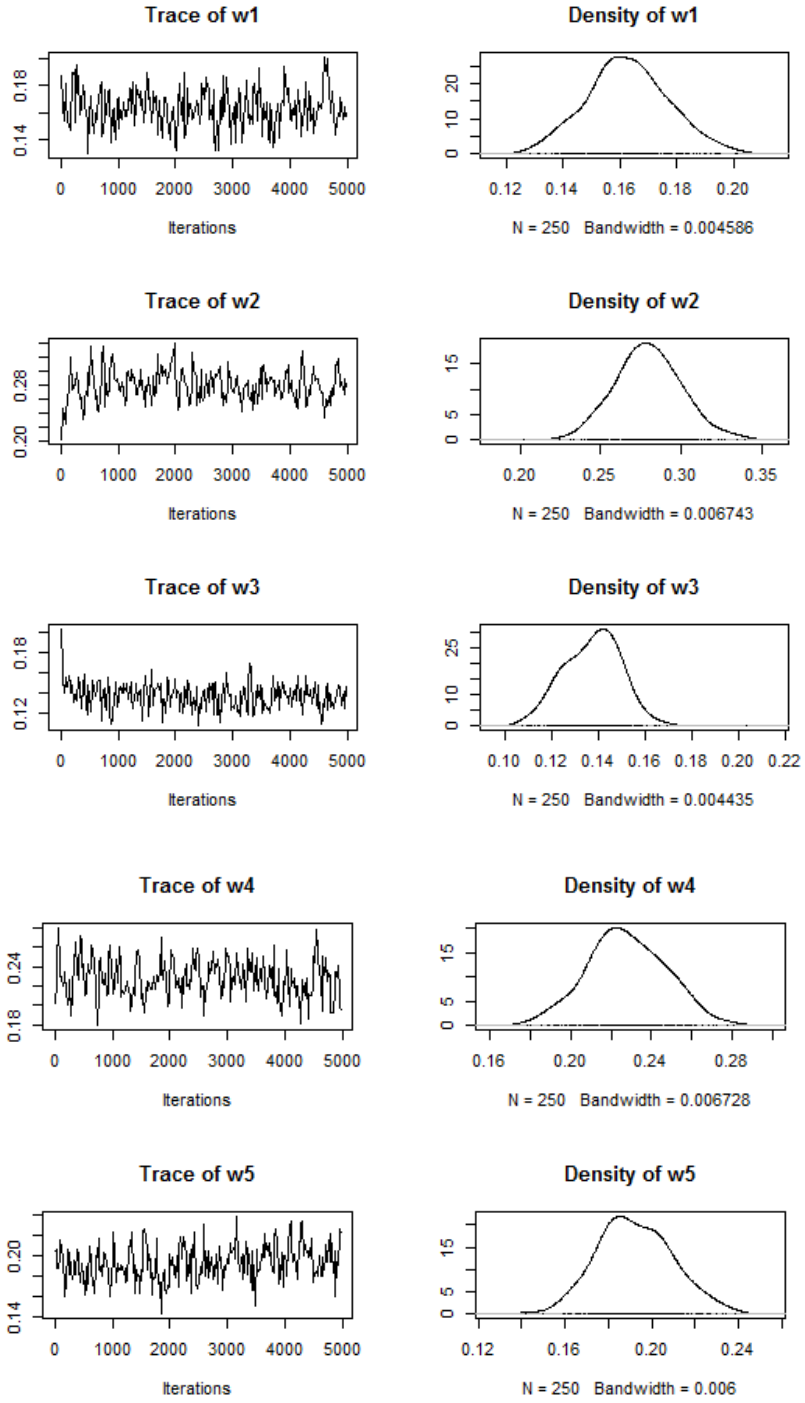


Figure 5.2 Convergence of weight parameters in LF model when true values are (.15,.25,.15,.25,.2) with starting values from (.2,.2,.2,.2,.2)



5.6 MCMC Convergence in the Model Fitting

We use Markov Chain Monte Carlo, specifically using a data-augmentation framework, to fit the LF Imputation Model while generating multiple imputations. Each data set corresponds to one Markov chain, resulting in 204 (replications) × 32 (scenarios) × 2 (missingness mechanisms) chains. As an example to illustrate the convergence properties of the procedure, we present results here from one replication in the scenario with 50 subjects, 20 variables, 3 factor scores, 5 time points, and missingness mechanism M1.

Figure 5.1 relates to the scenario where the true correlation between factor scores in adjacent time points is $\rho = .2$. We explored the impact of alternatively starting at -.5, -.3, -.1, .3 or .5. in the MCMC procedure. The Markov chains rapidly mix well and appear to stabilize around a 95% credible interval of (-.01,.26). The Gelman-Rubin potential scale reduction factor (Gelman and Rubin 1992) has a 95% upper limit of 1.01, which suggests satisfactory convergence for applied research.

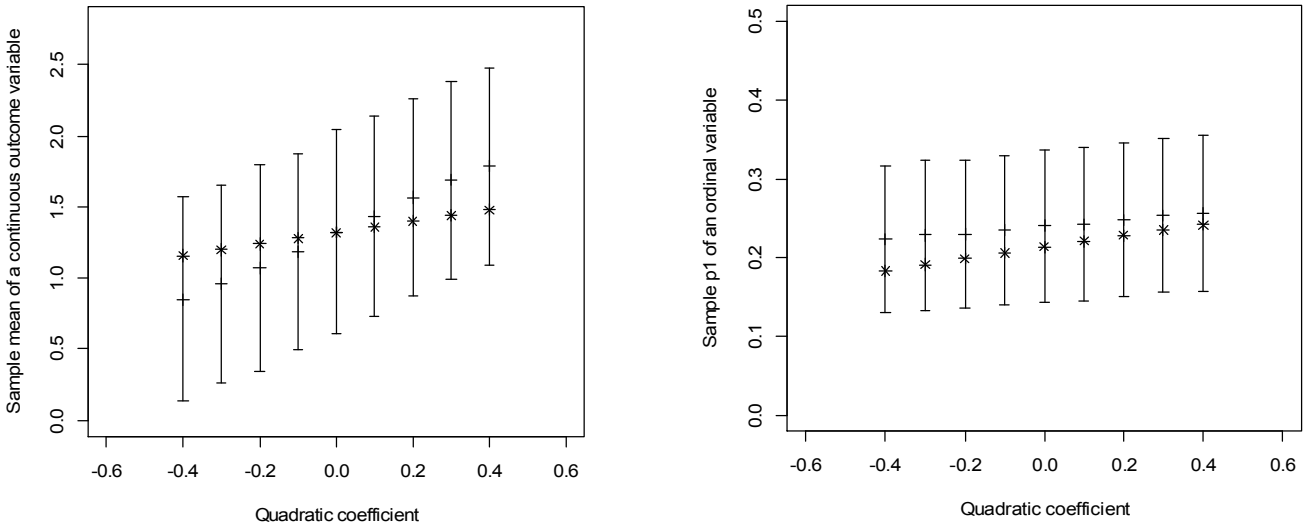
Another step in the MCMC estimation is to draw values from a weighted regression procedure. Starting from the null value (.2,.2,.2,.2,.2), the weights finally converge around the true values (.15,.25,.15,.25,.2) as shown in Figure 5.2.

Working with a version of a factor-analysis model where the outcome variables are not necessarily standardized, the factoring is being done on the covariance matrix rather than the correlation matrix of the multivariate observations at a given time point. This leads to the elements of the loading matrix possibly being outside of the range from -1 to 1. In an effort to start from an "over-dispersed" distribution, we used starting values

independently drawn from 3 discrete values, namely -2, 0 and 2, with equal probability for each free element in the loading matrix. Since the true values of the loading coefficients were drawn from a standard normal distribution, our choice of starting values makes the Markov chains start from possible values with considerable variation around the true values. We tested the 32 scenarios with 2 missing mechanisms using loading matrices with starting values chosen so that all parameters other than those whose convergence behavior was being studied were taken as starting from predicted null values, with the assumption regarding the number of factor scores to be used in the imputation being correct. As we hypothesized, the different starting points lead to convergence to the desired target distribution when the burn-in process is long enough.

Using 8000 iterations as a burn-in period for each chain, we found that analyses based on all chains are qualitatively identical in all scenarios for both of the missing-data mechanisms (M1 and M2). In Table 5.4, we summarized the changes of interval-estimate lengths (in percentages) comparing multiple-chain imputation and single-chain imputation across all 32 scenarios, across the two missing mechanisms (M1 and M2), and across the two types of outcomes being analyzed (continuous and ordinal). The averages of roughly 0 with ranges below ±4% in the changes of interval lengths in all four cells suggest that the single-chain imputations mixed well in convergence. Here multiple-chain imputation appropriately extends the interval estimates; on the other hand, the single-chain imputation we used in the simulations did not head to any invalid inferences regarding the factor structure.

Figure 5.3 Robustness of estimation to the quadratic coefficients where linear coefficient .5



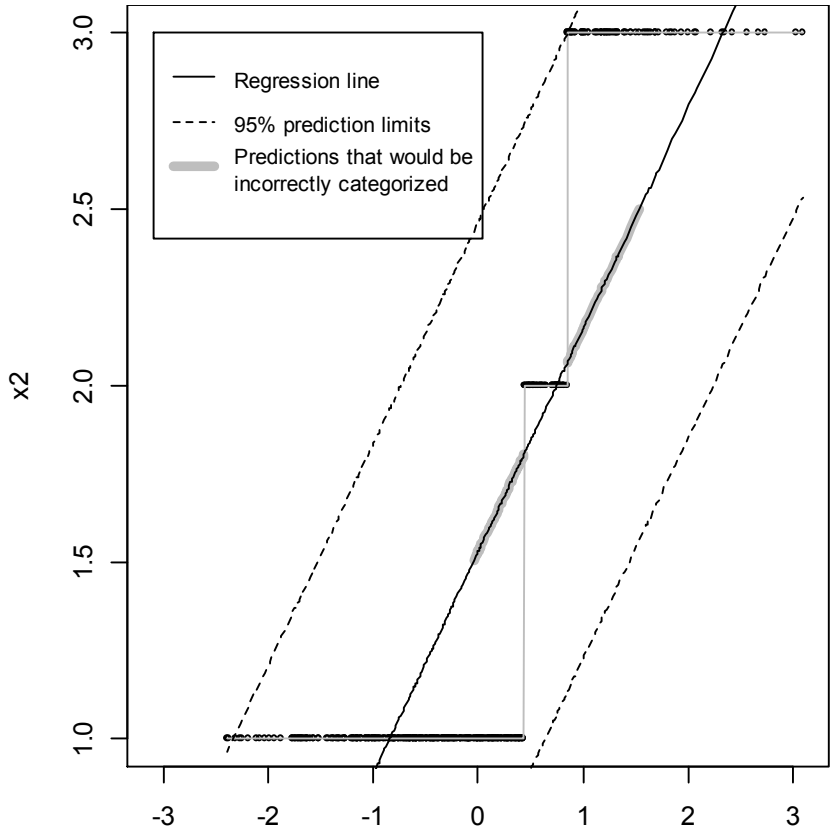
A B

A shows the average 95% confidence intervals of the mean of the 5th variable (continuous) using the LF Imputation Model assuming correct number of factors linear time trend only. The actual quadratic coefficient of time vary from -.4 to .4. The stars are approximate theoretical values based on 4000 true complete data with same parameters. The actual coverage rates drops below 90% only for the values -.4 and .4 of quadratic coefficients.

B shows the average 95% CI of p_1 of the 10th variable (ordinal), otherwise same as A. The actual coverage rates are above 90% for all values of quadratic coefficients.

In both A and B, M2 (100% missing of the variable in the analysis) of scenario with 5 time points, 10 outcome variables, 200 subjects and 5 factor scores is used, linear contribution of visit time are fixed to .5 for all 5 factors.

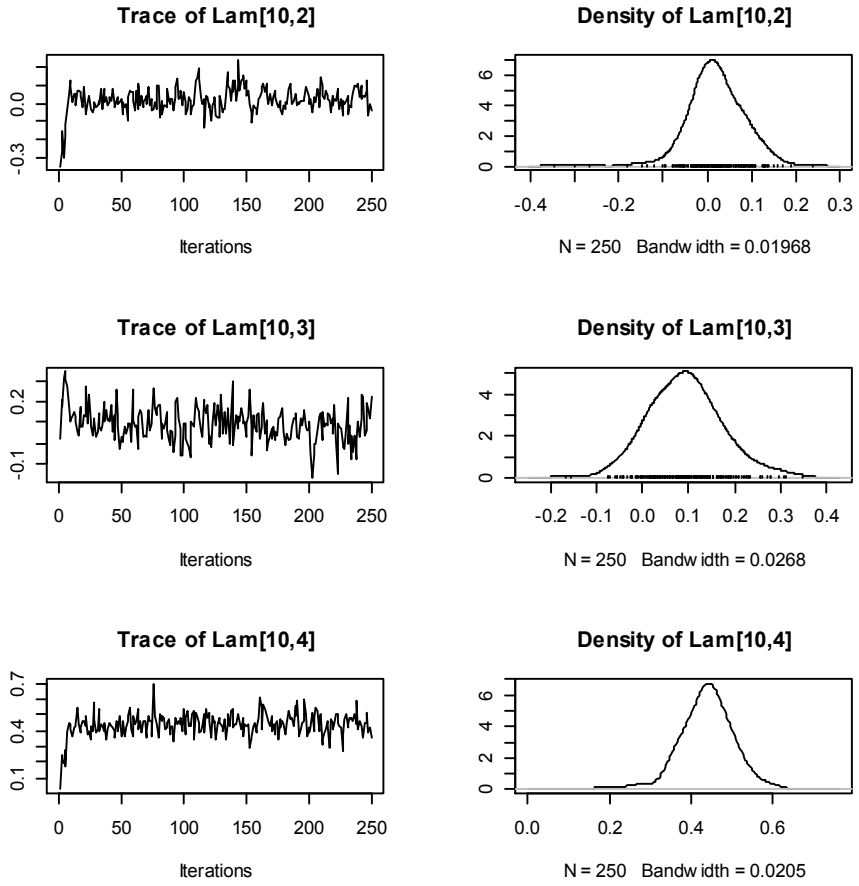
Figure 5.4 Non-linearity of an ordinal variable in relation to the latent variable



x_2 is the categorized version of x_1 by certain cut-points. A linear regression of x_2 on x_1 generates wrong prediction on the gray part if we round to the nearest integer.

Note, we connect the points or circles with gray vertical lines in order to help reader eliminate the perceptual illusion that the 3 categories are overlapped horizontally. Technically they do overlap by the diameter of a circle, but no more.

Figure 5.3 Convergence of loading coefficients for variable 10, the continuous variable under analysis, true value is (.09,.25,.41), start from (0,0,0)



Chapter 6
Application to Dentistry Data

As described in Chapter 1, a motivating example is provided by a dentistry data set from a study including 336 patients to compare two treatments for broken jaws. All patients are assigned to one of the two treatments: maxillomandibular fixation (MMF), or rigid internal fixation (RIF). The oral health of each patient is measured at 5 time points, namely just before discharge from the hospital, 10 days after discharge, 1 month after discharge, 6 months after discharge, and 12 months after discharge.

Because RIF is a surgical treatment, hence more expensive compared to MMF, patients are able to recover certain functions, such as eating more quickly than with MMF. Wang (2002) hypothesized that the RIF treatment group would show better recovery in terms of patients' general oral health assessment index (GOHAI), an oral-health quality-of-life composite measure derived from questionnaire items.

Due to the longitudinal nature of the study and some patients not being available at all visits, the data set is marked by considerable missingness. For example, the missingness rates of GOHAI measurements across the 5 visits are 2.1%, 14%, 14.6%, 43.2%, and 41.4%, which are high enough to induce concerns for some analyses. Using only subjects with 5 complete measurements, or using only available cases in a model that does not require subjects to have all measurements, there is some possibility that meaningful biases in estimates of quantities of interest would result due to selection effects. A sensible imputation strategy for the missing data could be expected to recover

the joint distribution based on associations among variables, resulting in analyses being more trustworthy than those based on either complete-case or available-case approaches, especially when the missingness rate is not negligible.

6.1 Treatment Difference in GOHAI Trajectories During Recovery

An important question in the applied context was whether there is any treatment difference between the treatment groups in terms of the slopes relating changes in GOHAI to time across the period from the patient's discharge from the hospital to the patient's one-year follow-up assessment. The researchers hypothesized that there is some difference between treatment groups. Wang (2002) multiply imputed the GOHAI variable, which takes on integer values in the range from 0 to 100 but was analyzed as a continuous variable, using two different models: a multivariate normal model fitted with the program NORM (Schafer 1997) and a forerunner of the LF Imputation M developed by Wang (2002), like the model introduced by this report, incorporated both longitudinal and cross-sectional associations but where the ordinal variables are treated as continuous variables in that version, with rounding to the nearest possible value. The analysis results based on different ways of handling missing observations are summarized in Table 6.1. In line with Wang (2002), we assumed the number of factors to be equal to 3 in the newly developed LF Imputation Method.

Table 6.1 Linear Trajectory of GOHAI

	Available-case	*Norm*	*Wang (2002)*	*LF Imputation Model*
B_{00} (MMF)	28.55 (26.24 , 30.92)	29.30 (26.35 , 32.33)	28.90 (26.45 , 31.20)	28.72 (26.09 , 31.35)
B_{01} (RIF-MMF)	-.292 (-4.674 ,4.049)	-4.241 (-7.183 , -1.437)*	-3.932 (-5.271 , -1.946)*	-3.87 (-5.68 , -2.06)*
B_{10} (MMF)	7.074 (4.780 ,9.237)	6.147 (1.902 , 9.794)	6.571 (2.238 , 9.342)	6.26 (2.39 , 10.12)
B_{11} (RIF-MMF)	1.859 (-2.418 , 5.956)	2.721 (.202 , 5.381)*	2.690 (.916 , 5.017)*	2.38 (.35 , 4.41)*

Each cell contains Posterior Mean (CI)
* $p<.05$

Working with imputed data sets, a linear statistical model with different intercept and slope is used for the trajectory of GOHAI scores. In Table 6.1, there is a tendency for available-case analysis yield longer interval estimates for all parameters, while the analyses based on the multiple imputation strategy of Wang (2002) and LF Imputation Model correspond to shorter confidence intervals, with that corresponding to NORM-based multiple imputation falling in between. In summary, the multiple imputation results give rise to a significant non-zero slope difference, which is not reflected in the available-case results.

These results match with the finding of our simulations in Chapter 5: the available-case analyses correspond to longer interval estimates and lower coverage rates. The columns of Table 6.1 corresponding to the method of Wang (2002) and LF Imputation Model, which make use of the trajectory the longitudinal data, correspond to shorter interval estimates and coverage rate as claimed. The MICE approach (Van Buuren and Groothuis-Oudshoorn 2011), a sequential regression approach implemented in R, is not shown in here. Based on our test, MICE and NORM perform similarly in

terms of actual coverage rates and lengths of interval estimates. The specification of conditional steps in MICE accommodates a broader range of scenarios than those implied by NORM. In line with other evaluations (e.g., Tang, Song, Belin and Unützer 2005) we find the flexibility of conditional specifications of MICE provides robustness against departures from normality, emerging as preferable to NORM in situations where both models yield similar interval lengths.

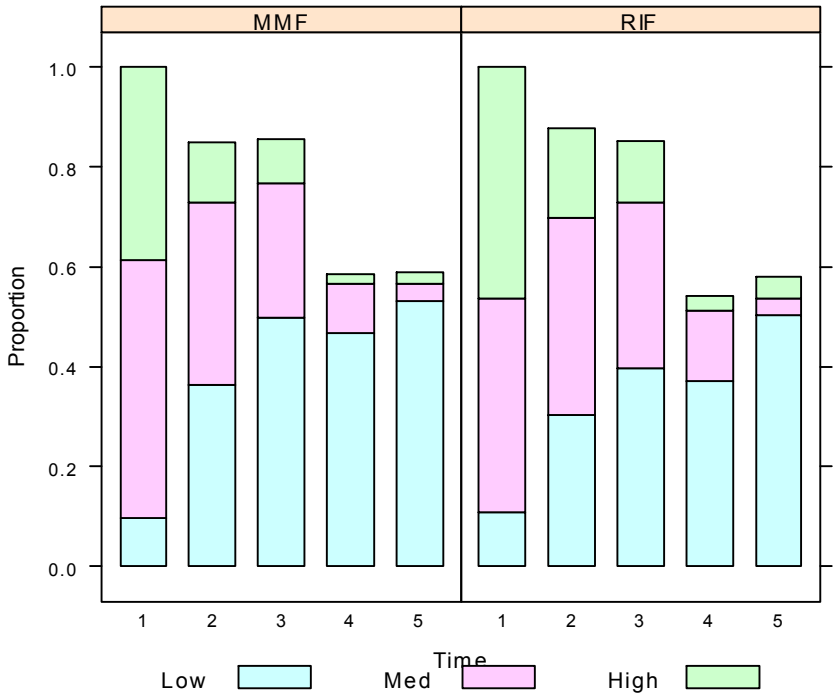
Our simulation results lend support to the analysis result of the dentistry data corresponding to the LF Imputation Model as being preferable to NORM or available-case analysis, in line with the conclusion of Wang (2002).

6.2 Treatment Difference in Patient Pain Levels

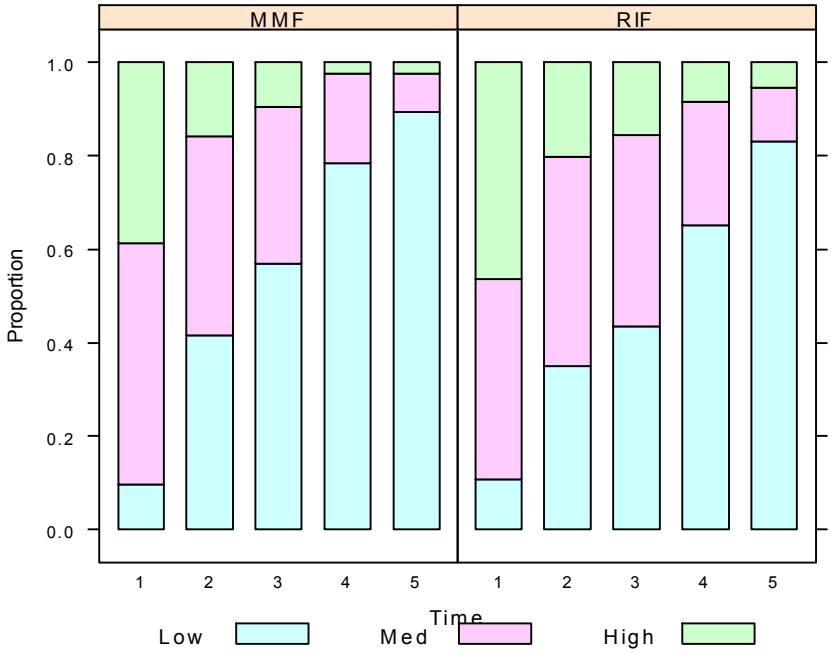
Another question of interest is whether the two treatment groups feel different pain levels during the course of the patients' recovery. We applied an ordinal logistic model on the pain level of a patient at each visit, after multiple imputation using the LF Imputation Model developed here.

Pain was originally measured on an integer scale from 0 – 10. As described in Chapter 1, we categorized the pain into three levels based on a desire to avoid either an 11-level categorical variable analysis, which seems too complicated as a starting point for the newly developed method, or adopting a normal assumption without considering problems associated with possible skewness.

Figure 6.1 Distribution of PAIN Level Before and After Imputation using LFA



A. Before imputation



B. After imputation

As shown in Figure 6.1 A, pain levels are completely observed in the initial visit of all patients, while considerable portions of patients are missing this variable in subsequent visits. Specifically, 14%, 14.6%, 43.2% and 41.4% of the pain measurements are missing across the 2nd through the 5th visits, respectively. The proportions of low, medium and high pain levels suggest a trend of patients generally experiencing lower pain levels over time, if we assume that high pain levels do not predominate in the missing observations. However, while prior knowledge suggests that patients who recover well are apt to be represented in the incomplete cases, prior knowledge also suggests that patients who are doing poorly might be disproportionally represented in the incomplete cases due to missed visits.

Figure 6.1 B shows the pain levels in one imputed data set based on the LF Imputation Model. The five data sets we prepared for a multiple imputation analysis using the LF Imputation Model differ little in their proportions, making the five box plots difficult to distinguish visually. In Table 6.2, we show the average proportions across imputations with standard errors obtained from multiple imputation frameworks.

Table 6.2 Proportions of Imputed PAIN Levels

		Visit 1	Visit 2	Visit 3	Visit 4	Visit 5
MMF	High	.41 (0)	.16 (.01)	.11 (.01)	.04 (.02)	.05 (.01)
	Med	.49 (0)	.45 (.01)	.37 (.02)	.22 (.03)	.07 (.02)
	Low	.10 (0)	.39 (.02)	.52 (.02)	.74 (.04)	.88 (.02)
RIF	High	.43 (0)	.16 (.02)	.13 (.01)	.03 (.01)	.03 (.01)
	Med	.47 (0)	.42 (.01)	.27 (.02)	.16 (.03)	.08 (.02)
	Low	.10 (0)	.42 (.01)	.61 (.02)	.80 (.03)	.89 (.02)

Average (SE based on to multiple imputation)

Figure 6.1 B shows a clear trend toward generally lower pain levels among patients. The imputation model incorporating factor analysis on all measured variables and a missing-at-random assumption for the missing-data mechanism suggests that the missing observations are mostly in the low and medium pain levels.

In order to answer the question whether the two treatments differ in patient pain levels, we fit an ordinal logistic model with a random effect using the package pglm (Paneled Generalized Linear Model) in R. The model incorporates a random intercept and a difference of intercepts for two treatments, as suggested in Chapter 1. As shown in Table 6.3, available-case analysis, MICE, and the LF Imputation Model all suggest significantly more pain associated with RIF.

Table 6.3 Ordinal Logistic Regression of PAIN

	Available-case	*MICE*	*LF Imputation Model*
B_{00} (MMF)	2.82 (2.45 , 3.18)	2.13 (1.61, 2.65）	2.21 (1.68 , 2.74)
B_{01} (RIF-MMF)	.42 (.11 , .74)*	.40 (.11, .68)*	.37 (.08 , .65)*
B_1 (slope)	-1.08 (-1.20 , -.96)*	-.93 (-1.04 , -.82)*	-.96 (-1.07 , -.85)*
Cut Points	0 (-) 2.32 (2.10 , 2.54)	0 (-) 2.21 (2.01 , 2.40)	0 (-) 2.21 (1.98 , 2.45)
SE of random intercept	1.23 (.93 , 1.53)	1.09 (.82 , 1.37)	1.04 (.72 , 1.38)

Chapter 7
Discussion and Future Research

Missing values sometimes lead inefficiency or bias in a statistical analysis when complete-case or available-case strategies are used. It is generally more desirable to address missing-data problems by applying multiple imputation to the full data set before fitting the statistical models for scientific research. Compared to the imputation based on part of the data set, imputation based on a full range of variables is more likely to carry the information needed to predict the missing values. In other words, the MAR assumption is apt to be more plausible on the full data set than on a subset of the data set.

We proposed an imputation method which is capable of handling continuous and ordinal variables to address the difficulty that mixed variable types may present in a multivariate data. However, there might be more data types like binary, nominal, and count variables. As proposed by Dunson and Herring (2005), a Baysian model assuming latent variables in an exponential family may be used to impute more types of variables.

The LF Imputation Model assumes the number of factors to be known. In Section 5.4, we showed that imputations assuming this number to be lower than the number leads to unstable results, while assuming this number to be slightly higher than the true value leads to reasonable results. It is difficult to determine the number of factors based on the analysis results. Song and Belin (2008) proposed to choose the number of factors based on the AIC and BIC in the fitting of the imputation model. The eigen-values and scree plots in their work show some evidence regarding the number of factors to use, although

finding the eigen-values might be challenging using a data set with missing values. It is also possible to add a reversible-jump MCMC step (Lopes and West 2004) to the imputation algorithm so that the number of factors can be fit in the context of an MCMC algorithm.

The LF Imputation Model assumes an AR(1) covariance structure of factor scores. In order to simplify the model, we assume all factor scores share the same correlation coefficient ρ. A natural extension is to assume different AR(1) structures, or as many ρ's as there are factors. Other covariance structures might be considered as well.

Unbalanced longitudinal data can be another challenge. The LF Imputation Model models the trend, the covariance within time points, and the covariance between time points separately, which makes it possible to produce imputations for time points where values are 100% missing. This feature can be used to handle unbalanced data at least to some extent. However, more work is needed if the degree to which the data are unbalanced is extreme.

The LF Imputation Model is closely related to the PAN model, which assumes a multivariate normal model on the data. LF is applicable to high-dimensional data, even with small numbers of subjects and high collinearity, while PAN crashes in these situations. The imputation of categorical variable is also improved since the LF Imputation Model assumes a probit link for latent variables underlying the categorical variables. There are ways to improve PAN such as incorporating a ridge prior in the error covariance. When multiple choices of imputation methods are available, researchers need to make their own decision. Alternatively, Siddique, Harel and Crespi (2012) proposed a nested imputation method to incorporate several plausible imputation models.

In some situations, the factor model itself is of interest to researchers. Erosheva and Curtis (2011) and Pape, Aßmann and Boysen-Hogrefe (2013) proposed relabeling algorithms, which may motivate further work to find suitable rotations for the series of draws of the loading matrices. Exploratory information of the full data set might be achieved this way to help make the factor structure interpretable.

In summary, factor-analysis methods proved useful in our investigation to help fit flexible models to highly multivariate data with a mix of data types. Future research could build on the methods developed here, including both MCMC methods to fit non-standard statistical models and parallel-computing methods to make model-fitting feasible. The ability to such method to large data sets and extend such method to large data sets and a broader mix of data types offers great promise for the future research.

References

Atchison KA, Dolan TA. Development of the geriatric oral health assessment index. *Journal of Dental Education* 1990; **54**: 680-687.

Belin TR, Hu MY, Young AS, Grusky O. Performance of a general location model with an ignorable missing‐data assumption in a multivariate mental health services study. *Stat Med* 1999; **18**: 3123-3135.

Bernaards CA, Belin TR, Schafer JL. Robustness of a multivariate normal approximation for imputation of incomplete binary data. *Stat Med* 2007; **26**: 1368-1382.

Casella G, Robert CP, Wells MT. Mixture models, latent variables and partitioned importance sampling. *Statistical Methodology* 2004; **1**: 1-18.

Chib S, Greenberg E. Analysis of multivariate probit models. *Biometrika* 1998; **85**: 347-361.

Collins LM, Schafer JL, Kam CM. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychol Methods* 2001; **6**: 330-351.

David M, Little RJ, Samuhel ME, Triest RK. Alternative methods for CPS income imputation. *Journal of the American Statistical Association* 1986; **81**: 29-41.

Demirtas H. A distance-based rounding strategy for post-imputation ordinal data. *Journal of Applied Statistics* 2010; **37**: 489-500.

Dunson DB, Herring AH. Bayesian latent variable models for mixed discrete outcomes. *Biostatistics* 2005; **6**: 11-25.

Erosheva EA, Curtis SM. Dealing with rotational invariance in Bayesian confirmatory factor analysis. *Technical Report 589, University of Washington* **2011**.

Gebregziabher M, Desantis SM. Latent class based multiple imputation approach for missing categorical data. *Journal of Statistical Planning and Inference* 2010; **140**: 3252-3262.

Gelfand AE, Smith AF. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* 1990; **85**: 398-409.

Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. *Statistical science* 1992: 457-472.

Gelman A, Speed T. Characterizing a joint probability distribution by conditionals. *Journal of the Royal Statistical Society. Series B (Methodological)* 1993: 185-188.

Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. *Bayesian data analysis.* Taylor & Francis, 2014.

Geweke J, Zhou G. Measuring the pricing error of the arbitrage pricing theory. *Review of Financial Studies* 1996; **9**: 557-587.

Heitjan DF, Little RJ. Multiple imputation for the fatal accident reporting system. *Applied Statistics* 1991: 13-29.

Hobert JP, Casella G. Functional compatibility, Markov chains, and Gibbs sampling with improper posteriors. *Journal of Computational and Graphical Statistics* 1998; **7**: 42-60.

Horton NJ, Lipsitz SR, Parzen M. A potential for bias when rounding in multiple imputation. *The American Statistician* 2003; **57**: 229-232.

Li K-H, Meng X-L, Raghunathan TE, Rubin DB. Significance levels from repeated p-values with multiply-imputed data. *Stat Sin* 1991; **1**: 65-92.

Li K-H, Raghunathan TE, Rubin DB. Large-sample significance levels from multiply imputed data using moment-based statistics and an F reference distribution. *Journal of the American Statistical Association* 1991; **86**: 1065-1073.

Lin S. *Joint Bayesian Modeling of Irregularly Measured Multivariate Longitudinal Nutrient Consumption and Longitudinal Outcome Data.* UNIVERSITY OF CALIFORNIA, LOS ANGELES, 2012.

Little RJ, Schluchter MD. Maximum likelihood estimation for mixed continuous and categorical data with missing values. *Biometrika* 1985; **72**: 497-512.

Liu J, Gelman A, Hill J, Su Y-S, Kropko J. On the stationary distribution of iterative imputations. *Biometrika* 2013: ast044.

Lopes HF, West M. Bayesian model assessment in factor analysis. *Stat Sin* 2004; **14**: 41-68.

Olkin I, Tate RF. Multivariate correlation models with mixed discrete and continuous variables. *The Annals of Mathematical Statistics* 1961: 448-465.

Pape M, Aßmann C, Boysen-Hogrefe J. The Directional Identification Problem in Bayesian Factor Analysis: An Ex-Post Approach. 2013.

Pemstein D, Quinn KM, Martin AD. The Scythe statistical library: An open source C++ library for statistical computation. *Journal of Statistical Software* 2011; **42**: 1-26.

Quinn KM. Bayesian factor analysis for mixed ordinal and continuous responses. *Political Analysis* 2004; **12**: 338-353.

Raghunathan TE, Lepkowski JM, Van Hoewyk J, Solenberger P. A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey methodology* 2001; **27**: 85-96.

Rubin DB. Inference and missing data. *Biometrika* 1976; **63**: 581-592.

Rubin DB Multiple imputations in sample surveys-a phenomenological Bayesian approach to nonresponse. *Proceedings of the survey research methods section of the American Statistical Association* 1978, American Statistical Association.

Rubin DB, Schenker N. Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association* 1986; **81**: 366-374.

Rubin DB. The calculation of posterior distributions by data augmentation: Comment: A noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are

modest: The SIR algorithm. *Journal of the American Statistical Association* 1987: 543-546.

Rubin DB. Multiple imputation after 18+ years. *Journal of the American Statistical Association* 1996; **91**: 473-489.

Saylor CF, Finch A, Baskin CH, Saylor CB, Darnell G, Furey W. Children's Depression Inventory: Investigation of procedures and correlates. *Journal of the American Academy of Child Psychiatry* 1984; **23**: 626-628.

Schafer JL. *Analysis of incomplete multivariate data.* CRC press, 1997.

Schenker N, Taylor JM. Partially parametric techniques for multiple imputation. *Computational Statistics & Data Analysis* 1996; **22**: 425-446.

Siddique J, Belin TR. Multiple imputation using an iterative hot‐deck with distance‐based donor selection. *Stat Med* 2008; **27**: 83-102.

Siddique J, Harel O. MIDAS: a SAS macro for multiple imputation using distance-aided selection of donors. *Journal of Statistical Software* 2009; **29**.

Song J, Belin TR. Imputation for incomplete high‐dimensional multivariate normal data using a common factor model. *Stat Med* 2004; **23**: 2827-2843.

Song J, Belin TR. Choosing an appropriate number of factors in factor analysis with incomplete data. *Computational Statistics & Data Analysis* 2008; **52**: 3560-3569.

Song R, Harrison KM, Hanson DL, Hall HI. Correction of Bias in Imputing Missing Values of Categorical Variables. *Communications in Statistics-Theory and Methods* 2009; **39**: 350-362.

Tang L, Song J, Belin TR, Unützer J. A comparison of imputation methods in a longitudinal randomized clinical trial. *Stat Med* 2005; **24**: 2111-2128.

Tanner MA, Wong WH. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association* 1987; **82**: 528-540.

Van Buuren S, Oudshoorn K. Flexible multivariate imputation by MICE. *Leiden, The Netherlands: TNO Prevention Center* 1999.

Van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software* 2011; **45**.

Vermunt JK. Latent class and finite mixture models for multilevel data sets. *Statistical Methods in Medical Research* 2008; **17**: 33-51.

Wang J. *Handling incomplete high dimensional multivariate longitudinal data by multiple imputation using a longitudinal factor analysis model.* UNIVERSITY OF CALIFORNIA, LOS ANGELES, 2002.

Ware Jr JE, Sherbourne CD. The MOS 36-item short-form health survey (SF-36): I. Conceptual framework and item selection. *Medical care* 1992: 473-483.

Zhang X, Boscardin WJ, Belin TR. Sampling correlation matrices in Bayesian models with correlated latent variables. *Journal of Computational and Graphical Statistics* 2006.

Zhang X, Boscardin WJ, Belin TR. Bayesian analysis of multivariate nominal measures using multivariate multinomial probit models. *Computational Statistics & Data Analysis* 2008; **52**: 3697-3708.