University of California

Los Angeles

Multiple Imputation of High-dimensional Mixed Incomplete Data

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Biostatistics

by

Ren He

2012

Abstract of the Dissertation

Multiple Imputation of High-dimensional Mixed Incomplete Data

by

Ren He

Doctor of Philosophy in Biostatistics

University of California, Los Angeles, 2012

Professor Thomas R. Belin, Chair

It is common in applied research to have large numbers of variables with mixed data types (continuous, binary, ordinal or nomial) measures on a modest number of cases. Also, even a simple imputation model can be overparameterized when the number of variables is moderately large. Finding a joint model to accommodate multivariate data with mixed data types is challenging. Here we develop two joint multiple imputation models. One is using multivariate normal components for continuous variables and latent-normal components for categorical variables. Following the strategy of Boscardin and Weiss (2003) and using Parameter-expanded Metropolis-Hastings estimation (Boscardin,Zhang and Belin 2008), we use a hierarchical prior for the covariance matrix centered around a parametric family. The second one is using a factor analysis model to impute missing items. It is an extension of Song and Belin (2004).

The report is organized as follows: Chapter 1 gives a brief introduction of the research problem. Chapter 2 lists the review of the background knowledge related to our two new approaches. We introduce two existing methods of handling high-dimensional continuous incomplete data in Chapter 3 and another two methods of handling mixed incomplete data in Chapter 4. Our newly developed methods are outlined in Chapter 5. In Chapter 6, simulations under various conditions are carried out to compare the results based on our approaches with the results from the rounding method (Bernaards et al. 2007) as well as

available-case analysis. In Chapter 7, our two approaches are applied to the California Health Interview Survey (CHIS) 2009 data set. Several possible extensions and further directions of our methods are discussed in Chapter 8.

The dissertation of Ren He is approved.

Karl Lorenz

Donatello Telesca

Catherine A. Sugar

Thomas R. Belin, Committee Chair

University of California, Los Angeles

2012

To my family and friends.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to thank my advisor Dr. Thomas Belin for his professional advice, dedicated guidance, constant support, collaboration and insight. I would also like to thank the rest of my committee, Dr. Catherine Sugar, Dr. Donatello Telesca and Dr. Karl Lorenz, for the time and interest required to review my dissertation. Aside from my advisor and committee members, I would like to thank several other statisticians. I must thank Dr. Juwon Song for helping me obtain the data set and offering helpful advice about my analysis. I want to thank Dr. Hongquan Xu who introduced me to the experimental design idea and helped me use it to improve my simulation plan. For their constant support and love, I would like to thank my family. I want to thank my friends from undergraduate and from high school, who are too many to name, for their lasting friendship and support without which I could not have made it this far. Moreover, without my friends in graduate school, I would certainly not have survived the ordeal.

Vita

1984	Born, P. R. China.
2006	B.S.(Statistics), University of Science and Technology of China, P. R. China.
2008	M.A.(Statistics), University of Missouri, Missouri, USA.
2008–2012	Research Assistant, Department of Biostatistics, University of California, Los Angeles, California, USA.

Publications

J.Qiao, R. He et al., Correlation of MRI Imaging Characteristics with Patterns of Progression in Patients with Recurrent Glioblastoma Treated with Bevacizumab, *Journal of Neuro-oncology*, to appear.

R. He and T. Belin, Multiple Imputation of High-dimensional Mixed Incomplete Data, *In: JSM Proceedings, Biometrics Section, 2011*

P. Chen, R. He, J. Sun and J. Shen, Regression Analysis of Right-Censored Failure Time Data with Missing Censoring Indicators, *Acta Mathematicae Applicatae Sinica* Volume 25, Number 3, July, 2009.

B. Kim and R. He, Survival Analysis of Erlotnib Use in the Veterans Affairs Hospital for Advanced Lung Cancer Patients, *submitted*.

B. Kim and R. He, Predictors of Erlotnib Use in the Veterans Affairs Hospital for Advanced Lung Cancer Patients, *submitted*.

CHAPTER 1

Introduction

In research on populations that have a shared characteristic which makes them interesting to study, such as that they have a particular illness or gene, it is standard practice to collect as much information on them as possible to address wide-ranging questions of scientific interest. By the term "high-dimensional data", we are thinking broadly of settings where traditional models allowing each variable to be correlated with each other variable would have too many parameters to estimate with precision. Beyond involving technical challenges, it can be cumbersome to organize analysis of such data sets one outcome variable at a time. When the number of variables is large relative to the number of cases, even a small number of missing items on each variable can result in a large number of incomplete cases. For example, with 20 variables on 100 cases, if 10 percent of the values on each variable are randomly missing, we would expect only about $100 \times 0.9^{20} \approx 12$ cases with complete records.

When applying multiple imputation to incomplete data sets, it is recommended to include available information to the fullest extent possible because systematic differences between completely and partially observed cases may be reduced by incorporating important covariate information (Rubin 1996). However, when the sample size is modest, even a simple model can be overparameterized when the number of variables is moderately large. For example, for 50 variables, $50 \times 49/2 = 1225$ correlation parameters would need to be estimated in a multivariate normal model with a general covariance matrix. Moreover, sometimes several variables are closely related to one another, which can cause problems with model stability.

Even if it is not overparameterized, a model for a large number of variables may include inestimable or unstable parameters due to the close relationships among variables, making the analysis impossible. Without additional structure to the model, treating data as

multivariate normal with a general covariance structure might give rise to a need to delete variables to avoid inestimable or unstable parameters or to engage in tedious and time consuming model checking. Moreover, data from applied research often include many correlated variables, but it is not always reasonable to analyze data after deleting closely related variables. In this case, we need to use proper priors to make all parameters estimable. The estimates of those parameters will then depend on the prior specification.

Beyond the challenges of modeling continuous data, one can expect to have different types of variables in applied settings, including continuous, binary, ordinal and nominal variables. The idea of developing methods for a joint model to accommodate multivariate data of mixed types presents considerable challenges but would be valuable to applied researchers. The goal of this dissertation is to develop joint modeling strategies that will accommodate realistic data structures involving large numbers of mixed types of variables and modest numbers of cases with general patterns of incomplete data.

The report is organized as follows: Chapter 2 gives a review of the background knowledge related to our two new approaches. We introduce two existing methods of handling high-dimensional continuous incomplete data in Chapter 3 and another two methods for handling mixed incomplete data in Chapter 4. However, the four methods listed in Chapter 3 and 4 are not quite applicable for high-dimensional mixed incomplete data. Our newly developed methods are outlined in Chapter 5. In Chapter 6, simulations under various conditions are carried out to compare the results based on our approaches with the results from the rounding method (Bernaards et al. 2007) as well as available-case analysis. In Chapter 7, our two approaches are applied to the California Health Interview Survey (CHIS) 2009 data set. Several possible extensions and further research directions of our methods are discussed in Chapter 8.

CHAPTER 2

Background review

2.1 Modeling assumptions for incomplete multivariate data

We represent a multivariate data set as a matrix Y with n rows and p columns, where n denotes the number of observations and p denotes the number of variables. The variables can be either continuous or categorical, with the latter including the possibility of binary, ordinal or nominal variables. It is assumed that observations are independently, identically distributed (iid) random draws from a joint multivariate distribution, with the n rows being exchangeable. Missing data can occur anywhere in the data set.

Analysis of incomplete data relies, whether explicitly or implicity, on underlying modeling assumptions. To help characterize important distinctions in the types of models that might be considered, Rubin (1976) classified missing data mechanisms as follows. When missing items do not depend upon both observed values of Y, denoted by Y_{obs}, and the missing values of Y, denoted by Y_{mis}, we say the missingness is missing completely at random (MCAR). Assume that R is a $n \times p$ missing indicator matrix, where $R_{ij} = 1$ means the value in the ith row and jth column is observed, and $R_{ij} = 0$ if that value is missing. If $P(R|Y_{obs}, Y_{mis}, \phi) = P(R|\phi)$, where ϕ refers to the parameters of the missing data mechanism, then the missing mechanism is MCAR. If $P(R|Y_{obs}, Y_{mis}, \phi) = P(R|Y_{obs}, \phi)$, then the mechanism is called missing at random (MAR). Conceptually, MAR allows missing values to depend on observed quantities, but after controlling for observed quantities there is no residual dependence on the underlying missing value.

When the missing data mechanism is either MCAR or MAR and the data Y and the missing data indicators R depend on distinct parameters θ and ϕ respectively, then likelihood-

based inferences about parameters of the data do not depend on the missing data mechanism, and we say the missingness is "ignorable".(Rubin 1976, Little and Rubin 2002). Although the assumption that the missing data mechanism is not always reasonable, it is hard to develop general-purpose missing-data models for nonignorable data, and many methods for nonignorable missingness build on approaches for ignorable missingness. The work presented here will focus on models for settings where the missing data mechanism can be assumed to be ignorable.

2.2 Multiple imputation

Multiple imputation (Rubin 1987) is a technique for imputing $m \geq 2$ plausible values for each missing item. The m plausible values are chosen to reflect the sampling variability of the missing items. Therefore, multiple imputation remains valid in settings where the missing data are MAR and imputations are"proper" in that they accurately represent the distribution of plausible values of unobserved values (Rubin 1987).

Multiple imputation results in m complete data sets. Simulation studies have shown that values of m between $m = 2$ and $m = 10$ give rise to satisfactory coverage when the percentage of missing information is not too large. The standard complete data analysis can be applied to each imputed data set, and the results of the analysis from each imputed data set can be combined to obtain an overall inference (Rubin 1987). Moreover, since the approach involves generating complete data sets, many different analysis can be applied to these data sets.

We denote a complete-data quantity of interest as Q. Building on theory that calls for obtaining draws from the posterior predictive distribution of missing values given observed values, one can multiply impute missing items and get m imputed data sets. The standard complete data analysis for each of m imputed data sets results in the parameter estimates, $\hat{Q}^{(1)}, ..., \hat{Q}^{(m)}$ where $\hat{Q}^{(i)} = \hat{Q}(Y_{obs}, Y_{mis}^{(i)})$, for $i = 1, ..., m$, and their corresponding variance

estimates $U^{(1)}, ..., U^{(m)}$. Then, the multiple imputation point estimate for Q is:

$$\bar{Q} = \frac{1}{m}\sum_{i=1}^{m}\hat{Q}^{(i)} \tag{2.1}$$

and the variance estimate for $\bar{Q}$ is:

$$T = \bar{U} + (1 + \frac{1}{m})B \tag{2.2}$$

where $\bar{U} = \frac{1}{m}\sum_{i=1}^{m}U^{(i)}$, reflecting the average "within-imputation" component of variance, and $B = \frac{1}{m-1}\sum_{i=1}^{m}(\hat{Q}^{(i)} - \bar{Q})^2$, reflecting the "between-imputation" component of variance. The factor $(1 + \frac{1}{m})$ is a correction for performing a finite rather than an infinite number of imputations. The term

$$r = (1 + \frac{1}{m})\frac{B}{\bar{U}} \tag{2.3}$$

is called the relative increase in variance due to nonresponse, and

$$\hat{\lambda} = \frac{(r+2)/(v+3)}{r+1},\ where\ v = (m-1)[1 + \frac{\bar{U}}{(1+1/m)B}]^2 \tag{2.4}$$

is an estimate of the fraction of missing information about Q. When Q is a scalar, inference about Q can be based on the approximation

$$T^{-1/2}(Q - \bar{Q}) \sim t_v \tag{2.5}$$

and a $100(1-\alpha)\%$ confidence interval for Q is $\bar{Q} \pm t_{v,1-\alpha/2}\sqrt{T}$. Inference for multidimensional Q involves matrix generalizations of these formulas (Rubin 1987). Rubin (1987) also shows that the efficiency based on m imputations is approximately $(1 + \frac{\lambda}{m})^{-1}$, where λ is the fraction of missing information for the quantity being estimated. Unless the rate of missing information is very high, there is little advantage to producing and analyzing more than a few imputed data sets.

When it comes to producing multiple imputations, one can consider alternative ap-

proaches which can be grouped under the following three general strategies:

(1) Joint modeling strategy: Develop a joint model for multivariate data and base draws for missing values on implied conditional distributions, often by using a Markov chain Monte Carlo (MCMC) procedure (e.g: Gibbs sampler, data augmentation, Metropolis-Hastings) to draw imputations conditional on drawn parameter values (Robert and Casella 2004). Such an approach is able to draw MCMC sequences which converge in distribution to draws from the desired posterior distribution. This strategy can impute values for multiple missing variables at one time, which saves computation time, especially for high-dimensional data.

(2) Sequential regression strategy (e.g: ICE/MICE: Van Buuren et al 1999; IVEWare: Raghunathan et al 2001): Motivated by analogy with MCMC methods, this idea involves specifying a set of overlapping conditional distributions, even though the collection of conditional distributions might not correspond to a well-defined joint model. As such, this method embraces an approximation at the modeling stage for the sake of flexibility and simplicity. It avoids specifying the covariance structure among variables and can handle mixed type incomplete data. Successful applications of the idea serve as motivating examples since MCMC theory does not apply directly.

(3) Implicit-model strategy (eg: hot-deck imputation (Rubin 1987)): This strategy involves borrowing imputed values from values observed on other cases within the same cell of a contingency table ("hot-deck cells"), which might be based on agreement among key observed characteristics (e.g: borrowing income from individuals in the same geographic area who have the same gender and education level) or based on close agreement among predicted values in a regression (e.g: predict income from multiple covariates, define hot-deck cells based on predicted values from the model, and borrow a value from a case with an observed income whose predicted income placed it in the same hot-deck cell (Schenker and Taylor, 1996)). Like sequential regression imputation, building up an implicit-model approach for an entire multivariate data set tends to proceed one variable at a time. Siddique and Belin (2008) describe a distance-based donor selection approach with an approximate Bayesian Bootstrap (ABB) where donors are selected with probability inversely proportional to their distance from the donee. A SAS macro called MIDAS is available for multiple imputation using

distance-aided selection of donors by Siddique and Harel (2009). This strategy guarantees the imputed values are reasonable since the imputations are from those observed real values.

The above three strategies all have their own criticisms. For the joint modeling strategy, it is not always easy to find a joint model for a mixture of continuous, categorical, semi-continuous and Poisson count variables. For the sequential regression strategy, the order of imputed variables (e.g: Which variable is imputed first?) sometimes affect the output Also there is no good theory to support this method. For hot-deck methods, sometimes it is impossible to impute some missing values. For example, the weight of a patient may be missing because his weight is above the upper-limit of the scale. However, the imputed value using hot-deck method is always below the upper-limit. Thus the imputed value is not realistic. Generally speaking, which multiple imputation strategy to use always depends upon the data set itself and the reason of missingness.

2.3 Data augmentation

Tanner and Wong (1987) proposed the data augmentation approach to estimate parameters when joint distribution of the data is not analytically tractable and is hard to simulate. Assume $z = (x, y)$, and it is difficult to sample (x,y) from the joint distribution p(z). Sampling from the conditional distributions $p(x|y)$ and $h(y|x)$, however may be relatively simple. Let $(x_1^{(t)}, ..., x_m^{(t)}, y_1^{(t)}, ..., y_m^{(t)})$ be a random sample with size m at the tth iteration. Then the (t+1)th step is made up of two steps:

(1). First draw $x_i^{(t+1)}, i = 1, ..., m$ from $p(x|y_i^{(t)})$.

(2). Define $\bar{h}(y|x^{(t+1)}) = \frac{1}{m}\sum_{i=1}^{m} h(y|x_i^{(t+1)})$ to be the equally weighted mixture of the $h(y|x_i^{(t+1)})$, and then draw $y_i^{(t+1)}, i = 1, ..., m$ from $\bar{h}$. Tanner and Wong (1987) proved that the distribution of $(x_1^{(t)}, ..., x_m^{(t)}, y_1^{(t)}, ..., y_m^{(t)})$ converges to p(z) when $t \to \infty$. Notice that when $m = 1$, the data augmentation algorithm reduces to the Gibbs sampler.

To apply the data augmentation algorithm to generate imputations, assume that $y = (y_{mis}, y_{obs})$, where y_{mis} and y_{obs} are the missing and observed part of y respectively. If the

observed-data posterior $p(\theta|y_{obs})$ is hard to simulate, then data augmentation can be used. In the (t+1)th iteration, first draw $y_{mis}^{(t+1)}$ from $p(y_{mis}|y_{obs}, \theta^{(t)})$, and then draw $\theta^{(t+1)}$ from $p(\theta|y_{obs}, y_{mis}^{(t+1)})$. Then one can repeatedly draw values until the sequence is satisfactorily close to a stationary distribution (Gelman and Rubin 1992).

2.4 Factor analysis

Factor analysis was first suggested by Galton (1888) and came into wider use when Spearman (1904) with Pearson applied a single factor idea to intelligence test scores. This simple model was extended later to include multiple common factors. Bartholomew(1987) motivates factor analysis by pointing out that our ability to visualize relationships is often limited to two or three dimensions. Factor analysis has generally been used to explain the relationships among variables with a small number of factors which contain as much information as possible. In multivariate studies with large number of variables, even simple models for the data may contain inestimable or poorly estimated parameters, so that dimension reduction techniques such as factor analysis have the potential to be very useful.

In the linear factor model, observed values of variables can be expressed as a linear function of a smaller number of composites, or factors, which explain the inter relationships among variables, along with a residual, or uniqueness, which reflects characteristics specific to that variable. The model is as follows:

$$Y = \alpha + Z\beta + \epsilon \tag{2.6}$$

where Y is an $n \times p$ observed data matrix, Z is an $n \times k$ unobserved factor-score matrix where $k \leq p$, α is a $1 \times p$ mean vector, β is a $k \times p$ factor-loading matrix, $\epsilon \sim (0, \tau^2)$, where $\tau^2 = diag(\tau_1^2, ..., \tau_p^2)$, and Z and ϵ are independent. Usually, a normal distribution for ϵ is assumed to simplify the estimation process.

It is worth noting that β cannot be determined uniquely. Actually, it is only uniquely determined up to an orthogonal transformation. In the linear factor model, the variance of

Y, $Var(Y)$, can be expressed as $Var(Y) = \beta'\beta + \tau^2$. Assuming that an orthogonal matrix T exists, then, since $TT' = I$, we have:

$$Var(Y) = \beta'\beta + \tau^2 = \beta'TT'\beta + \tau^2 = (T'\beta)'(T'\beta) + \tau^2 \tag{2.7}$$

That is, $\beta^* = T'\beta$ implies the same variance-covariance structure. Therefore, many rotation methods have been developed for the better interpretation of the factor loadings; for example, varimax rotation favors factor loadings near 0 or 1 in absolute value over intermediate values.

Another difficulty comes from the uniqueness terms, τ^2. By definition, τ^2 should be positive, but some methods may yield zero or negative estimates of τ^2 in some cases. In these so-called Heywood cases, the parameter estimates are on the boundary when τ^2 becomes zero. Several methods have been suggested for dealing with this problem. First, one can use a Bayesian approach with a prior distribution for τ^2 (Martin and McDonald 1975). Second, one may stop the iterate process for estimating parameters when any element of τ^2 becomes zero. Third, one can add some small positive number to τ^2, so it becomes positive.

There are controversies about the appropriate sample size needed for factor analysis. In earlier years, it was thought the number of observations should be related to the number of variables, and very large numbers of observations were recommended. However, Aleamoni (1976) argued that the underlying number of factors rather than the number of variables should primarily determine the number of observations needed. Still, large numbers of observations help stabilize estimates of the parameters in factor analysis.

Sometimes it is necessary to find the maximum likelihood estimator (MLE) of the factor loadings. Rubin and Thayer (1982) discuss how the EM algorithm can be used to perform maximum likelihood factor analysis by viewing factor scores as missing data. However, since missing data is a common problem in many data sets, Rubin and Thayer's method needed to be extended to the case where there are missing observations, as well as factors. Jamshidian (1997) introduced EM algorithm for maximum likelihood factor analysis with missing data. Song and Belin (2008) use Jamshidian's method to find the appropriate number of factors for a factor model with incomplete data.

2.5 Multivariate probit model

We now review the multivariate probit model as described in Chib and Greenberg (1998). This modeling technique allows modeling of longitudinal or clustered binary data, ordinal data, which may be useful to multiple impute incomplete binary or ordinal variables.

Suppose we have n subjects measured at each of p occasions or each of p attributes. Let $Y_1, ..., Y_n$ be multivariate binary outcome variables with $Y_i = (Y_{i1}, ..., Y_{ip})^T$ for $i = 1, ..., N$ and $X_{ij} = (X_{ij1}, ..., X_{ijt})^T$ is a $t \times 1$ vector of observed covariates for each subject i and each measurement occasion $j = 1, ..., p$. We assume the following model structure. Each Y_{ij} is distributed Bernoulli with probability of success π_{ij} which is assumed to follow a probit model, i.e. $\pi_{ij} = \Phi(X_{ij}^T\beta)$, where $\Phi(.)$ is the cumulative standard normal distribution function and β is a $t \times 1$ vector of unknown regression parameters.

Let $X_i = (X_{i1}, ..., X_{ip})^T$ be the design matrix for the i-th subject. We introduce n latent variables $Z_1, ..., Z_n$, where the $Z_i = (Z_{i1}, ..., Z_{ip})^T$ are independent $N_p(X_i\beta, R)$, and R is sometimes called the tetrachoric or polychoric correlation of the Y_i (Drasgow, 1986). By defining $Y_{ij} = 1$ if $Z_{ij} > 0$ and $Y_{ij} = 0$ otherwise, it can be easily shown that, marginally, the Y_{ij} are Bernoulli random variables with $\pi_{ij} = P(Y_{ij} = 1) = \Phi(X_{ij}^T\beta)$.

When $Y_1, ..., Y_n$ are multivariate ordinal variables, the element Y_{ij} takes values on the discrete set $0, 1, ..., J_j - 1$, we can still use the above set-up except define $Y_{ij} = l$ if and only if the latent variable Z_{ij} is in the range $(\gamma_{j,l-1}, \gamma_{j,l}]$ where $\gamma_{j,l}$ are the set of cut-points, for $j = 1, ..., p$ and $l = 0, ..., J_j - 1$. Usually, we set $\gamma_{j,0} = -\infty$,$\gamma_{j,J_j-1} = +\infty$ and $\gamma_{j,1} = 0$ for identifiability of the cut-points. Thus we extend the multivariate probit model to ordinal variable case.

Multivariate probit model is difficult to fit because of constraints of covariance matrix (covariance matrix has to be a correlation matrix). It is hard to put a reasonable and convenient prior distribution on correlation matrix. Moreover, in general the full conditional distribution of correlation matrix is not directly available. Zhang, Boscardin and Belin (2004) proposed a parameter-extended Metropolis-Hastings algorithm (PX-MH) for sampling R in Bayesian models with correlated latent variables. The idea in their method is that instead

of a marginal prior for R, they specified a joint prior for R and D (unidentified marginal variances) derived from some inverse Wishart distribution of $\Sigma = DRD$ in model estimation. Then sampling (R;D) jointly was accomplished through a Metropolis Hastings algorithm by first drawing Σ from a pre-specified Wishart distribution with degrees of freedom being the sample size and the scale matrix being the current value of Σ. Using this method, all components of R are drawn at one time. The details of the PX-MH algorithm will be given in Section 5.1.

2.6 Multivariate multinomial probit model

The multinomial probit (MNP) model provides a framework for representing association between levels of a multinomial outcome. Letting $i = 1, ..., n$ index subjects and $g = 1, ..., G$ index levels of a multinomial outcome having G levels. When subject i has outcome g, we define $y_{ig} = 1$, otherwise define $y_{ig} = 0$, thus $y_i = (y_{i1}, ..., y_{iG})$, $i = 1, ..., n$, becomes a multinomial $1 \times G$ vector. Then we define $d = (d_1, ..., d_n)$, let $d_i = g$, if $y_{ig} = 1$. $u_i = (u_{i1}, ..., u_{iG})$ is corresponding latent vector. Due to two identification problems (additive redundance and multiplicative redundance (Zhang et al 2008)), we define the MNP model as follows:

$$z_i = X_i\beta + \epsilon_i \tag{2.8}$$

where $\epsilon_i \sim N(0, \Sigma)$ independently, we restrict the first diagonal element of Σ, σ_{11}, to be equal to 1. $z_{ij} = u_{ij} - u_{iG}$. The model can be described as:

$$d_i = \begin{cases} 0 & if \quad max_{(1 \le l \le G-1)} z_{il} < 0 \\ g & if \quad max_{(1 \le l \le G-1)} z_{il} = z_{ig} > 0 \end{cases}$$

We can extend the MNP model to multivariate nominal measures in straightforward fashion. Assume we have p nominal measures for each subject i,the first measure has G_1 levels, the second has G_2 levels, and so on up to the last, which has G_p levels. Assume

each of the p nominal measures follows a MNP model, and define $d_i = (d_{i1}, ..., d_{ip})$ to be the values of those p measures of the i-th subject. Then we can define the MVMNP model for the p measures as follows:

$$z_i = X_i\beta + \epsilon_i \tag{2.9}$$

where $z_i^T = (z_{i1}, ..., z_{ip})$ with $z_{iq} = (z_{iq1}, ..., z_{iq(G_q-1)})$, $\epsilon_i \sim N(0, \Sigma)$ with $\sigma_{qq} = 1$, where $q = 1, G_1, (G_1+G_2-1), (G_1+G_2+G_3-2)..., (G_1+G_2+...+G_{p-1}-p+2)$. We then specify:

$$d_i = \begin{cases} 0 & if \quad max_{(1\le l\le G_q-1)} z_{iql} < 0 \\ g & if \quad max_{(1\le l\le p_q-1)} z_{iql} = z_{iqg} > 0 \end{cases} \tag{2.10}$$

for $i = 1, ..., n$ and $q = 1, .., p$

Zhang, Boscardin and Belin (2008) describe an MCMC procedure for fitting such a model, where a key step is drawing a correlation matrix for the latent continuous variables that reflects association between nominal categorical variables. For example, correlation among latent continuous variables can induce association between responses to a question about ethnicity and a question about primary language spoken at home (which is apt to be disproportionately Spanish among people who report Hispanic ethnicity).

2.7 Modeling technique for the covariance matrix of high-dimensional longitudinal data

Structured covariance matrices have become extremely popular in recent years for modeling high-dimensional longitudinal data. This approach is appealing as it offers a substantial reduction in the dimensionality of the parameter space leading to more precisely estimated parameters. However, the structured covariance matrices make strong assumptions about the data variances and correlations. When the number of time points is large, the assumptions of a parsimonious covariance model will be untenable. Moreover, if the structured covariance matrix does not fit the data well, this will adversely affect the standard errors of the mean

function and predictions.

The alternative is to use unstructured covariance matrices. Two major problems exist with unstructured covariance models: (i) they involve too many parameters and some of them may be inestimated or poorly estimated, contributing to estimation and prediction variance, and (ii) it can be impossible to estimate a covariance matrix at all unless observations are taken at a small set of times specified by design.

Now the question is: Are we able to find a new Bayesian model that can combine the strengths of structured and unstructured matrices? The framework I plan to pursue in this dissertation is to make use of an idea described by Boscardin and Weiss (2001), who proposed a hierarchical prior distribution for covariance matrix Σ that is centered around a parametric family. This offers a substantial reduction in the dimension of the parameter space, offering the prospect of more precisely estimated parameters. But it also allows flexibility for the data to depart from a tightly structured covariance matrix. Briefly, the idea is to choose a parametric family $\Omega(\theta)$ that reflects anticipated features of Σ. Basic examples would include a first-order autoregressive, or AR(1), process, where correlation between observations at different times goes down geometrically based on the time-lag between them in a manner governed by a parameter left to be estimated or a compound symmetry model where all pairs of the observations have the same correlation. After defining the hierarchical prior distribution, we can carry out an MCMC algorithm, as described in further detail in Section 5.2.

2.8 Summary

After reviewing the related knowledge, we are trying to find a way to multiple impute high-dimensional mixed incomplete data. The central idea in the present work is to introduce continuous latent variables linked to binary, ordinal or nominal variables based upon a multivariate probit model or a multivariate multinomial probit model. Then we can fit a joint model of continuous variables and continuous latent variables for multiple imputation. Ideas of making the covariance matrix of the joint multivariate normal model to be centered around

a restricted parametric family or using a factor model can be applied for reducing the dimension of parameter space of high-dimensional missing data. Under the joint modeling strategy, a Markov Chain Monte Carlo (MCMC) algorithm can be applied to get the parameter estimates and multiple imputation. That is, based on the assumed model structure, the model parameters and missing items can be drawn randomly from conditional distributions with other parameters fixed. The details are described in Section 5.2 and Section 5.3.

CHAPTER 3

Existing multiple imputation approaches for high-dimensional continuous incomplete data

In practice, it is very common to have strongly related variables in the data sets with large numbers of variables. In such cases, the sample variance-covariance matrix becomes singular or almost singular. Thus imputation analysis based on non-informative priors is often impossible. One alternative is to use an informative prior such as a ridge prior (Schafer 1997). The ridge prior is a limiting case of a joint normal inverted-Wishart prior. Another alternative is to use a factor model to reduce the dimension of the paramter space (Song and Belin 2004).

3.1 Ridge prior method

We denote the complete data by $Y = (Y_{obs}, Y_{mis})$, where Y_{obs} and Y_{mis} are the observed and missing portion of the data matrix, let y_{ij} denote an individual element of Y, $i = 1, ..., n$,$j = 1, ..., p$. The i-th row of Y is expressed as a column vector $y_i = (y_{i1}, ..., y_{ip})^T$, we assume:

$$y_1, y_2, ..., y_n|\theta \sim iidN(\mu, \Sigma) \tag{3.1}$$

where $\theta = (\mu, \Sigma)$ is the unknown parameter. Let us apply the following distribution. Suppose that, given Σ, μ is assumed to be conditionally multivariate normal,

$$\mu|\Sigma \sim N(\mu_0, \tau^{-1}\Sigma) \tag{3.2}$$

where $\mu_0 \in R^p$ and $\tau > 0$ are fixed and known. Moreover, suppose that Σ is inverted-Wishart,

$$\Sigma \sim W^{-1}(m, \Lambda) \tag{3.3}$$

for fixed hyperparameters $m \geq p$ and $\Lambda > 0$.

Suppose that we adopt the limiting form of the normal inverted-Wishart prior as $\tau \to 0$ for some m and Λ. The posterior becomes:

$$\mu|\Sigma, Y \sim N(\bar{y}, n^{-1}\Sigma) \tag{3.4}$$

$$\Sigma|Y \sim W^{-1}(m+n, [\Lambda^{-1} + nS]^{-1}) \tag{3.5}$$

which is proper provided that $m + n \geq p$ and $\Lambda^{-1} + nS > 0$. It is called ridge prior by analogy with ridge regression where the strategy to avoid variance inflation in regression coefficient estimates is to add a small positive quantity to the sum-of-squares-and-cross-products matrix, thereby inducing slight bias in the estimated regression coefficients but also inducing a reduction in the variance of those estimates that can improve overall precision. Notice that now the covariance matrix Σ has been smoothed toward a matrix proportional to Λ^{-1}.

The idea of the ridge prior is related to the ridge regression estimator (Hoerl and Kennard 1970). It adds small amount of the variance term to a nearly singular variance-covariance matrix to facilitate taking the inverse of the variance-covariance matrix.

3.2 Song and Belin's factor analysis method

Song and Belin (2004) raised a method to impute incomplete high-dimensional multivariate normal data using common factor analysis model. Let $Y_i, i = 1, ..., n$, denote the i-th observation of Y representing an i.i.d random draw from an underlying sampling distribution, let $y_{ij}, j = 1, ..., p$, denote the j-th variable on the i-th observation of Y. Let $Y = (Y_{obs}, Y_{mis})$. The factor model with k underlying factors can be defined as:

$$Y_i = \alpha + Z_i\beta + \epsilon_i \tag{3.6}$$

where α is a $1 \times p$ mean vector, Z_i is $1 \times k$ factor score vector,β is a $k \times p$ factor loading matrix, $\epsilon_i \sim N(0, \tau^2)$, $\tau^2 = diag(\tau_1^2, ..., \tau_p^2)$, and Z_i and ϵ_i are independent.

One can assume an inverse Gamma distribution $IG(v_j/2, b_j/2)$ as prior distribution for each τ_j^2,$j = 1, ...p$, which is a conjugate prior. We also can assign conjugate priors for α_j and β_j, namely:

$$\alpha_j|\tau_j^2 \sim N(\alpha_{0j}, \frac{1}{n_\alpha}\tau_j^2) \quad for\ j = 1, ..., p \tag{3.7}$$

$$\beta_j|\tau_j^2 \sim N(\beta_{0j}, \frac{1}{n_\beta}\tau_j^2 I_k) \quad for\ j = 1, ..., p \tag{3.8}$$

We can carry out the following Gibbs sampler to impute missing values as well as to estimate parameters:

(1). $Y_{i(mis)}|Y_{i(obs)}, \alpha, \beta, \tau^2 \sim N(a_{mis,obs} + Y_{i(obs)}b_{mis,obs}, \Sigma_{mis,obs}),\ i = 1, ..., n$ Where $a_{mis,obs}$ is a $1 \times (p - p_1)$ intercept vector of the regression of Y_{mis} on Y_{obs} when p_1 variables are observed and $p - p_1$ variables are not observed, $b_{mis,obs}$ is the regression coefficients matrix with dimension $p_1 \times (p - p_1)$, $\Sigma_{mis,obs}$ is the residual matrix.

(2). $Z_i|Y_{i(obs)}, Y_{i(mis)}, \alpha, \beta, \tau^2 \sim N((Y_i - \alpha)(\beta\beta' + \tau^2)^{-1}\beta', I_k - \beta(\beta'\beta + \tau^2)^{-1}\beta'),\ i = 1, ..., n$
Then, transform α to $\alpha^* = \alpha + \bar{Z}\beta$

(3). $\tau_j^2|Y_{obs}, Y_{mis}, Z \sim IG(\frac{n+v_j}{2}, \frac{b'_j}{2})$ $j = 1, ..., p$ where the explicit form of b'_j is not given here due to its complexity, but you can find it from Song and Belin (2004).

(4). $\alpha_j^*|\tau_j^2, Y_{obs}, Y_{mis}, Z \sim N(\frac{n\bar{y_j}+n_\alpha\alpha_{0j}^*}{n+n_\alpha}, \frac{\tau_j^2}{n+n_\alpha})$, $j = 1, ..., p$

(5). $\beta_j|\tau_j^2, Y_{obs}, Y_{mis}, Z \sim N((\sum_{i=1}^{n}(Z_i-\bar{Z})'(Z_i-\bar{Z})+n_\beta I_k)^{-1}(\sum_{i=1}^{n}(Z_i-\bar{Z})'(Y_{ij}-\bar{Y_j})+n_\beta\beta_{0j}), (\sum_{i=1}^{n}(Z_i-\bar{Z})'(Z_i-\bar{Z})+n_\beta I_k)^{-1}\tau_j^2$ Then transform α^* to α by $\alpha = \alpha^* - \bar{Z}\beta$

This algorithm is actually an application of data augmentation, and we can apply multiple imputation after imputing Y_{mis}. The dimension of parameter space is reduced by introducing the common factor model. The simulation results of Song and Belin (2004) show that this algorithm tend s to have less bias compared to Schafer's ridge prior method when we can find the correct number of factors for the factor model. However, this algorithm only works for high-dimensional continuous incomplete data. Modification is needed to extend this algorithm to accommodate mixed incomplete data.

CHAPTER 4

Existing multiple imputation approaches for incomplete data with mixed data types

4.1 General location model

Little and Schluchter (1985) used a general location model to analyze mixed missing data. Assume there are p_1 continuous variables (X) and p_2 categorical variables (Y). The j-th categorical variable has I_j levels. Thus we can define the categorical variables as a contingency table with $C = \sum_{j=1}^{p_2} I_j$ cells. For subject i, let x_i be the $1 \times p_1$ vector of continuous variables and y_i be the $1 \times p_2$ vector of categorical variables. Also from y_i we define $1 \times C$ vector w_i, which equals D_c if subject i falls into cell c. D_c is a $1 \times C$ vector with 1 as the c-th entry and 0 elsewhere. The general location model (Olkin and Tate 1961) is as follows:

$$P(w_i = D_c) = \pi_c, \ c = 1, ..., C, \ \sum \pi_c = 1 \tag{4.1}$$

$$x_i | w_i = D_c \sim N_K(\mu_c, \Omega) \tag{4.2}$$

the unknown parameters are $(\pi_1, ..., \pi_C, \mu_c, \Omega)$. Once we have the model, Little used an EM algorithm to calculate the estimators for those unknown parameters. See Little and Schluchter (1985) for details. In practice, the covariance matrix of the continuous variables is usually assumed to be constant across cells. However, this assumption is not always realistic in real-life applications. Also, the general location model does not perform well when there are sparse cells in the contingency table of the categorical data. If some cells are sparse or even empty, the corresponding parameters are poorly estimated or even inestimated.

4.2 Sequential regression method

To analyze mixed missing data, Raghunathan et al (2001) raised an approach named sequential regression multiple imputation (SRMI). Assume we have k variables $Y = (Y_1, ..., Y_k)$ with missing values, ordered by the amount of missing values, from least to most. Let X be the matrix containing all the variables without missing values. Both X and Y can include continuous,binary, ordinal or mixed variables. The imputation steps are as follows:

(1). Regress the most observed variable Y_1 on X, imputing the missing values under the appropriate regression model, then regress Y_2 on X and Y_1,...., regress Y_k on $Y_1, ...Y_{k-1}$ and X, then we have a complete data set (X,Y).

(2). Then regress Y_1 on X and $Y_2, ..., Y_k$, update the imputed missing values by this regression model, regress Y_2 on X and $Y_1, Y_3, ..., Y_k$ and so on. Repeat the steps until stable imputed values occur.

This method avoids joint modeling strategy and it does not require specifying the covariance structure. Also it's easy to carry out since this method is already embedded in several statistical softwares such as Stata and R. However, all above approaches are either complicated or not applicable when we encounter high-dimensional situation.

4.3 Boscardin, Zhang and Belin's method

Boscardin, Zhang and Belin (2006) develop a modeling technique for data that are a mixture of ordinal and continuous components. They also plug the missing data imputation step in their MCMC algorithm so their model can also be used to impute missing values of mixed continuous and ordinal variables. Their idea is:

1. Treat the ordinal data in the multivariate probit model framework, assuming there is an underlying normal latent variable for each ordinal variable.

2. Use a multivariate normal distribution for latent continuous variables (corresponding to the ordinal measures) and the continuous variables.

3. Assign prior distributions to sample parameters, randomly draw sample parameters from their full conditional distributions (conditional on other parameters, observed values and missing values). Since there are constraints for the variance-covariance matrix of multivariate probit model (the part of the covariance matrix corresponding to ordinal variables has to be a correlation matrix), the Parameter-extended Metropolis-Hastings (PX-MH) algorithm is used to generate the variance-covariance matrix.

4. Randomly draw the missing values from its full conditional distribution (conditional on all the model parameters and observed values)

5. Repeat step 3 and step 4 until convergence is obtained.

This approach does not work well when the dimension of the data set is large. Some of the parameters may be inestimable or poorly estimated. Modification of this approach is given in Section 5.2.

CHAPTER 5

Newly Proposed Methods of Imputation of High-dimensional Mixed Incomplete Data

In this section, we propose two new methods for handling a high-dimensional data set with both continuous and categorical variables. We build the first method on the Parameter-extended Metropolis-Hastings algorithm outlined by Zhang et al (2006) and adapt it to settings with multiple types of variables. The second method is based upon a factor analysis model adapted with a mixture of continuous and binary variables.

5.1 Parameter-extended Metropolis-Hastings (PX-MH) algorithm

As discussed earlier, the multivariate probit model is a useful tool to analyze binary or ordinal data. The covariance matrix of the multivariate probit model is in fact a correlation matrix. However, there are not many well-known families of density functions for correlation matrices, which motivated Zhang et al (2006) developed a flexible solution for placing a prior density on the correlation matrix using a separation strategy. If R is the corresponding correlation matrix, we define $W = D^{\frac{1}{2}} R D^{\frac{1}{2}}$, where W is a positive definite covariance matrix, and $D^{\frac{1}{2}}$ means a diagonal matrix of the standard deviations. That is:

$$W = \begin{pmatrix} w_{11} & ...w_{1p} \\ ... & ..... \\ w_{p1} & ...w_{pp} \end{pmatrix} \tag{5.1}$$

$$R = \begin{pmatrix} 1 & r_{12} & ... & r_{1p} \\ r_{21} & 1 & ... & r_{2p} \\ ... & ... & ... & ... \\ r_{p1} & r_{p2} & ... & 1 \end{pmatrix} \tag{5.2}$$

$$D = \begin{pmatrix} w_{11} & 0 & ... & 0 \\ 0 & w_{22} & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & w_{pp} \end{pmatrix} \tag{5.3}$$

where $w_{ij} = \sqrt{w_{ii}}\sqrt{w_{jj}}r_{ij}$, for $i \neq j$. Thus we expand the parameter space with a new parameter D. The Jacobian transformation of $(W \rightarrow R, D)$ is equal to:

$$(\prod_{i=1}^{p} w_{ii})^{\frac{p-1}{2}} \tag{5.4}$$

We can assume a prior distribution p(W) for W, then we have $p(R, D) = Jacobian(W \rightarrow R, D) \times p(W)$, assume $p(R, D|Y)$ is the posterior distribution of (R, D). Sampling (R, D) is accomplished through a Metropolis-Hastings algorithm by sampling W. Zhang et al(2006) named it as parameter extended Metropolis-Hastings algorithm (PX-MH). There are two generic priors for (R, D). If we assume W follows $Wishart_p(m, \Sigma)$, then we say (R, D) have a PXW prior, if W follows $Wishart_p^{-1}(m, \Sigma)$, we say (R, D) have a PXIW prior. The steps of PX-MH algorithm are listed as follows:

(1). set initial value of (R^0, D^0) through setting $W^0 = D^{0^{1/2}} R^0 D^{0^{1/2}}$ to an initial covariance matrix. For $i = 1, ..., M$:

(2). generate $W^* = D^{*^{1/2}} R^* D^{*^{1/2}} \sim Wishart_p(m, W^i)$., thus we have (R^*, D^*).

(3).

$$(R^{i+1}, D^{i+1}) = \begin{cases} (R^*, D^*) & with\ probability\ \alpha \\ (R^i, D^i) & otherwise \end{cases} \tag{5.5}$$

where $\alpha = min(\frac{p(R^*,D^*|Y)\times q(W^i|W^*)}{p(R^i,D^i|Y)\times q(W^*|W^i)}, 1)$. $q(.|W^i)$, the proposal density, is the product of Jacobian term given in (5.24) and the Wishart density with m degrees of freedom and scale matrix equal to W^i.

5.2 Modeling high-dimensional mixed incomplete data using a parametric family for the covariance matrix

In this section, we borrow the idea of PX-MH algorithm and modeling covariance matrix techniques in Section 2.7 to come up with a new algorithm that can handle high-dimensional mixed incomplete data problems. For the high-dimensional problems we aim to address, we can assume $\Sigma|\nu, \theta \sim$Inv-Wishart$(\nu, (\nu\Omega(\theta))^{-1})$, where ν is the degree of freedom parameter. The Inverse Wishart distribution is a conjugate family for modeling covariances in a Bayesian framework. In the setting here, there is only partial conjugacy given the complexities outlined by Zhang et al (2006) for converting a drawn covariance matrix into a drawn correlation matrix. Actually, ν can be treated as a tuning parameter for the amount of smoothing performed. When ν value is large, Σ is constrained to the parametric family $\Omega(\theta)$, the deviation will be very small. When $\nu \to \infty$, Σ exactly belongs to $\Omega(\theta)$. When the ν value is small, then Σ can deviate a lot from the parametric family $\Omega(\theta)$, when ν is close to 0, Σ can have an arbitrary structure. However, our data setting here is cross-sectional, not longitudinal. In this case, commonly used structured matrices, such as autoregressive structures or compound symmetry structure, may not be appropriate to be used as a centered parametric family. But compound symmetry (CS) structure is still a reasonable option.

For the purpose of fixing ideas, we assume here a scenario with only continuous and binary data, although the intention is to expand the idea to incorporate ordinal and nominal

categorical data using multivariate probit and multivariate multinomial probit modeling techniques. Let $T_i^T = (v_i^T, c_i^T), i = 1, ..., n$ consists of a continuous proportion $v_i^T = (v_{i1}, ..., v_{ip_1})$ with length p_1 and a binary portion $c_i^T = (c_{i1}, ..., c_{ip_2})$ with length p_2, $p_1 + p_2 = p$. Assume $v_i \sim N_{p_1}(X_i^{(1)}\beta, \Sigma_{vv})$, where $X_i^{(1)}$ is a $p_1 \times t$ predictor matrix and β is a $t \times 1$ regression parameter, Σ_{vv} is a $p_1 \times p_1$ covariance matrix.

We treat the binary variables in the multivariate probit model framework in section 1.4. Let z_i is the corresponding latent vector for c_i, z_i is a $p_2 \times 1$ vector. z_i follows a multivariate normal distribution $N_{p_2}(X_i^{(2)}\beta, R_{zz})$, where $X_i^{(2)}$ is a $p_2 \times t$ predictor matrix for the linear model, R_{zz} is a $p_2 \times p_2$ correlation matrix. Note that the use of a correlation matrix R_{zz} for the latent normal variables addresses the identifiability problem for the multivariate probit model but presents challenges for an MCMC estimation procedure given the difficulty of drawing correlation matrices. In the present development, we will assume $X_i^{(1)}$ and $X_i^{(2)}$ are completely observed, relaxing this assumption would involve an extension. Let Λ denote the covariance matrix of the $y_i^T = (v_i^T, z_i^T)$:

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

where R_{vv} is the corresponding correlation matrix for Σ_{zz}, and D_{vv} is the diagonal matrix of variances of Σ_{vv}.

Given the latent variable z_i, $y_i^T = (v_i^T, z_i^T) \sim N_p(X_i\beta, \Lambda)$, where $X_i = (X_i^{(1)}, X_i^{(2)})$. From a Bayesian perspective, we need to define the joint prior distribution $p(\beta, \Lambda)$ for (β, Λ). For simplicity, we assume $p(\beta, \Lambda) = p(\beta) \times p(\Lambda)$. Let $\beta \sim N_k(b_0, B)$, where b_0 and B are known mean and covariance matrix. Let $Z = (z_1, ..., z_n)^T$,$V = (v_1, ..., v_n)^T$,$C = (c_1, ..., c_n)^T$, also let $v_i^T = (v_{i,obs}^T, v_{i,mis}^T)$, $v_{i,obs}$ is the observed part of v_i, $v_{i,mis}$ is the missing part of v_i. Similarly, $c_i^T = (c_{i,obs}^T, c_{i,mis}^T)$. Corresponding to c_i, I write $z_i^T = (z_{i,obs}^T, z_{i,mis}^T)$ although z_i are all not observed. $z_{i,obs}$ is corresponding to $c_{i,obs}$, while $z_{i,mis}$ is corresponding to $c_{i,mis}$. The following MCMC algorithm can be used to impute missing values and estimate unknown parameters:

- $\beta|\Lambda, Z, V, C \sim N(\hat{\beta}, V_\beta)$ where

$$V_\beta = ((\sum_{i=1}^{n}(X_i^T\Lambda^{-1}X_i)) + B^{-1})^{-1} \tag{5.6}$$

$$\hat{\beta} = V_\beta(\sum_{i=1}^{n}(X_i^T\Lambda^{-1}y_i) + B^{-1}b_0) \tag{5.7}$$

- For the missing continuous variables $v_{i,mis}$, we have:

$$v_{ij,mis}|\beta, \Lambda, z_i, v_{i,obs}, v_{ik,mis}, k \neq j \sim N(\hat{v}_{ij}, s_{ij}) \tag{5.8}$$

Where $\hat{v_{ij}}$ and s_{ij} are usual conditional mean and variance. To give explicit formulas for these, recall that $y_i^T = (v_i^T, z_i^T)$. Without loss of generality, we assume in v_i, the first p_{1i} variables are observed and the rest are missing. Thus $v_{ij,mis}$ is the $(j' = p_{1i} + j)$th element of y_i. Let $X_{ij'}$ be the $t \times 1$ covariate vector for the ith subject's j'th continuous variable. We have:

$$\hat{v}_{ij} = X_{ij'}^T\beta + \Lambda_{j',-j'}\Lambda_{-j',-j'}^{-1}(y_{i,-j'} - X_{i,-j'}\beta) \tag{5.9}$$

$$s_{ij} = \Lambda_{j',j'} - \Lambda_{j',-j'}\Lambda_{-j',-j'}^{-1}\Lambda_{-j',j'} \tag{5.10}$$

where $\Lambda_{j',-j'}$ means the vector of the j'th row of Λ without its j'th column element, $\Lambda_{-j',j'}$ is the transpose of $\Lambda_{j',-j'}$, $\Lambda_{-j',-j'}$ is the Λ matrix without its j'th row and j'th column, $X_{i,-j'}$ is the matrix X_i without the j'th row, and $y_{i,-j'}$ is the vector y_i without its j'th element.

- For the latent variable Z_i,

$$z_{ij,obs}|\beta, \Lambda, v_{i,mis}, v_{i,obs}, z_{i,mis}, z_{ik,obs}, j \neq k \propto I_{ij} \times N(\hat{z}_{ij}, t_{ij}) \tag{5.11}$$

where $I_{ij} = 1_{(c_{ij}=1)}1_{(z_{ij}>0)} + 1_{(c_{ij}=0)}1_{(z_{ij}\leq 0)}$, $\hat{z}_{ij}$,t_{ij} are conditional mean and variance of

$z_{ij,obs}$ given v_i,$z_{i,mis}$,$z_{ik,obs}$,$k \neq j$. So this is actually a truncated normal distribution.

$$z_{ij,mis}|\beta, \Lambda, v_{i,mis}, v_{i,obs}, z_{i,obs}, z_{ik,mis}, j \neq k \sim N(\tilde{z}_{ij}, \tilde{t}_{ij}) \tag{5.12}$$

It is just the normal univariate conditional distribution without truncation. I assume the length of the vector $z_{i,obs}$ is p_{2i}, then $z_{ij,obs}$ is the $(j'' = p_1 + j)$th element of y_i, while $z_{ij,mis}$ is the $(j''' = p_1 + p_{2i} + j)$th element of y_i. Thus $\hat{z}_{ij}, t_{ij}, \tilde{z}_{ij}, \tilde{t}_{ij}$ have the similar forms of $\hat{v}_{ij}, s_{ij}$ in step (2), we only need to change the corresponding subscripts.

-

$$p(\Lambda|\beta, Z, Y, C) \propto p(\Lambda) \times \prod_{i=1}^{n} N(y_i|X_i\beta, \Lambda) \tag{5.13}$$

This is the most difficult part of the algorithm, since Λ is a covariance matrix with the lower diagonal part to be a correlation matrix. Thus we cannot use either approaches for sampling an unrestricted covariance matrix nor approaches for sampling a true correlation matrix. Here we use the generalized parameter-extended Metropolis-Hastings algorithm (Boscardin et al 2008) to sample Λ.

First we expand Λ to Σ which include a diagonal scale matrix D_{zz} for z_i.

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

Then we can write $\Sigma = D^{1/2} R D^{1/2}$, where

$$D = \begin{pmatrix} D_{vv} & 0 \\ 0 & D_{zz} \end{pmatrix} \; and \; R = \begin{pmatrix} R_{vv} & R_{vz} \\ R_{vz}^T & R_{zz} \end{pmatrix}$$

In order to reduce the high dimension of the parameter space, we then place a Wishart prior distribution on Σ with ν degree of freedom and scale matrix $\Omega(\theta)/\nu$, so that Σ has prior mean $\Omega(\theta)$. For this cross-sectional data setting, usually we use compound symmetry

(CS) matrix as a possible option for $\Omega(\theta)$, the diagonal elements of the CS matrix can be considered as the average variance across the p variables, while the off-diagonal elements can be considered as the average covariance across the p variables. To gain more flexibility, the degree of freedom parameter ν should be treated as unknown, although we have concerns about the estimation procedure if we were to assign ν a noninformative prior. So we plan to start by using a gamma prior left truncated at p with known hyperparameters α_ν and β_ν. The reason of the truncation at p is to make sure the Wishart prior distribution is well defined. Now we know the prior distribution for (R, D) is $p(R, D) = Jacobian(\Sigma \rightarrow R, D) \times Wishart_\nu(\Sigma|\Omega(\theta)/\nu)$. The conjugacy that is gained in the covariance matrix situation from using the inverse-Wishart distribution is unfortunately lost because of this Jacobian term. We will use the Wishart distribution in what follows for simplicity. Define $\tilde{Y} = Y - X\beta$, then we have:

$$p(R, D, \theta, \nu|Y) \propto (2^{vp/2}\Gamma_p(\nu/2))^{-1}|R|^{-n/2}exp(-\frac{1}{2}tr(R^{-1}\tilde{Y}^T\tilde{Y})) \tag{5.14}$$

$$(\prod_{j=1}^{p} D_{jj})^{(p-1)/2}|\Omega(\theta)/\nu|^{-\nu/2}|D^{1/2}RD^{1/2}|^{(\nu-p-1)/2} \tag{5.15}$$

$$exp(-\frac{1}{2}tr(D^{1/2}RD^{1/2}(\Omega(\theta)/\nu)^{-1}))p(\theta)p(\nu) \tag{5.16}$$

where $\Gamma_p(\nu/2) = \prod_{j=1}^{p} \Gamma((\nu + 1 - j)/2)$. However, when the value of p goes large,$\Gamma_p(\nu/2)$ may be a huge number which cannot be stored in modern statistical softwares. For example, $\Gamma_p(\nu/2) = \infty$ in Matlab software when $p = 100$ although the real value is not infinity, which may cause computational problem. It's because the real value of $\Gamma_p(\nu/2)$ is a way big number which can not be saved by the software. So for large p, I use a Sterling formula to approximate $\Gamma_p(\nu/2)$.

We need to add the following steps in the MCMC algorithm by using the PX-MH algorithm described in Section 5.1:

- generate $\Sigma^* = (D^*)^{1/2}R^*(D^*)^{1/2} \sim Wishart_{\nu_0}(\Sigma^{(m)}/\nu_0)$, we accept $(R^{(m+1)}, D^{(m+1)}) =$

(R^*, D^*) with probability α, where

$$\alpha = min(\frac{p(R^*, D^*, \nu^{(m)}, \theta^{(m)}|\tilde{Y}) \times q(\Sigma^m|\Sigma^*)}{p(R^m, D^m, \nu^{(m)}, \theta^{(m)}|\tilde{Y}) \times q(\Sigma^*|\Sigma^m)}, 1). \tag{5.17}$$

$q(\Sigma^*|\Sigma^m) = Jacobian(\Sigma^* \to R^*, D^*)Wishart_{\nu_0}(\Sigma^*|\Sigma^m/\nu_0)$ is the jumping kernel. Once we get $\Sigma^{(m+1)}$, we can have $\Lambda^{(m+1)}$.

- sample θ from $p(\theta|\Sigma^{m+1}, \nu^{m+1}, \beta, Z, C, V)$, this probability may not have analytical form. We can use a Metropolis-Hastings algorithm here.

- sample ν from $p(\nu|\Sigma^{(m+1)}, \theta^{m+1}, \beta, Z, C, V)$ also using a Metropolis-Hastings algorithm.

So now we have an entire MCMC algorithm, repeating the steps until convergence is obtained. We anticipate that it will be straightforward to extend this algorithm to ordinal cases by adding the cutoff points and that we can apply this approach to nominal variables as well using multivariate multinomial probit model.

Now we extend our algorithm to the ordinal variable case. Besides the continuous and binary variables, assume we also have p_2 ordinal variables $s_i^T = (s_{i1}, ..., s_{ip_2})$,$i = 1, ..., n$. The element s_{ij} takes values on the discrete set $0, 1, ..., J_i - 1$, the corresponding latent variables are $\omega_i^T = (\omega_{i1}, ..., \omega_{ip_2})$. According to the multivariate probit model, $s_{ij} = l$ if and only if ω_{ij} is in the range $(\gamma_{j,l-1}, \gamma_{j,l}]$ where $\gamma_{j,l}$ are the set of cut-points, for $j = 1, ..., p_2$ and $l = 0, ..., J_j - 1$. As mentioned in previous chapter, we set $\gamma_{j,0} = -\infty, \gamma_{j,J_j-1} = \infty, \gamma_{j,1} = 0$ for notation simplicity and identifiability. Let's assume $\omega_i \sim X_i^{(3)}\beta$. Given the latent variable z_i, ω_i, now $y_i^T = (v_i^T, z_i^T, \omega_i^T) \sim N_p(X_i\beta, \Lambda)$, where $X_{newi} = (X_i^{(1)}, X_i^{(2)}, X_i^{(3)})$. Then we can copy the above MCMC steps while using X_{newi} instead of X_i, also we need plug in three more steps:

(a). generate $w_{i,mis} \sim P(w_{i,mis}|v_i, z_i, w_{i,obs}, \beta, \Gamma)$, which is a multivariate normal distribution with regular conditional mean and variance

(b). generate $w_{ij,obs} \sim P(w_{i,obs}|v_i, z_i, w_{i,obs}, \beta, \Gamma)$, which is a multivariate normal distribu-

tion truncated by interval $(\gamma_{j,k-1}, \gamma_{j,k}]$ if $s_{ij} = k$

(c). generate $\gamma_{j,l}$ from $p(\gamma_{j,l}|\beta, X_i, \omega_i, z_i, \Gamma, \gamma_{j,k}, k \neq l)$
$\propto U(\gamma_{j,l}|max\{max\{\omega_{ij} : s_{ij} = l\}, \gamma_{j,l-1}\}, min\{min\{\omega_{ij} : s_{ij} = l+1\}, \gamma_{j,l+1}\})$

5.3 Modeling high-dimensional mixed continuous and binary data using a factor analysis strategy for the covariance matrix

In Section 3.2, I introduced Song and Belin's factor analysis approach for high-dimensional continuous data. It is natural to extend this method to high-dimensional mixed type data situation. However, standard factor analysis models are not designed to accommodate mixed data. So here normal latent variables are used to accommodate binary or categorical data. Some constraints are also added to the model to make sure the model is identifiable. For simplicity, here I only consider the mixture of continuous and binary variables.

Let's use the settings in the last section. But I extend the assumption that the data set also includes binary variables. $T_i^T = (v_i^T, c_i^T), i = 1, ..., n$ include both continuous and binary variables. z_i is the latent vector corresponding to the binary part c_i. Let $y_i^T = (v_i^T, z_i^T)$ with length p. Assume we have p_1 continuous variables and p_2 binary variables. Then $p = p_1 + p_2$. We divide v_i into two parts. $v_i = (v_{i,obs}, v_{i,mis})$, where $v_{i,obs}$ denotes the observed part of v_i, $v_{i,mis}$ denotes the missing part of v_i. Similarly, we can define $z_i = (z_{i,obs}, z_{i,mis})$ and $c_i = (c_{i,obs}, c_{i,mis})$. The factor model is:

$$y_i = \alpha + \phi_i \Lambda + \epsilon_i \tag{5.18}$$

where α is a $1 \times p$ intercept vector, Λ is a $k \times p$ factor loading matrix. ϕ_i is a $1 \times k$ factor score and $\phi_i \sim N(0, I)$, $\epsilon_i \sim_{iid} N(0, \tau)$, $\tau = diag(\tau_1^2, ..., \tau_p^2)$. The diagonal elements of τ that correspond to binary and ordinal response variables are constrained to be equal to 1 for identification purpose. That is, $\tau_j^2 = 1, j = p_1 + 1, ..., p$. Now the question is: How can we assign proper prior information for the unknown parameters?

For the prior distribution of the j-th diagonal unconstrained element of τ, denoted τ_j^2, we assume an inverse gamma distribution $IG(\nu_j/2, b_j/2)$,
$j = 1, ..., p_1$. We start with this prior distribution due to its convenience as a conjugate form.

Meanwhile, conjugate prior distributions can be assigned for α_j and β_j, namely:

$$\alpha_j|\tau_j^2 \sim N(\alpha_{0j}, \frac{1}{n_\alpha}\tau_j^2) \quad for\ j = 1, 2, ..., p_1 \tag{5.19}$$

$$\alpha_j|\tau_j^2 \sim N(\alpha_{0j}, \frac{1}{n_\alpha}) \quad for\ j = p_1 + 1, ..., p \tag{5.20}$$

$$\Lambda_j|\tau_j^2 \sim N(\Lambda_{0j}, \frac{1}{n_\Lambda}\tau_j^2 I_k) \quad for\ j = 1, 2, ..., p_1 \tag{5.21}$$

$$\Lambda_j|\tau_j^2 \sim N(\Lambda_{0j}, \frac{1}{n_\Lambda} I_k) \quad for\ j = p_1 + 1, ..., p \tag{5.22}$$

where α_{0j} and Λ_{0j} are prior means, n_α and n_Λ can be viewed as additional prior degrees of freedom for inference about α and Γ respectively, and I_k is a $k \times k$ identity matrix. With these specifications, we can derive the following MCMC algorithm:

- Simulate the missing values of continuous variables from

$$v_{ij,mis}|v_{i,obs}, z_i, \alpha, \Lambda, \phi_i, \tau^2 \sim N(a_j, \tau_j^2), j \in F_v(i) \tag{5.23}$$

  where $F_v(i)$ denotes the missingness position index set for $v_{i,mis}$. For example, if v_{22}, v_{25} are missing, then $F_v(2) = 2, 5$. Note that each $v_{ij,mis}$ is independent to other $v'_{ij,mis}s$ and z_i when conditional on the factor score ϕ_i.

- Simulate the latent variables corresponding to the missing part of binary variables from

$$z_{ij,mis}|v_i, z_{i,obs}, \alpha, \Lambda, \phi_i, \tau^2 \sim N(a_j, \tau_j^2), j \in F_{z_1}(i) \tag{5.24}$$

  where $F_{z_1}(i)$ denotes the index set for $z_{i,mis}$.

- Simulate the latent variables corresponding to the observed part of binary variables

from

$$z_{ij,obs}|v_i, z_{i,mis}, z_{iL,obs}, L \neq j, \alpha, \Lambda, \phi_i, \tau^2 \sim [I_{(z_{ij,obs}>0)}I_{(c_{ij}=1)}+$$
$$I_{(z_{ij,obs}<=0)}I_{(c_{ij}=0)}] \times N(a_j, \tau_j^2), j \in F_{z_2}(i) \tag{5.25}$$

which are truncated univariate normal distributions. $F_{z_2}(i)$ denotes the index set for $z_{i,obs}$. $z'_{ij,obs}s$ are all independent to each other when conditional on the factor score ϕ_i.

- Simulate factor scores from

$$\phi_i|y_i, z_i, \alpha, \Lambda, \tau^2 \sim N((y_i - \alpha)(\Lambda'\Lambda + \tau^2)^{-1}\Lambda', I_k - (\Lambda'\Lambda + \tau^2)^{-1}\Lambda') \tag{5.26}$$
$$\tag{5.27}$$

Then, transform α to $\alpha^* = \alpha + \bar{\phi}\Lambda$, where $\bar{\phi} = \frac{1}{n}\sum_{i=1}^{n}\phi_i$. The reason for making this transformation is to reduce the high autocorrelation between α and other parameters.

- Simulate uniqueness terms from

$$\tau_j^2|y_i, z_i, \phi, \alpha, \Lambda \sim IG(\frac{n+\nu_j}{2}, \frac{b'_j}{2}),\ j = 1, ..., p_1 \tag{5.28}$$

- Simulate mean estimates from

$$\alpha_j^*|\tau_j^2, Y_{obs}, Y_{mis}, Z \sim N(\frac{n\bar{y_j} + n_\alpha\alpha_{0j}^*}{n+n_\alpha}, \frac{\tau_j^2}{n+n_\alpha}),\ j = 1, ..., p \tag{5.29}$$

where the the explicit formula for term $'_{-j}$ can be found from Song and Belin (2004), I won't give the details here due to its complicated form.

- Simulate the factor loading from

$$\Lambda_j|\tau_j^2, Y_i, Z_i, \phi_i, \alpha \sim N((\sum_{i=1}^{n}(\phi_i - \bar{\phi})'(\phi_i - \bar{\phi}) + n_\beta I_k)^{-1}(\sum_{i=1}^{n}(\phi_i - \bar{\phi})'(Y_{ij} - \bar{Y_j})$$
$$+n_\beta\beta_{0j}),(\sum_{i=1}^{n}(\phi_i - \bar{\phi})'(\phi_i - \bar{\phi}) + n_\beta I_k)^{-1}\tau_j^2) \tag{5.30}$$

Then transform α^* to α by $\alpha = \alpha^* - \bar{Z}\beta$, and transform $z_{1i,mis}, z_{2i,mis}$ to $c_{1i,mis}, c_{2i,mis}$ using multivariate logit model.

This algorithm is actually an application of Gibbs sampler. The transformation we made in step (4) is designed to avoid the slow convergence due to high correlation between α and Λ (Song and Belin 2004). The convergence of our MCMC algorithm can be monitored by the time-series plots of all parameters or Gelman-Rubin statistics.

When there are more than one mode of the likelihood, the Gibbs sampler may not mix values across separate regions of appreciable posterior density. In this case, we can draw values from multiple chains based on multiple starting values from a over-dispersed distribution.

It is possible that sometimes the generated uniqueness term in the iteration of Gibbs sampler is close to zero, resulting in a so-called Heywood case. We can use a proper prior distribution for τ_j^2 to avoid the Heywood case.

The factor model proposed here has much fewer parameters than a multivariate normal model in high-dimensional setting. In the factor model, we assume the number of factors, k, is already known in advance. The effect of various choices of k is considered in our simulation study. Song and Belin (2008) discussed the possible approaches for choosing an appropriate k value for high-dimensional continuous case. We may extend their idea to mixed data situation. We can incorporate ordinal and nominal variables each has different levels in our model using the multivariate logit and multinomial logit model in Chapter 2.5, 2.6.

CHAPTER 6

Simulation Studies

In this chapter, we carry out two sets of simulation studies to evaluate the validity of the two proposed approaches. The goal will be to recover parameter values used to generate the data based on inference from the incomplete data sets where a missingness pattern has been introduced. In Section 6.1, a simulation study is based on two data sets generated from a multivariate normal structure with the covariance matrix either to be an unstructured or a compound symmetry matrix. After choosing missing items from different missing data mechanisms, we apply our parametric-family method centered around a compound symmetry matrix to impute missing items. In Section 6.2 , a simulation study is based on data generated from a simple factor structure. Also after choosing missing items from a MCAR or MAR missing data mechanism with different missing rates, we apply factor models either with a correct number of factors or with an incorrect number of factors to impute missing items.

After establishing the validity of the approaches, we plan to compare the proposed methods developed for a mixed of variable types with potential competitor methods. For example, the multivariate normal model approach of Schafer (1997) could be applied to binary data, with imputed values rounded to the nearer of 0 or 1, in line with the approach considered by Bernaards, Belin and Schafer (2007). Bernaards et al (2007) found that rounding normal imputations to produce binary imputations tended to work better with underlying proportions close to 0.5 than with underlying proportions close to 0 or 1 to produce close-to normal coverage. Accordingly, we plan to vary underlying proportions for binary variables in the simulations, with some assumed to be 0.7 and some assumed to be 0.1, by making the mean of the latent variables not to be 0.

Five multiply imputed data sets are generated from each of the parametric family model

and factor model. To reflect the broader concerns with a large number of variables, we propose to start with $p = 50$ variables. This is a realistic number to measure in applied investigations. Specifically, we plan to generate 25 continuous variables and 25 binary variables. Let $y_1, ..., y_{25}$ be continuous variables and $y_{26}, ..., y_{50}$ be binary variables. Comparisons are based on the inferences about the means of y_{25} and y_{50}, where the underlying values are known in both cases. In addition to results from both the factor method and the parametric family model and the rounding method, we also present the results from available-case analysis. We can compare the standard deviations, the lengths of confidence intervals and the coverage probabilities.

6.1 Simulation Studies for Parametric Family Approach

In section 5.3, I proposed a new method for handling high-dimensional mixed missing data. Here I use all the definitions defined there. Because the prior distribution of augmented covariance matrix Σ is centered around a parametric family $\Omega(\theta)$, it is natural to consider how the choice of the parametric family will affect the inference.

Suppose Σ is truly an compound symmetry (CS) covariance matrix. As I discussed in section 5.3, using an CS hierarchical center for the prior distribution of Σ with large v value will be well-supported by the data. In contrast, if Σ is truly an unstructured covariance matrix, then a compound symmetry hierarchical center with small v value may have good behavior.

I choose the sample size $n = 100$ or $n = 300$ which denote moderate and relatively larger sample size situations respectively. Specifically, we plan to generate 25 continuous variables and 25 binary variables based on the augmented covariance matrix Σ_1, where Σ_1 is a 50×50 CS matrix. In the CS covariance structure, let $\Sigma_{1,ij} = \sigma^2\rho^{1(i\neq j)}$ with $\rho = 0.1$,$\rho = 0.5$,$\rho = 0.8$ and $\sigma^2 = 3$. They represent small correlation, moderate correlation and strong correlation, respectively. The mean model is $Ey_{ij} = \beta_0$,$i = 1, ..., n, j = 1, ..., p$. For simplicity, here I choose $\beta_0 = 0$. In another setting, we define Σ_1 to be a 50×50 unstructured covariance matrix. The algorithm we use to generate such an unstructured Σ_1 is as follows:

(1). generate a 50×50 matrix A whose elements are random numbers with uniform distribution $U(-1000, 1000)$

(2).do a singular value decomposition to A, that is $A = S * V * D$, where S and D are 50×50 orthogonal matrices, V is a 50×50 diagonal matrix

(3). generate a 50×50 diagonal matrix V^* whose diagonal elements $V^*_{ii}, i = 1, ..., 50$ are random numbers with uniform distribution $U(0.1, 1000)$

(4). Let $B = S * V^* * D$, then B is a positive definite matrix with eigenvalues $V^*_{ii}, i = 1, ..., 50$

(5). Let $\Sigma_1 = \frac{B+B'}{2}$, this guarantees Σ_1 is a symmetric positive definite matrix

I specified the CS hierarchical center prior distributions for Σ. Prior information for the unknown parameters are given as follows:

$$\beta_0 \sim N(0, 1000) \tag{6.1}$$

$$p(\sigma^2) \sim Inv - gamma(1, 1) \tag{6.2}$$

$$p(\rho) \propto 1 \tag{6.3}$$

$$v \sim I_{(v>100)}\Gamma(1000, 10000) \tag{6.4}$$

$$p(\beta_0, \sigma^2, \rho, v) \propto p(\beta_0)p(\sigma^2, \rho)p(v) \tag{6.5}$$

Then we explore two missing data mechanisms. In the first mechanism M1, the first 24 continuous variables $y_1, y_2, ..., y_{24}$ and the first 24 categorical variables $y_{26}, ..., y_{49}$ are missing 25% of the time completely at random, while y_{25} and y_{50} are missing according to a logistic

regression model. Specifically, I assume:

$$
\begin{aligned}
&p(y_1 = missing) = 0.25 \\
&p(y_2 = missing) = 0.25 \\
&... \\
&p(y_{24} = missing) = 0.25 \\
&logit[p(y_{25}\ missing)] = l_0 + l_1 y_1 + ... + l_{24} y_{24} \qquad (6.6) \\
&among\ observed\ y_i's, i = 1, ...24 \\
&p(y_{26} = missing) = 0.25 \\
&... \\
&p(y_{49} = missing) = 0.25 \\
&logit[p(y_{.50}\ missing)] = r_0 + r_1 y_{25} + ... + r_{24} y_{49} \qquad (6.7) \\
&among\ observed\ y_i's, i = 1, ...49
\end{aligned}
$$

where l_i, r_i,$i = 1, ..., 24$ are drawn from N(0,1) and then fixed throughout the simulation. l_0 and r_0 are constants that can be used to adjust the missing rates of y_{25} and y_{50}. Here we choose l_0 and r_0 to assure that the missing rates of y_{25} and y_{50} are around 25%. Since the role of l_0 and r_0 is to manage total missing percentage while keeping the order of plausibility of missingness, the following ad-hoc method can be used. Let's take l for example, If we know all the values of $l's$, we can calculate $p(y_{25} = missing)$ for each y's. Therefore, we can get the mean of $p(y_{25} = missing) = 0.5$ by choosing $l_0 = log(\frac{0.5}{1-0.5}) = 0$.

However, choosing $l_0 = log(\frac{p(y_{25}=missing)}{1-p(y_{25}=missing)})$ may not be appropriate for the case that $p(y_{25} = missing) \neq 0.5$. It is because the distribution of $y_{25} = missing$ is not symmetric but right-skewed considering $p(y_{25} = missing) < 0.5$. So the mean of $p(y_{25} = missing)$ tends to be larger than the median of $p(y_{25} = missing)$. On the other hand, the distribution of $y_{25} = missing$ is left-skewed if $p(y_{25} = missing) < 0.5$. So the mean of $p(y_{25} = missing)$ tends to be smaller than the median of $p(y_{25} = missing)$. Therefore, we should consider the gap between the mean and median of $p(y_{25} = missing)$ to adjust the value of l_0. We define

the adjusting value $\omega = Z_\omega.\sigma_\omega$, where σ_ω is the standard deviation of $l_0 + l_1y_1 + ... + l_{24}y_{24}$. Z_ω is multiplied by σ_ω to standardize the unit of $l_0 + l_1y_1 + ... + l_{24}y_{24}$. Then l_0 can be taken from $l_0 = log(\frac{p(y_{25}=missing)}{1-p(y_{25}=missing)}) + \omega$. Please note that we may need to take several ad-hoc trials to make the adjustment since the real percentage of missing items can vary randomly across different data sets. However, this adjustment approach is useful because it can handle any choice of $l's$. For the r's, we can also apply the above approach to choose the value of r_0. Technically, M1 is an Missing At Random (MAR) mechanism. However,since the correlations between y's are all positive and the coefficients of the logistic regression are distributed symmetrically around zero, prediction errors in one direction tended to be canceled by prediction errors in the other direction, thus the missing data mechanism M1 is actually close to MCAR. We need to consider an ignorable missing data mechanism that departs more substantially from MCAR. Section 6.1.2 illustrates this fact according to the simulation output. The second mechanism, M2, is similar to M1 except we use absolute values of normal random numbers to be the logistic regression coefficients, that is:

$$
\begin{aligned}
&p(y_1 = missing) = 0.25 \\
&p(y_2 = missing) = 0.25 \\
&... \\
&p(y_{24} = missing) = 0.25 \\
&logit[p(y_{25}\ missing)] = l_0 + |l_1|y_1 + ... + |l_{24}|y_{24} \qquad (6.8) \\
&among\ observed\ y_i's, i = 1, ...24 \\
&p(y_{26} = missing) = 0.25 \\
&... \\
&p(y_{49} = missing) = 0.25 \\
&logit[p(y_{50}\ missing)] = r_0 + |r_1|y_{25} + ... + |r_{24}|y_{49} \qquad (6.9) \\
&among\ observed\ y_i's, i = 1, ...49
\end{aligned}
$$

where l_i, r_i,$i = 1, ..., 24$ are drawn from N(0,0.5) and then fixed throughout the simulation.

We take the absolute values of $l_i's$ and $r_i's$ to avoid a canceling effect across variables. As before, l_0 and r_0 are constants that can be used to adjust the missing rates of y_{25} and y_{50} to be around 25%. All $l_i's$ and $r_i's$ are fixed throughout the simulation process.

The following table shows the combinations used in the simulation study.

6.1: Combinations of the simulation

# of observations (n)	# of variables (p)	centered var-cov matrix (CS,UN)	missingness mechanisms (M1,M2)
100	50	CS($\rho = 0.1, 0.5, 0.8$),UN	M1,M2
300	50	CS($\rho = 0.1, 0.5, 0.8$),UN	M1,M2

According to the above table, we may need to carry out $4 \times 2 \times 2 = 16$ combinations of simulation, which will induce high computation burden to our simulation process. Due to the cost of computation time, we plan to perform only 75 replications, which can be expected to produce a margin of error of $1.96 \times \sqrt{\frac{0.95 \times (1-0.95)}{75}} \approx 4.9\%$ for 95% coverage statistics. For each of the simulated data sets, 90000 iterations of the Metropolis-Hastings algorithm are generated with the maximum likelihood estimate as a start point. The first 30000 iterations is used as the "burn-in" period. We can use Gelman-Rubin statistic and time-series plot to monitor the convergence of the MCMC algorithm. To "thin" the chain, we take every 3 iterations in order to reduce the sample autocorrelations, which leaves us 20001 iterations. Five imputed data values are taken at iterations 18000,18500,19000,19500 and 20000 of the Metropolis-Hastings algorithm after earlier exploration revealed the autocorrelation between the Metropolis-Hastings algorithm draws of lag 500.

Since the percentage of missing items is chosen to be 25% for all variables, the number of complete cases is very small. When $n = 100$, around 2/3 of the seventy-five data sets do not include any complete cases. Even when $n = 300$, less than half of the seventy-five data sets have complete cases. So it's not possible to apply complete-case analysis in the data sets if we want to include all the variables in the analysis.

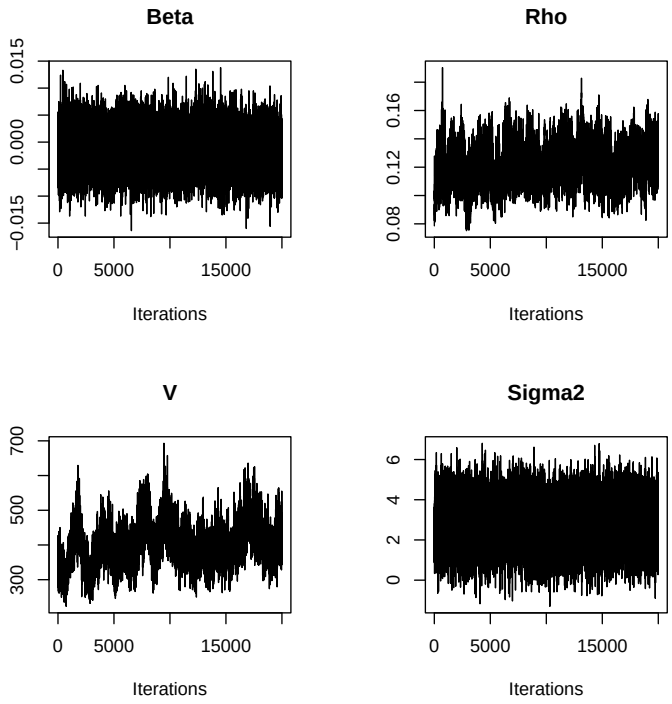
Then we can use the inferences about the mean of y_{25}, the proportion of 1's of y_{50} to

check the efficiency of our multiple imputation algorithm. In addition to results from both our parametric family approach and the rounding method (Bernaards et al 2007), we also present results from available-case analysis. The result based on complete-case analysis is not included because of very small sample size. Moreover, it is common that the imputation model is larger than the analysis model, so available-case analysis can also be considered as complete-case analysis from the analyst's viewpoint.

6.1.1 Simulation findings with moderate sample size

Figure 6.1 and 6.2 show the time-series plots of parameters related to y_{25} and y_{50} with $n = 300, p = 50$, compound symmetry (CS) generating structure with $\rho = 0.1$ and missing data mechanism M1. Since we have more than 50 parameters in our parametric family model, it is burdensome to list the plots of all parameters, but all other parameters display similar patterns as seen in these plots. Plots from other data generated under different covariance structure or different missing data mechanisms show similar patterns.

6.1: Time-series plots of parameters related to y_{25} and y_{50} with $n = 300, p = 50$,CS structure with $\rho = 0.1$, and missing data mechanism M1



6.2: Time-series plots of parameters related to y_{25} and y_{50} with $n = 300, p = 50$,CS structure with $\rho = 0.1$, and missing data mechanism M1

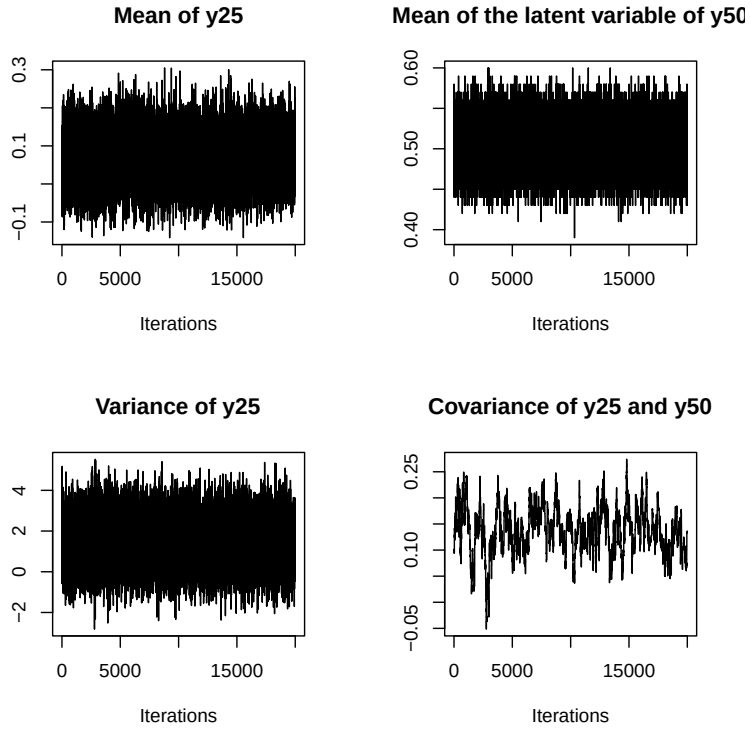


Table 6.2 shows results of inference on the mean of y_{25} and y_{50} while data are generated based on an unstructured covariance matrix under missing data mechanisms M1. The first and second columns represent the Monte Carlo mean and standard error. The Monte Carlo mean and standard error for both the parametric family approach and rounding approach are calculated based on multiple imputation inference (Rubin 1987). The third column displays an actual 95% coverage rate measured by the number of data sets whose 95% confidence interval covers a true parameter value. For missing data mechanism M1, mean estimates from the available-case analysis are more biased than those from the two imputation methods, resulting in lower 95% coverage rates for both mean of y_{25} and mean of y_{50}. Imputations based

on the rounding method and our parametric family method centered around a compound symmetry structure show little bias in mean estimates with good 95% coverage probabilities. However, the results of our method are slightly better. For the inference on the mean of y_{50}, the rounding method tends to have larger bias and lower coverage rate than those of parametric family method.

6.2: The means of y_{25} and y_{50} with n=300,p=50, missing data mechanism M1 and data are generated based on an unstructured covariance matrix

	y_{25}			y_{50}		
	M.C. Mean	M. C. S.E	Act 95% Coverage	M. C. Mean	M. C. S.E	Act 95% Coverage
True	0.0000			0.7		
All data	0.0079	0.0545	0.93	0.73	0.0556	0.94
Available Cases	0.0888	0.0587	0.77	0.80	0.0688	0.83
Rounding	0.0108	0.0701	0.92	0.78	0.0685	0.88
Parametric Family	0.0142	0.0689	0.94	0.74	0.0504	0.94

Table 6.3, 6.4 and 6.5 list the results of inference on the mean of y_{25} and y_{50} while data are generated based on an compound symmetry (CS)covariance matrix with correlation parameter ρ values to be 0.1,0.5 and 0.8 under missing data mechanism M1. For the inference of y_{25}, we can find that the available-case approach tends to induce larger bias and lower 95% coverage rate compared with other two approaches. Our parametric family method performs a little better than the rounding method although the difference is not big. However, when it comes to the inference of y_{50}, the binary variable, our parametric family approach results in much less bias for mean estimates and closer to 95% coverage probability than other two methods. The Monte Carlo standard deviations of the three approaches are not quite different.

Table 6.6, 6.7, 6.8 and Table 6.9 list the results of inference on the mean of y_{25} and y_{50} under missing data mechanism M2. They have similar trend to those in Table 6.2, 6.3, 6.4 and Table 6.5 but the available-case approach create larger bias and worse 95% coverage

6.3: The means of y_{25} and y_{50} with n=300,p=50, missing data mechanism M1 and data are generated based on an CS covariance matrix with $\rho = 0.1$

	y_{25}			y_{50}		
	M.C. Mean	M. C. S.E	Act 95% Coverage	M. C. Mean	M. C. S.E	Act 95% Coverage
True	0.0000			0.7		
All data	0.0016	0.0554	0.94	0.68	0.0585	0.94
Available Cases	0.0778	0.0601	0.81	0.56	0.0707	0.74
Rounding	0.0132	0.0652	0.91	0.60	0.0669	0.81
Parametric Family	0.0093	0.0636	0.94	0.67	0.0684	0.93

6.4: The means of y_{25} and y_{50} with n=300,p=50, missing data mechanism M1 and data are generated based on an CS covariance matrix with $\rho = 0.5$

	y_{25}			y_{50}		
	M.C. Mean	M. C. S.E	Act 95% Coverage	M. C. Mean	M. C. S.E	Act 95% Coverage
True	0.0000			0.7		
All data	0.0033	0.0571	0.94	0.71	0.0498	0.93
Available Cases	0.0864	0.0612	0.79	0.68	0.0716	0.95
Rounding	0.0065	0.0627	0.94	0.62	0.0682	0.88
Parametric Family	0.0048	0.0584	0.93	0.69	0.0613	0.94

6.5: The means of y_{25} and y_{50} with n=300,p=50, missing data mechanism M1 and data are generated based on an CS covariance matrix with $\rho = 0.8$

	y_{25}			y_{50}		
	M.C. Mean	M. C. S.E	Act 95% Coverage	M. C. Mean	M. C. S.E	Act 95% Coverage
True	0.0000			0.7		
All data	0.0062	0.0664	0.95	0.73	0.0715	0.96
Available Cases	0.0759	0.0769	0.83	0.56	0.0717	0.85
Rounding	0.0125	0.0652	0.91	0.60	0.0589	0.81
Parametric Family	0.0094	0.0621	0.94	0.67	0.0784	0.93

probability than those under mechanism M1.

6.6: The means of y_{25} and y_{50} with n=300,p=50, missing data mechanism M2 and data are generated based on an unstructured covariance matrix

	y_{25}			y_{50}		
	M.C. Mean	M. C. S.E	Act 95% Coverage	M. C. Mean	M. C. S.E	Act 95% Coverage
True	0.0000			0.7		
All data	0.0048	0.0504	0.96	0.69	0.0613	0.94
Available Cases	0.1697	0.0669	0.70	0.60	0.0672	0.73
Rounding	0.0218	0.0825	0.93	0.62	0.0851	0.83
Parametric Family	0.0125	0.0797	0.93	0.68	0.0719	0.94

6.1.2 Simulation findings with small sample size

It is often the case that some model parameters can be inestimable or almost inestimable when the number of observations is small compared to the number of variables. When we study scenarios with 100 observations and 50 variables, our MCMC algorithm may fail to converge or converge slowly with non-informative priors. Therefore, we need to apply a more

6.7: The means of y_{25} and y_{50} with n=300,p=50, missing data mechanism M2 and data are generated based on a CS covariance matrix with $\rho = 0.1$

	y_{25}			y_{50}		
	M.C. Mean	M. C. S.E	Act 95% Coverage	M. C. Mean	M. C. S.E	Act 95% Coverage
True	0.0000			0.7		
All data	0.0034	0.0664	0.95	0.73	0.0715	0.96
Available Cases	0.1844	0.0669	0.71	0.59	0.0738	0.74
Rounding	0.0150	0.0652	0.91	0.63	0.0796	0.87
Parametric Family	0.0125	0.0721	0.93	0.69	0.0721	0.95

6.8: The means of y_{25} and y_{50} with n=300,p=50, missing data mechanism M2 and data are generated based on a CS covariance matrix with $\rho = 0.5$

	y_{25}			y_{50}		
	M.C. Mean	M. C. S.E	Act 95% Coverage	M. C. Mean	M. C. S.E	Act 95% Coverage
True	0.0000			0.7		
All data	0.0010	0.0693	0.95	0.68	0.0642	0.94
Available Cases	0.1771	0.0979	0.79	0.58	0.0799	0.76
Rounding	0.0085	0.0656	0.92	0.66	0.0772	0.89
Parametric Family	0.0123	0.0784	0.94	0.66	0.0831	0.92

6.9: The means of y_{25} and y_{50} with n=300,p=50, missing data mechanism M2 and data are generated based on a CS covariance matrix with $\rho = 0.8$

	y_{25}			y_{50}		
	M.C. Mean	M. C. S.E	Act 95% Coverage	M. C. Mean	M. C. S.E	Act 95% Coverage
True	0.0000			0.7		
All data	0.0058	0.0586	0.94	0.72	0.0552	0.94
Available Cases	0.2565	0.0795	0.65	0.51	0.0753	0.70
Rounding	0.0154	0.0736	0.92	0.59	0.0689	0.83
Parametric Family	0.0068	0.0725	0.93	0.68	0.0726	0.96

informative prior to the parametric family model. The new priors are chosen as follows:

$$\beta_0 \sim N(0, 100) \tag{6.10}$$

$$p(\sigma^2) \sim Inv - gamma(2, 2.5) \tag{6.11}$$

$$p(\rho) \propto 1 \tag{6.12}$$

$$v \sim I_{(v>100)}\Gamma(50, 50) \tag{6.13}$$

$$p(\beta_0, \sigma^2, \rho, v) \propto p(\beta_0)p(\sigma^2, \rho)p(v) \tag{6.14}$$

We also tried several other different informative priors and find the results are similar.

Figure 6.3 and 6.4 show time series plots of parameters related to y_{25} and y_{50} with $n = 100, p = 50$, compound symmetry (CS) generating structure with $\rho = 0.1$ and missing data mechanism M1. All time-series plots show that the parameters all converge after 90000 iterations. Autocorrelation plots also show fast disappearing dependence. Plots from other simulation settings display similar patterns.

6.3: Time-series plots of parameters related to y_{25} and y_{50} with $n = 100, p = 50$,CS structure with $\rho = 0.1$, and missing data mechanism M1

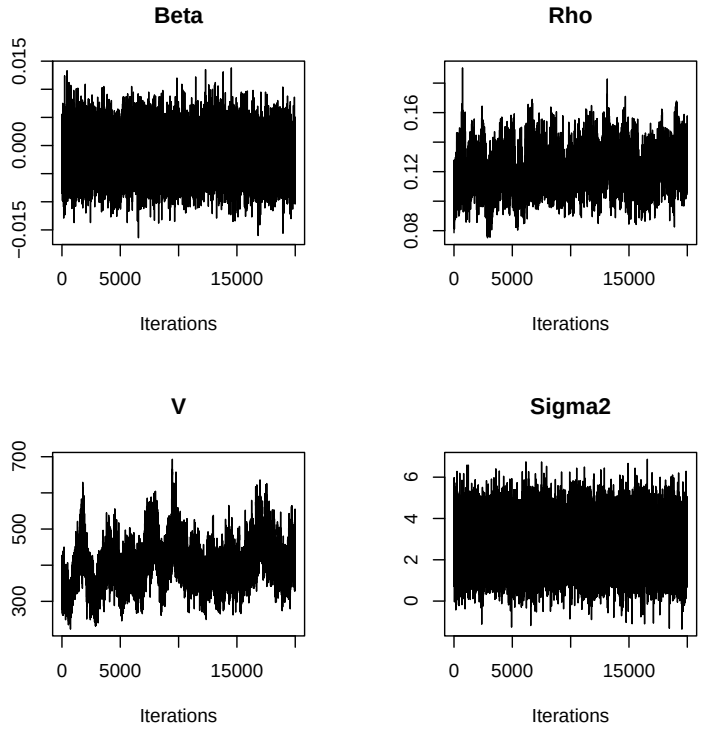


Table 6.10, 6.11, 6.12 and Table 6.13 show inference about the means of y_{25} and y_{50} with n=100,p=50 and missing data mechanisms M1. The M.C. standard errors are about more than twice wide as those from data with 300 observations due to smaller sample size. As for the inference about the mean of y_{25}, even available-case method shows good 95% coverage probability. However, available-case method gives better 95% coverage probability than rounding method for the inference about the mean of y_{50} in some tables. This may be due to the skewed distribution of the binary variable y_{50}.

The results for available-case analysis suggest that the missing data mechanism M1 applied to these data sets is "close" to MCAR in the sense that a procedure (available-case

6.10: The means of y_{25} and y_{50} with n=100,p=50, missing data mechanism M1 and data are generated based on an unstructured covariance matrix

	y_{25}			y_{50}		
	M.C. Mean	M. C. S.E	Act 95% Coverage	M. C. Mean	M. C. S.E	Act 95% Coverage
True	0.0000			0.7		
All data	0.0146	0.1112	0.96	0.68	0.0682	0.95
Available Cases	-0.0116	0.1385	0.93	0.62	0.0797	0.89
Rounding	-0.0095	0.1356	0.92	0.57	0.0813	0.87
Parametric Family	0.0138	0.1291	0.94	0.71	0.0711	0.96

6.11: The means of y_{25} and y_{50} with n=100,p=50, missing data mechanism M1 and data are generated based on a CS covariance matrix with $\rho = 0.1$

	y_{25}			y_{50}		
	M.C. Mean	M. C. S.E	Act 95% Coverage	M. C. Mean	M. C. S.E	Act 95% Coverage
True	0.0000			0.7		
All data	0.0090	0.1011	0.96	0.73	0.0673	0.93
Available Cases	0.0148	0.1765	0.95	0.59	0.0855	0.87
Rounding	0.0115	0.1284	0.92	0.55	0.0772	0.85
Parametric Family	0.0064	0.1445	0.94	0.72	0.0733	0.92

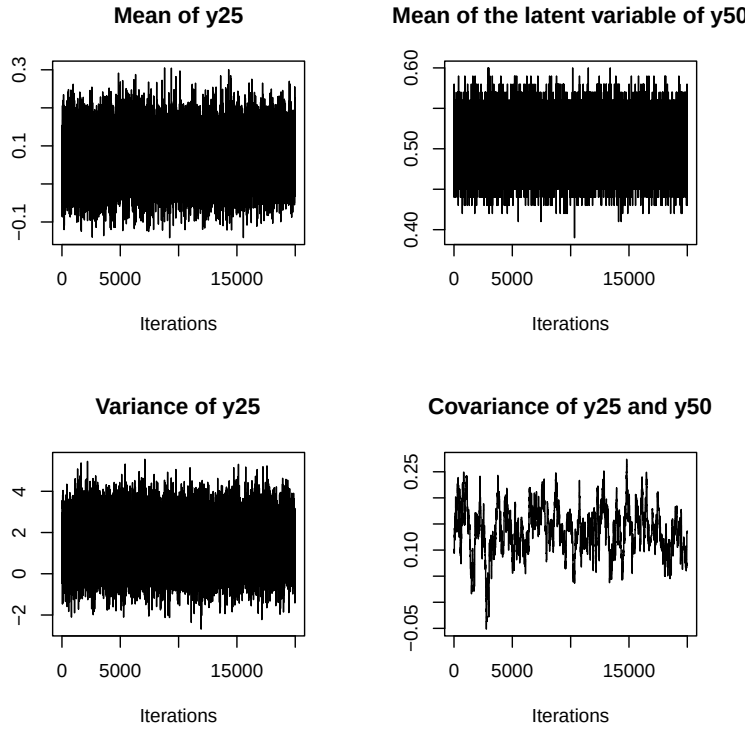
6.12: The means of y_{25} and y_{50} with n=100,p=50, missing data mechanism M1 and data are generated based on a CS covariance matrix with $\rho = 0.5$

	y_{25}			y_{50}		
	M.C. Mean	M. C. S.E	Act 95% Coverage	M. C. Mean	M. C. S.E	Act 95% Coverage
True	0.0000			0.7		
All data	0.0025	0.0978	0.96	0.68	0.0562	0.93
Available Cases	0.1459	0.2857	0.91	0.55	0.0997	0.91
Rounding	0.0191	0.1636	0.93	0.55	0.0772	0.88
Parametric Family	0.0113	0.1492	0.94	0.71	0.0711	0.96

6.13: The means of y_{25} and y_{50} with n=100,p=50, missing data mechanism M1 and data are generated based on a CS covariance matrix with $\rho = 0.8$

	y_{25}			y_{50}		
	M.C. Mean	M. C. S.E	Act 95% Coverage	M. C. Mean	M. C. S.E	Act 95% Coverage
True	0.0000			0.7		
All data	0.0068	0.1394	0.96	0.68	0.0682	0.94
Available Cases	0.0771	0.1999	0.91	0.58	0.0799	0.76
Rounding	0.0265	0.1756	0.94	0.62	0.0754	0.89
Parametric Family	0.0151	0.1593	0.95	0.68	0.0692	0.95

6.4: Time-series plots of parameters related to y_{25} and y_{50} with $n = 100, p = 50$,CS structure with $\rho = 0.1$, and missing data mechanism M1



analysis) designed to work under MCAR mechanism actually works well with this missing data mechanism M1. Missing values on $y_1, ..., y_{24}, y_{26}, ..., y_{50}$ are generated from an MCAR mechanism. In addition, the randomly generated $l's$ and $r's$ describing the missingness on y_{25} and y_{50} are allowed to be either positive or negative, so that these coefficients could have a canceling effect on another in this case where all variables are positively related according to the covariance structure being explored. Therefore we need to consider the missing data mechanism M2.

6.14: The means of y_{25} and y_{50} with n=100,p=50, missing data mechanism M2 and data are generated based on an unstructured covariance matrix

	y_{25}			y_{50}		
	M.C. Mean	M. C. S.E	Act 95% Coverage	M. C. Mean	M. C. S.E	Act 95% Coverage
True	0.0000			0.7		
All data	0.0081	0.0904	0.96	0.69	0.0723	0.94
Available Cases	0.2287	0.1169	0.55	0.60	0.0829	0.76
Rounding	0.0508	0.1565	0.93	0.62	0.0851	0.83
Parametric Family	0.0452	0.1787	0.95	0.68	0.0819	0.94

6.15: The means of y_{25} and y_{50} with n=100,p=50, missing data mechanism M2 and data are generated based on a CS covariance matrix with $\rho = 0.1$

	y_{25}			y_{50}		
	M.C. Mean	M. C. S.E	Act 95% Coverage	M. C. Mean	M. C. S.E	Act 95% Coverage
True	0.0000			0.7		
All data	0.0103	0.1167	0.94	0.67	0.0706	0.93
Available Cases	0.2022	0.1377	0.71	0.57	0.0788	0.74
Rounding	0.0262	0.1575	0.91	0.60	0.0738	0.85
Parametric Family	0.0271	0.1290	0.93	0.69	0.0774	0.94

6.16: The means of y_{25} and y_{50} with n=100,p=50, missing data mechanism M2 and data are generated based on a CS covariance matrix with $\rho = 0.5$

	y_{25}			y_{50}		
	M.C. Mean	M. C. S.E	Act 95% Coverage	M. C. Mean	M. C. S.E	Act 95% Coverage
True	0.0000			0.7		
All data	-0.0021	0.0993	0.95	0.71	0.0692	0.94
Available Cases	0.2971	0.1439	0.75	0.46	0.0814	0.69
Rounding	0.0365	0.1556	0.93	0.57	0.0838	0.81
Parametric Family	-0.0283	0.1394	0.94	0.66	0.0859	0.92

6.17: The means of y_{25} and y_{50} with n=100,p=50, missing data mechanism M2 and data are generated based on a CS covariance matrix with $\rho = 0.8$

	y_{25}			y_{50}		
	M.C. Mean	M. C. S.E	Act 95% Coverage	M. C. Mean	M. C. S.E	Act 95% Coverage
True	0.0000			0.7		
All data	0.0064	0.1147	0.94	0.68	0.0722	0.95
Available Cases	0.2563	0.1191	0.61	0.48	0.0885	0.65
Rounding	0.0373	0.1342	0.93	0.54	0.0753	0.80
Parametric Family	0.0344	0.1266	0.92	0.70	0.0811	0.95

Table 6.14, 6.15, 6.16 and Table 6.17 show the results of inference about the means of y_{25} and y_{50} under the missing data mechanism M2. According to these tables, available-case analysis are more biased in the mean estimates than other methods. The available-case analysis also has lower coverage rates. The mean estimates of y_{25} from the parametric family method and the rounding method show little bias and good coverage rates. But for the inference about the mean of y_{50}, the rounding method gives larger bias and worse coverage probability than those from our parametric family approach. No matter what the correlation is in the compound symmetry structure, the parametric family approach has a good behavior. Even if the real covariance matrix is unstructured, the parametric family approach using a compound symmetry center can still work well.

6.2 Simulation studies for the factor model

In this section, we carry out a simulation study based on data generated from a simple factor structure. After choosing missing items from an ignorable or nonignorable missing data mechanism, we applied factor models both with a correct number of factors and with an incorrect number of factors to impute missing items. The results are then compared with the results from available-case analysis and multiple imputation based on a multivariate normal model using rounding strategy.

For the simulation we choose a simple factor structure for data and check how the factor model works if we correctly specify the number of factors or if we incorrectly specify the number of factors. Because data are generated to be consistent with the model underlying the proposed imputation method, this case should be especially favorable for the proposed method when the number of factors assumed is also correct.

To represent this situation, we choose a simple factor structure only with high loadings (0.8) and zero loadings (0). For example, if we assume a five-factor structure, we divide the number of variables (p) by the number of factors (k). Then we make the first p/k variables have high loadings on the first factor, the second p/k variables have high loadings on the

second factor, and so on. So the factor loading matrix is as follows:

$$\Lambda = \begin{pmatrix} 0.8 & ... & 0.8 & 0 & ... & 0 & 0 & ... & 0 \\ 0 & ... & 0 & 0.8 & ... & 0.8 & 0 & ... & 0 \\ ... & ... & ... & ... & ... & ... & ... & ... & ... \\ 0 & ... & ... & ... & ... & 0 & 0.8 & ... & 0.8 \end{pmatrix} \tag{6.15}$$

In addition, we generate the data by a multivariate normal distribution with the mean 0 and variance-covariance matrix $\Lambda'\Lambda + \tau$. Here I choose the diagonal elements of τ to be 1.

To represent a moderate or large sample size, we assumed that the sample size to be 100 or 300. Following the routine of section 6.1, we assume $p = 50$ variables are measured. The 50 variables are made up of 25 continuous and 25 binary variables. We also assume that true underlying factor structure includes 5 or 10 factors. In a real application, we usually don't know the correct number of factors, so it is possible to use an incorrect number of factors in the model. Therefore, we can explore the performance of the factor model based on 10 factors applied to data generated by 5 true factors as well as the performance of the factor model based on 5 factors applied to data based on 10 true factors. These represent the case that our imputation model is underparameterized or overparameterized, respectively. Meanwhile, we can explore the performance of the factor model based on correct factor numbers as well. We explore the performance of the factor model approach under two missingness patterns: M1 and M2. M1 and M2 are defined in section 6.1. The following table shows the combinations used in the simulation study.

6.18: Combinations of the simulation

# of observations (n)	# of variables (p)	# of true factors	# of assumed factors	missingness mechanisms (M1,M2)
100	50	5	5,10	M1,M2
		10	5,10	M1,M2
300	50	5	5,10	M1,M2
		10	5,10	M1,M2

For simplicity, we just follow the simulation settings listed in section 6.1, 75 replications are generated due to the computation burden. 75 data sets are expected to have an error standard deviation of 4.9% for 95% coverage of true parameters. For each of simulated data sets, 12000 iterations of Gibbs sampler are generated with the maximum likelihood estimate as a starting point. The first 2000 iterations is treated as a "burn-in" period. Five imputed data values are taken at iterations 11000,11250,11500,11750 and 12000 of the Gibbs sampler after earlier exploration revealed little autocorrelation between Gibbs sampler draws of lag 250. The inferences about the mean of y_{25}, the proportion of 1's of y_{50} is used to check the validity of the factor analysis approach. The result is compared with those of rounding method or available-case analysis. If $n = 300$, it is possible to apply the factor model with noninformative priors. However, when $n = 100$, more informative priors are necessary for the Gibbs sampler to work.

6.2.1 Simulation findings with moderate sample size

Figure 6.5 shows time-series plots of parameters related to y_{25} and y_{50} under the factor model with ten factors for a data set with five factors. Since there are so many simulation combinations and parameters in this model, it is not possible to show all parameters here, but all other parameters displayed similar patterns as seen in these plots. In the factor model, there is not a unique maximum likelihood estimate for ϕ, the factor score matrix, because ϕ multiplies any orthogonal matrix will produce the same variance-covariance matrix. It means that generated $\phi's$ are only stable up to multiplication of an orthogonal matrix. Instead of checking the convergence of ϕ directly, we check $\phi\phi'$ which is unique. Plots from other data generated under either five or ten factors and plots from factor models assuming the incorrect number of factors showed similar patterns.

Table 6.19 and Table 6.20 show results of inferences on the means of y_{25} and y_{50} under the factor model with n=300, p=50, k=5. The first and second columns represent the Monte Carlo mean and standard error. The Monte Carlo means and standard errors for both the factor model and the rounding method are calculated based on the multiple imputation

inference (Rubin 1987). The third column denotes an actual 95% coverage rate measured by the number of data sets whose 95% confidence interval covers a true parameter value. Under missing data mechanism M2, mean estimates from the available-case analysis are biased more than those from imputation methods and the standard errors are not much different, resulting lower 95% coverage rates. However, the available-case analysis has small bias and acceptable coverage rate under missing data mechanism M1. It's probably due to M1's "close to MCAR" property. The rounding method gives small bias and good 95% coverage rate on the inference about the mean of y_{25} under both missing data mechanisms. but not on that of y_{50}. However, we can find imputations based on the factor model with the correct number of factors show less bias in mean estimate of y_{50} and better 95% coverage than all other scenarios.

6.5: Time-series plots of parameters related to y_{25} and y_{50} under the factor model with n=300, p=50, k=10, and missing data mechanism M1

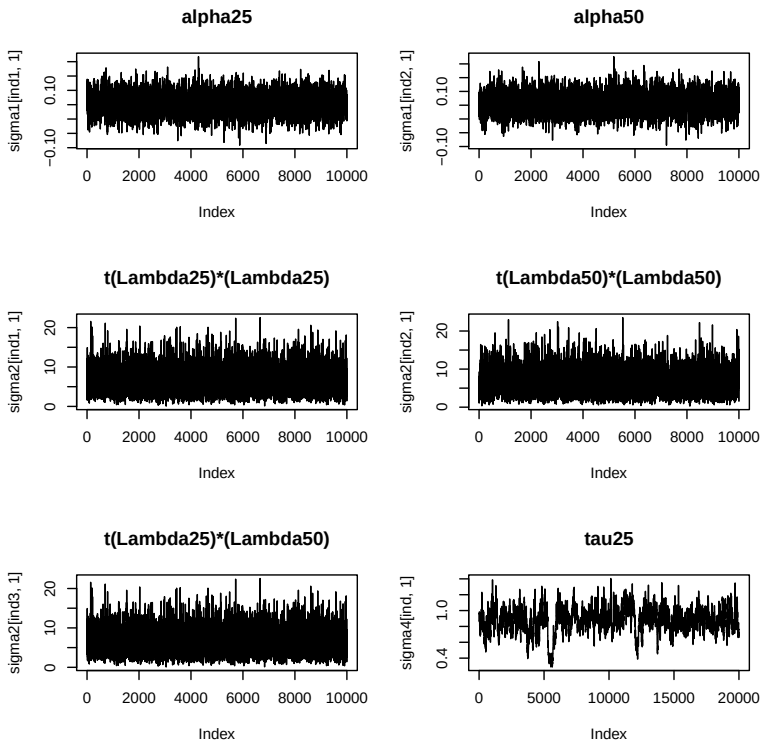


Table 6.21 and Table 6.22 show results of inferences on the means of y_{25} and y_{50} under the factor model with n=300, p=50, k=10. Compared with Table 6.19 and Table 6.20, we can find an overparameterized model (Table 6.19,6.20) results in little bias for mean estimates of y_{25} and y_{50} but an underparameterized model (Table 6.21,6.22) results in more biased mean estimates with lower than the nominal 95% coverage rate when we apply the incorrect number of factors in our model. But the underparameterized factor model still has better behaviors than the rounding method on the inference of y_{50}.

6.19: The means of y_{25} and y_{50} under the factor model with n=300, p=50, k=5, and missing data mechanism M1

	y_{25}			y_{50}		
	M.C. Mean	M. C. S.E	Act 95% Coverage	M. C. Mean	M. C. S.E	Act 95% Coverage
True	0.0000			0.7		
All data	0.0073	0.0495	0.93	0.68	0.0601	0.93
Available Cases	0.0222	0.0519	0.89	0.65	0.0653	0.91
Rounding Factor	0.0106	0.0530	0.92	0.60	0.0647	0.88
True (k=5)	0.0035	0.0516	0.93	0.66	0.0701	0.92
False (k=10)	0.0089	0.0582	0.91	0.65	0.0656	0.91

6.20: The means of y_{25} and y_{50} under the factor model with n=300, p=50, k=5, and missing data mechanism M2

	y_{25}			y_{50}		
	M.C. Mean	M. C. S.E	Act 95% Coverage	M. C. Mean	M. C. S.E	Act 95% Coverage
True	0.0000			0.7		
All data	-0.0052	0.0506	0.95	0.72	0.0586	0.93
Available Cases	-0.2678	0.0627	0.63	0.49	0.0676	0.71
Rounding Factor	0.0066	0.0633	0.94	0.57	0.0628	0.83
True (k=5)	-0.0049	0.0704	0.94	0.69	0.0633	0.94
False (k=10)	0.0036	0.0688	0.92	0.66	0.0645	0.92

6.21: The means of y_{25} and y_{50} under the factor model with n=300, p=50, k=10, and missing data mechanism M1

	y_{25}			y_{50}		
	M.C. Mean	M. C. S.E	Act 95% Coverage	M. C. Mean	M. C. S.E	Act 95% Coverage
True	0.0000			0.7		
All data	0.0018	0.0580	0.96	0.69	0.0595	0.93
Available Cases	0.0567	0.0591	0.84	0.56	0.0613	0.78
Rounding Factor	0.0147	0.0649	0.92	0.58	0.0659	0.86
False (k=5)	0.0389	0.0646	0.88	0.65	0.0624	0.91
True (k=10)	0.0112	0.0611	0.92	0.67	0.0677	0.93

6.22: The means of y_{25} and y_{50} under the factor model with n=300, p=50, k=10, and missing data mechanism M2

	y_{25}			y_{50}		
	M.C. Mean	M. C. S.E	Act 95% Coverage	M. C. Mean	M. C. S.E	Act 95% Coverage
True	0.0000			0.7		
All data	0.0029	0.0547	0.94	0.71	0.0574	0.95
Available Cases	0.2271	0.0684	0.55	0.50	0.0613	0.67
Rounding Factor	0.0125	0.0609	0.96	0.54	0.0676	0.80
False (k=5)	0.0539	0.0576	0.78	0.62	0.0654	0.89
True (k=10)	0.0090	0.0594	0.93	0.69	0.0622	0.93

6.2.2 Simulation findings with small sample size

As we describe in Section 6.1.2, some parameters can be inestimable or almost inestimable when the sample size is relatively small compared to the number of variables. When we study scenarios with 100 observations, our MCMC algorithm may fail to converge or converge very slowly. To make our MCMC algorithm converge faster, we use an informative prior for the factor model. Priors are chosen as follows:

$$\tau_j^2 \sim Inverse - Gamma(\frac{3}{2}, \frac{0.3}{2}) \tag{6.16}$$

$$\Lambda_j|\tau_j^2 \sim N(\Lambda_0, \frac{1}{3}\tau_j^2 I) \tag{6.17}$$

$$\alpha_j|\tau_j^2 \sim N(0, \frac{1}{3}\tau_j^2) \tag{6.18}$$

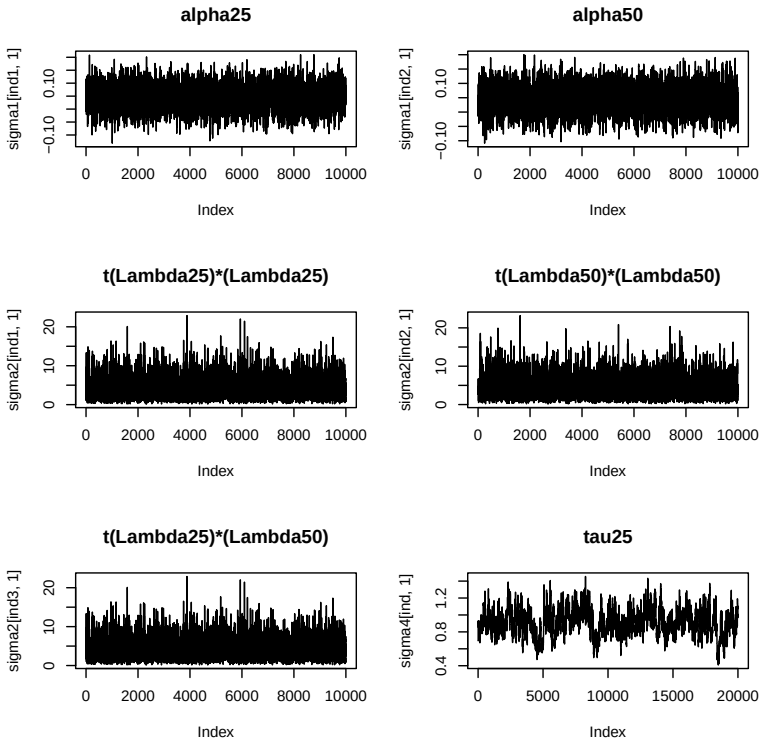
where Λ_0 is from the first k principal components. We also have tried several different for τ_j^2 and find that the results are similar, since these priors are quite weak compared to the total number of observations.

Figure 6.6 shows time series plots of parameters related to y_{25} and y_{50} under the five-factor model for a data set with 10 true factors and missing data mechanism M1. All time-series plots show that the parameters converged within 20000 iterations. Autocorrelation plots also show fast disappearing dependence. Plots from data generated from ten true factors and plots from other overparameterized or underparameterized factor model display similar patterns.

Table 6.23 and Table 6.24 show inferences about the means of y_{25} and y_{50} under the factor model with n=100, p=50, k=5. The standard errors are about two times of those with sample size 300. Under the missing data mechanism M1, all methods even available-case analysis show small biases and good 95% coverage probabilities on the inference about the mean of y_{25}. That again reveals the "close to MCAR" property of the missing data mechanism M1. But for the inference of y_{50}, both available-case analysis and rounding method give smaller nominal 95% coverage rates. The tables also show that the factor model creates little bias and good 95% coverage rate even under overparameterized scenarios. However, the factor

model with correct number of factors performs best among all the models we apply here.

6.6: Time-series plots of parameters related to y_{25} and y_{50} under the factor model with n=100, p=50, k=10, and missing data mechanism M1



6.23: The means of y_{25} and y_{50} under the factor model with n=100, p=50, k=5, and missing data mechanism M1

	y_{25}			y_{50}		
	M.C. Mean	M. C. S.E	Act 95% Coverage	M. C. Mean	M. C. S.E	Act 95% Coverage
True	0.0000			0.7		
All data	0.0070	0.1018	0.96	0.72	0.0766	0.94
Available Cases	0.0321	0.1245	0.92	0.59	0.0802	0.86
Rounding	0.0110	0.1296	0.95	0.62	0.0791	0.90
Factor True (k=5)	0.0085	0.1318	0.94	0.72	0.0821	0.94
False (k=10)	0.0134	0.1255	0.93	0.67	0.0815	0.93

6.24: The means of y_{25} and y_{50} under the factor model with n=100, p=50, k=5, and missing data mechanism M2

	y_{25}			y_{50}		
	M.C. Mean	M. C. S.E	Act 95% Coverage	M. C. Mean	M. C. S.E	Act 95% Coverage
True	0.0000			0.7		
All data	-0.0068	0.1174	0.93	0.69	0.0775	0.94
Available Cases	-0.5963	0.1248	0.47	0.53	0.0813	0.71
Rounding	-0.0137	0.1339	0.94	0.58	0.0826	0.83
Factor True (k=5)	0.0025	0.1276	0.94	0.67	0.0825	0.93
False (k=10)	-0.0190	0.1313	0.93	0.66	0.0792	0.92

Table 6.25 and Table 6.26 show inferences about the means of y_{25} and y_{50} under the factor model with n=100, p=50, k=10. The results are similar to those from Table 6.23 and Table 6.24. But we can find the underparameterized case tends to have large bias and worse 95% coverage rate than those from the overparameterized case. The factor model with correct factor number tends to have the best performance.

6.25: The means of y_{25} and y_{50} under the factor model with n=100, p=50, k=10, and missing data mechanism M1

	y_{25}			y_{50}		
	M.C. Mean	M. C. S.E	Act 95% Coverage	M. C. Mean	M. C. S.E	Act 95% Coverage
True	0.0000			0.7		
All data	0.0045	0.1107	0.95	0.68	0.0783	0.94
Available Cases	0.0247	0.1191	0.92	0.62	0.0822	0.89
Rounding Factor	0.0193	0.1235	0.93	0.60	0.0817	0.87
False (k=5)	0.0684	0.1288	0.90	0.70	0.0855	0.93
True (k=10)	0.0074	0.1249	0.94	0.66	0.0849	0.96

6.26: The means of y_{25} and y_{50} under the factor model with n=100, p=50, k=10, and missing data mechanism M2

	y_{25}			y_{50}		
	M.C. Mean	M. C. S.E	Act 95% Coverage	M. C. Mean	M. C. S.E	Act 95% Coverage
True	0.0000			0.7		
All data	0.0118	0.1189	0.94	0.67	0.0775	0.93
Available Cases	0.6177	0.1256	0.59	0.51	0.0816	0.66
Rounding Factor	0.0233	0.1248	0.92	0.56	0.0834	0.85
False (k=5)	0.0761	0.1263	0.90	0.63	0.0822	0.91
True (k=10)	0.0195	0.1291	0.94	0.66	0.0883	0.93

6.2.3 Regression analysis for simulation output

To further investigate the statistical properties of the procedure (e.g., bias in estimates of the mean of y_{25}, bias in estimates of the mean of y_{50}, 95% coverage rate of the mean of y_{25}, 95% coverage rate of the mean of y_{50}, etc.), we build a regression model for the analysis. The following tables list the response variables and possible covariates for the parametric approach.

6.27: Response variables

Response variables	Description
$bias\_y25$	The absolute value of the bias of the mean estimator for y_{25}
$bias\_y50$	The absolute value of the bias of the mean estimator for y_{50}
cp_y25	The 95% coverage probability of the mean estimator for y_{25}
cp_y50	The 95% coverage probability of the mean estimator for y_{50}

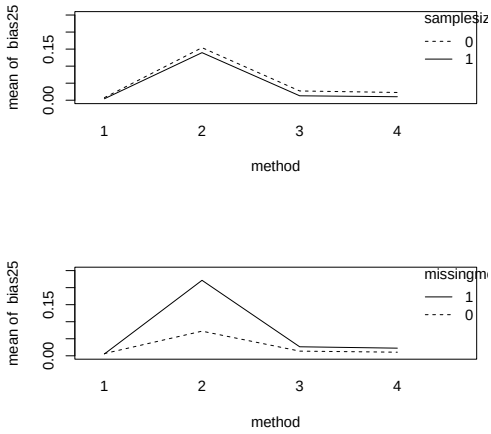
6.28: Regression covariates for parametric family approach

Regression covariates	Description
method	1 =all data (reference group) 2 =available-case method 3 =rounding method 4 =parametric family
samplesize	1 =sample size 300 0 =sample size 100
covmatrix	1=unstructured 0 = compound symmetry
missingmech	1 = missingness mechanism M2 0 = missingness mechanism M1

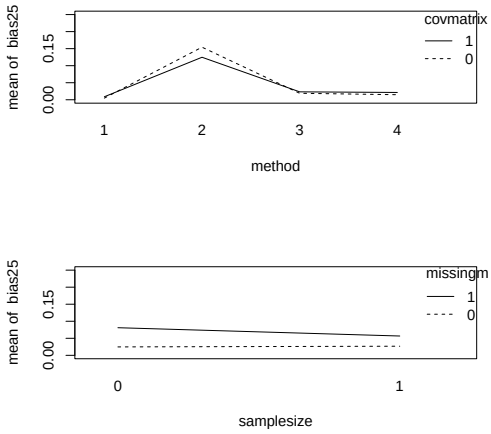
For each response variable, there is a corresponding regression model. The simulation outputs in Table 6.2-Table 6.17 are used to build the regression models. Before fitting the model, we need to check all the possible 2-way interaction effects. So we can decide whether

to include them in the model or not. For the model simplicity, we don't include any 3-way interaction terms in our model. Figure 6.5, 6.6 and Figure 6.7 show the interaction plots for the regression model of biasy25. Interaction plots for other regression models have similar patterns. From these interaction plots, we can find there may be strong interaction between method and missingness patterns. So we include this interaction term in our regression models. Since "method" is a categorical variable, we introduce dummy variables "method2","method3" and "method4", denoting available-case method, rounding method and parametric family method, respectively. Table 6.20,6.21,6.22 and Table 6.23 show the estimates of coefficients and corresponding p-values for the regression models. These tables display that the available-case method tend to have large biases and bad 95% coverage rates for the means of both continuous variable y_{25} and binary variable y_{50}. And all tables show that the available-case method has significant interaction with missing data mechanisms. The rounding method works well for the inference on mean of y_{25}, but the rounding method introduces more bias and worse 95% coverage rate on the mean of the binary variable y_{50}. Moreover, the parametric family approach is as efficient as all-case method for the means of both continuous variable y_{25} and binary variable y_{50}. We also find that the sample size factor has a somehow significant effect(p-values are a bit larger than 0.05) on the biases and 95% coverage rates of y_{25} and y_{50}. Larger sample size is more likely to obtain smaller bias and better 95% coverage rate. All above conclusions tally with the simulation output we have in Section 6.1.1 and Section 6.1.2.

6.7: Interaction plot between method and samplesize & Interaction plot between method and missingmech

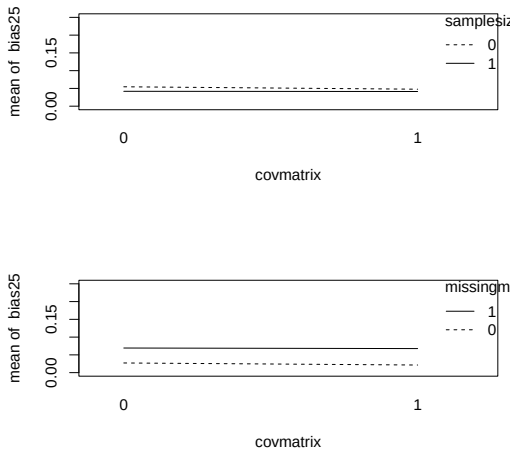


6.8: Interaction plot between method and covmatrix & Interaction plot between samplesize and missingmech



For factor model approach, we use the same definitions for response variables, but the covariates are a little bit different. Table 6.33 shows the regression covariates. Figure 6.8 and

6.9: Interaction plot between covmatrix and samplesize & Interaction plot between covmatrix and missingmech



6.29: Results of the regression model for bias_y25 (parametric family)

parameter	estimate	p-value
Intercept	0.0129	0.1458
aval-case	0.0658	$\underline{< 0.0001}^*$
rounding	0.0072	0.5337
para-family	0.0041	0.7264
samplesize	-0.0111	0.0583
missingmech	-0.0013	0.9140
covmatrix	-0.0037	0.5806
aval-case*missingmech	0.1505	$\underline{< 0.0001}$
rounding*missingmech	0.0140	0.3942
para-family*missingmech	0.0131	0.4247

* the highlighted and underscore type signifies a variable that is significant at $\alpha = 0.05$

Figure 6.9 show the interaction plots for the regression model of biasy25. From the figures we can find strong interaction between method and missingmech. Also there exists interaction between method and samplesize. We should include the two 2-way interaction terms in our regression models. Interaction plots for other response variables display similar patterns. Table 6.34, 6.35, 6.36 and Table 6.37 show the estimates of coefficients and corresponding p-values for the regression models. For simplicity, I don't list the insignificant interactions in

6.30: Results of the regression model for bias_y50 (parametric family)

parameter	estimate	p-value
Intercept	0.0336	0.0048
aval-case	0.0850	$\underline{< 0.0001}^*$
rounding	0.0863	$\underline{< 0.0001}$
para-family	-0.0013	0.9336
samplesize	-0.0147	0.0544
missingmech	-0.0038	0.8027
covmatrix	-0.0152	0.0835
aval-case*missingmech	0.0475	$\underline{< 0.0287}$
rounding*missingmech	-0.0088	0.6804
para-family*missingmech	0.0025	0.9062

* the highlighted and underscore type signifies a variable that is significant at $\alpha = 0.05$

6.31: Results of the regression model for cp_y25 (parametric family)

parameter	estimate	p-value
Intercept	0.0001	0.9276
aval-case	0.0775	$\underline{< 0.0001}^*$
rounding	0.0288	0.1446
para-family	0.0001	0.9896
samplesize	0.0014	0.1625
missingmech	-0.0003	0.8476
covmatrix	0.0018	0.1159
aval-case*missingmech	0.1825	$\underline{< 0.0001}$
rounding*missingmech	-0.0008	0.7858
para-family*missingmech	0.01	0.7172

* the highlighted and underscore type signifies a variable that is significant at $\alpha = 0.05$

these tables. The tables show that the available-case method tend to have large biases and bad 95% coverage rates for the means of both continuous variable y_{25} and binary variable y_{50}. And all tables show that the available-case method has significant interaction with missing data mechanisms. The rounding method works well for the inference on mean of y_{25}, but the rounding method introduces more bias and worse 95% coverage rate on the mean of the binary variable y_{50}. Moreover, the factor model approach is as good as all-case method for the means of both continuous variable y_{25} and binary variable y_{50}. Both overparameterized and underparameterized models create small bias and good 95% coverage rate. We also find

6.32: Results of the regression model for cp_y50 (parametric family)

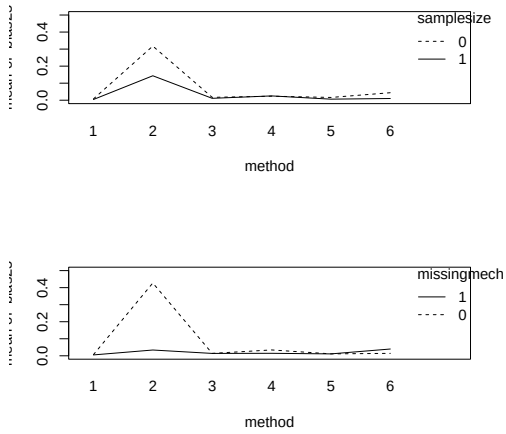
parameter	estimate	p-value
Intercept	0.0158	0.2240
aval-case	0.0937	$\underline{< 0.0001}^*$
rounding	0.0750	$\underline{< 0.0001}$
para-family	0.0013	0.9410
samplesize	-0.0019	0.8250
missingmech	-0.0025	0.8830
covmatrix	-0.0096	0.3290
aval-case*missingmech	0.1250	$\underline{< 0.0001}$
rounding*missingmech	0.0263	0.2760
para-family*missingmech	0.0013	0.9580

that the sample size and the missing data mechanism are not significant. However, further analysis may be necessary to investigate how underparameterized model can affect bias and 95% coverage probability.

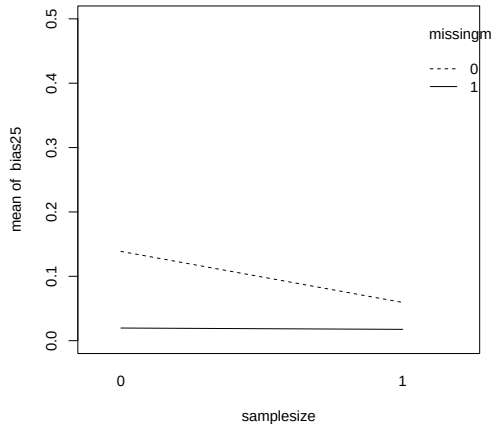
6.33: Regression covariates for factor model approach

Regression covariates	Description
aval-case	1 =available-case method
	0 =all other methods
rounding	1 =rounding method
	0 =all other methods
correct-factor	1 =factor model with correct factor number
	0 =all other methods
overparameterized	1 =overparameterized factor model
	0 =all other methods
underparameterized	1 =underparameterized factor model
	0 =all other methods
samplesize	1 =sample size 300
	0 =sample size 100
missingmech	1 = missingness mechanism M2
	0 = missingness mechanism M1

6.10: Interaction plot between method and samplesize & Interaction plot between method and missingmech



6.11: Interaction plot between missingmech and samplesize



6.34: Results of the regression model for biasy25 (factor model)

parameter	estimate	p-value
Intercept	0.0083	0.8211
aval-case	0.5060	$< 0.0001^*$
rounding	0.0086	0.8682
correct-factor	0.0329	0.5279
overparameterized	0.0080	0.8998
underparameterized	0.0076	0.9042
samplesize	-0.0032	0.9392
missingmech	-0.0015	0.9712
aval-case*missingmech	-0.3917	< 0.0001
aval-case*samplesize	-0.1710	0.0084
...	...	

6.35: Results of the regression model for biasy50 (factor model)

parameter	estimate	p-value
Intercept	0.02	0.8211
aval-case	0.1662	$< 0.0001^*$
rounding	0.1087	< 0.0001
correct-factor	0.02	0.3113
overparameterized	0.015	0.5322
underparameterized	0.016	0.5319
samplesize	-0.0005	0.7539
missingmech	0.0001	0.9926
aval-case*missingmech	-0.0975	0.0002
...*...		

6.36: Results of the regression model for cpy25 (factor model)

parameter	estimate	p-value
Intercept	0.0088	0.728
aval-case	0.3825	$< 0.0001^*$
rounding	0.0038	0.916
correct-factor	0.0325	0.365
overparameterized	0.0087	0.841
underparameterized	-0.0012	0.977
samplesize	0.0025	0.931
missingmech	-0.0025	0.931
aval-case*missingmech	-0.33	< 0.0001
...*...		

6.37: Results of the regression model for cpy50 (factor model)

parameter	estimate	p-value
Intercept	0.0112	0.4910
aval-case	0.2450	$< 0.0001^*$
rounding	0.1012	0.0002
correct-factor	0.015	0.5159
overparameterized	0.0137	0.626
underparameterized	0.0037	0.894
samplesize	0.0025	0.894
missingmech	0.0025	0.893
aval-case*missingmech	-0.1750	< 0.0001
...*...		

CHAPTER 7

Application

When data include moderate number of observations and large number of mixed variables, complete-case analysis is very inefficient or sometimes impossible even if the missingness mechanism is MCAR. Meanwhile, multiple imputation based on the multivariate normal model may fail without appropriate restrictions and prior information. To handle this difficulty, in Chapter 5 we propose two multiple imputation algorithms. In this chapter, the two algorithms are applied to a real data set.

7.1 CHIS 2009 data

The California Health Interview Survey (CHIS) is a population-based telephone survey of Californias population conducted every other year since 2001. CHIS is the largest health survey conducted in any state and one of the largest health surveys in the nation. CHIS is conducted by the UCLA Center for Health Policy Research (UCLA-CHPR) in collaboration with the California Department of Public Health, the Department of Health Care Serivces, California Department of Mental Health, First 5 California, The California Endowment, the National Cancer Institute, and Kaiser Permanente. CHIS collects information for all age groups on health status, health conditions, health-related behaviors, health insurance coverage, access to health care services, and other health and health related issues. Within each household, separate interviews are conducted with a randomly selected adult (age 18 and over), adolescents (ages 12-17), and parents of children (ages 0 to 11). CHIS 2009 is the fifth CHIS data collection cycle and was conducted between September 2009 and April 2010. CHIS 2009 has 47614 observations and 518 variables based on sample survey questions.

Diabetes is a lifelong (chronic) disease in which there are high levels of sugar in the blood. Diabetes can be caused by too little insulin, resistance to insulin, or both. After many years, diabetes can lead to other serious problems:

- You could have eye problems,including trouble seeing (especially at night) and light sensitivity.
- Your feet and skin can develop sores and infections. After a long time, your foot or leg may need to be removed. Infection can also cause pain and itching in other parts of the body.
- High blood sugar and other problems can lead to kidney damage.

It is well known that diabetes may be correlated with some health behaviors such as fruit taken or vegetable taken. But we are not aware of the exact relationship. Meanwhile, the health condition related to diabetes among some specific ethnic groups is also interesting. The main goal of our research is to investigate the relationship between diabetes and 18 predictors among Filipinos in Southern California. Since the CHIS 2009 data include hundreds of variables and different ethnic groups, we only use a subset of the data. The 18 predictors are made up of demographic variables and variables related to health behaviors. Part of them are continuous variables while part of them are binary variables. Including the response variable, most of 19 variables have missing items due to the non-response of the corresponding survey questions. Among 47614 observations, 430 of them are Filipinos. So the sample size we use is 430. Table 7.1 gives the brief description of the data set we use. "mis-rate" denotes the missing rate of each variable. Only 283 out of 430 observations are fully observed. Since some of the variables such as "fruit" and "fry" are not really continuous and normally distributed, we need to do the log transformation to make them accommodate the imputation models.

7.1: Data set description

variables	description	mis-rate
diabetes	whether this observation has diabetes	0%
age	age (in years)	0%
gender	gender of the observation 1 if female, 0 if male	0%
weight	body weight (in pounds)	0%
employed	working status of the observation 1 if employed, 0 if unemployed	0%
bsorabove	educational attainment 1 if bachelor degree or above, 0 elsewhere	0%
walk	whether walked at least 10 min in the past 7 days 1 if yes, 0 if no	3%
fruit	# of times ate fruit in the past month	5%
fry	# of times ate French fries in the past month	2%
vegetable	# of times ate vegetable in the past month	2%
soda	# of times drank soda in the past month	3%
juice	# of times drank fruit juice in the past month	2%
cakeorcookie	# of times ate cake or cookies in the past month	1%
icecream	# of times ate ice cream in the past month	1%
coffeeortea	# of times drank sweetened coffee or tea in the past month	5%
energy	# of times drank energy drink in the past month	1%
alcohol	whether this observation had alcohol in the past 12 month 1 if yes, 0 if no	8%
smoke	whether this observation smoked 100 or more cigarettes in his life 1 if yes, 0 if no	1%
sunburn	# of times had sunburned in the past 12 months	3%

7.2 Analysis of the data set

This data set example highlights several advantages of our parametric family modeling strategy. Following the definition of Section 5.3, the unrestricted variance-covariance matrix Λ contains $19 * (19 + 1)/2 = 190$ parameters which is a relatively large number to estimate accurately with 283 complete realizations of y. On the other hand, the structured variance-covariance matrix $\Omega(\theta)$ has only a few parameters and the data are sufficient to estimate them, but the parametric model might not be appropriate. A compromise is arrived at in

which the unrestricted matrix is pulled towards the parametric structure, yet the data are free to help determine important characteristics such as heteroscedasticity.

To apply the parametric family model to the data set, we choose the initial values of the parameters to be their maximum likelihood estimators obtained by the EM algorithm. This will accelerate the convergence. We make $\Omega(\theta)$ to be a 19×19 compound symmetry matrix. For the data set, since the sample size is relatively larger compared with the number of the variables, and the missing rates for all the variables are lower than 10%, we can apply less informative priors to the parameters. The following priors are used for the imputation:

$$\beta_0 \sim N(0, 1000) \tag{7.1}$$

$$p(\sigma^2) \sim Inv - gamma(2.1, 5) \tag{7.2}$$

$$p(\rho) \propto 1 \tag{7.3}$$

$$v \sim I_{(v>19)}\Gamma(6, 100) \tag{7.4}$$

$$p(\beta_0, \sigma^2, \rho, v) \propto p(\beta_0)p(\sigma^2, \rho)p(v) \tag{7.5}$$

We generate 31000 iterations from the MCMC algorithm with the first 1000 iterations treated as a burn-in period. Time series plots can be used here to check the convergence of the algorithm. The 10 imputations are taken from every 50th iterations since the 30550th iteration due to the auto-correlation plots. After the Multiple Imputation step, a logistic regression model will be applied to each imputed data set. Then we use Rubin's rule to combine the results of different imputed data sets. The combined results are listed in the following table:

To apply factor model to the data, it is very important to find an appropriate number of factors. Checking the eigenvalues of the estimated covariance matrix may not work since some of the variables are not continuous but binary. However, simulations in Section 6.2 show that overparameterization of factor model still gives small bias and good 95% coverage rate. So here I use a 18-factor model. For simplicity, we use the prior distributions defined in Section 5.4. 31000 iterations are generated and the first 1000 iterations are treated as a burn-in period. The 10 imputations are taken from every 20th iterations since the 30820th

7.2: Results of the logistic regression

parameter	estimate	p-value	parameter	estimate	p-value
Intercept	−8.4908	$\underline{< 0.0001}^*$	soda	−0.0343	0.0799
age	0.0530	$\underline{0.0003}$	energy	−0.0482	0.3188
gender	0.2840	0.5147	juice	0.0038	0.8132
weight	0.0289	$\underline{< 0.0001}$	coffeandtea	−0.0248	$\underline{0.0139}$
employed	0.3170	0.3914	cakeorcookie	0.0088	0.6607
bsorabove	0.1884	0.5983	icecream	−0.0717	0.1193
walk	−0.1399	0.6492	sunburn	−0.7277	0.0632
fruit	−0.0041	0.4411	smoke	0.8856	$\underline{0.0236}$
fry	−0.0144	0.4276	alcohol	−0.8014	$\underline{0.0304}$
vegetable	0.0073	0.3050			

* the highlighted and underscore type signifies a variable that is significant at $\alpha = 0.05$

iteration due to the auto-correlation plots. After generating multiple complete data sets, Rubin's rule is used to combine the logistic regression estimates. Table shows the combined results. The two approaches show similar results. From Table 7.2 and Table 7.3, we find

7.3: Results of the logistic regression

parameter	estimate	p-value	parameter	estimate	p-value
Intercept	−7.6277	$\underline{< 0.0001}^*$	soda	−0.0343	0.0790
age	0.0505	$\underline{0.0004}$	energy	−0.0508	0.2399
gender	0.2877	0.5030	juice	0.0051	0.6999
weight	0.0269	$\underline{< 0.0001}$	coffeandtea	−0.0256	$\underline{0.0108}$
employed	0.2982	0.5819	cakeorcookie	0.0072	0.8865
bsorabove	0.1870	0.8030	icecream	−0.0720	0.1060
walk	−0.0942	0.7508	sunburn	−0.7303	0.0667
fruit	−0.0062	0.2473	smoke	0.8326	$\underline{0.0307}$
fry	−0.0136	0.4639	alcohol	−0.7913	$\underline{0.0270}$
vegetable	0.0074	0.2965			

* the highlighted and underscore type signifies a variable that is significant at $\alpha = 0.05$

that older Filipinos are more likely to have diabetes than younger people. The risk of getting diabetes is higher among heavier Filipinos. Drinking coffee or tea can help Filipinos reduce the risk of getting diabetes. Moreover,smoking has a significant effect on increasing the likelihood of diabetes. All above conclusions are in accord with our common sense. It is interesting that the logistic regression outputs indicate that drinking alcohol will be beneficial

to reducing the risk of diabetes. One possible reason is there may exist quadratic effect of alcohol use. Another reason may be we should categorize alcohol use to be moderate use and heavy use. Thus this point is worth further research.

CHAPTER 8

Discussion and Future Research

In the analysis of incomplete data with large number of variables, the modest number of cases and mixed variable types, the complete-case analysis is inefficient and may result in biased estimates. Since we have large number of variables in hand, it may be reasonable to view the missing data mechanism for the data as MAR and to use the multiple imputation technique to obtain estimates that make use of all observed data. We introduce the latent variables for binary, ordinal or nominal variables so we can use a multivariate normal joint modeling to multiple impute the missing data. However, it is very common that some data sets include count variables or semi-continuous variables. To incorporate these variables in a joint modeling is challenging. Dunson (2005) proposed a latent variable model for mixed count, binary and ordinal data by using Poisson underlying latent variables. We may be able to tailor this Poisson latent variable model to handle the mixed continuous, count and categorical variables.

From Section 6.2 we know underparameterization of factor model can result in biased estimates, it seems better to choose enough number of factors to assure inclusion of all important variations. On the other hand, it is generally desirable to have a parsimonious model so that fewer parameters need to be estimated. Since the application of the factor model depend upon the number of factors in use, it would be of interest to develop an adaptive procedure to find an appropriate number of factors.

Choosing the appropriate number of factors is always a subjective matter, not to mention there are missing items and mixed variable types in the data. Since the number of variables is large and the number of observations is moderate, a large-sample test statistic for choosing the number of factors may not be appropriate. A common way to choose the number of

factors is using the scree plot. However, when there are many variables, it is sometimes hard to find a suitable choice from the scree plot. Moreover, it has been known the criterion to choose the number of factors as the number of eigenvalues equals to or greater than one sometimes can lead to the overestimation of the number of factors when there are large number of variables in the model. Song and Belin (2008) developed a new method of choosing the number of factors. First they apply EM algorithm to estimate the parameters in the factor model. Then they use AIC or BIC to choose the appropriate number of factors. But their approach need to be extended to handle the mixed variables scenario. A reversible-jump MCMC algorithm was proposed by Lopes and West (2004) to find the correct number of factors. It is possible to modify their algorithm (e.g., adding one step of missing data imputation) to accommodate the mixed incomplete data situation.

It would also be interesting to extend our parametric family approach in longitudinal data setting. Then we can have more candidates to use as the centered parametric family. Besides compound symmetry (CS), AR(1) can be another option. However, a lot of work need to be done if the longitudinal data is unbalanced.

It is natural to treat the design of simulation set ups as an example of experimental design. So a lot of experimental design techniques can be applied to our simulation procedure to modify the simulation. We have 16 simulation combinations for parametric family method and 16 combinations for factor model as well. The entire computation is burdensome. In Section 6.1.3, we find most of the two-way interaction terms of the simulation factors are not significant. We can then use orthogonal arrays design to reduce the number of simulation combinations. Other than this, the response surface modeling strategy can be used to find the best simulation set ups that can achieve minimal bias and best 95% coverage rate.

The simulation results in Section 6.1 and Section 6.2 indicate that our two approaches work well assuming the missing data mechanism is missing at random (MAR). However, although MAR is a helpful working assumption, sometimes data are not missing at random (NMAR) in practice. When data are NMAR, that is, the probability that a value is missing does depend on unobserved information, the model for generating imputations must not only depend on unobserved data but it must also take into account the process that gave rise to

the missing data. Carpenter, Kenward and White (2007) develop a reweighting approach to investigate the influence of departures from the MAR assumption on parameter estimates. Siddique , Harel and Crespi (2012) proposed an approach using several imputation models instead of unique model to reflect the imputation model uncertainty. Then they transform imputed ignorable variables to create nonignorable values. We may borrow the ideas and modify our models to make them accommodate the NMAR mechanism.

References

[1] Bartholomew, D. J. (1987). *Latent Variable Models and Factor analysis.* New York: Oxford University Press.

[2] Bernaards, C. A., Belin, T. R and Schafer, J. L (2007). *Robustness of a multivariate normal approximation for imputation of incomplete binary data.* Statistics in Medcine 26. 1368-1382.

[3] Boscardin, W. J. and Weiss, R (2001).*Models for the covariance matrix of multivariate longitudinal and repeated measures data.* Proceedings of American Statistical Association, Section on Bayesian Statistical Science.

[4] Boscardin, W. J.,Zhang, X., Belin, T. R (2008). *Modeling a mixture of ordinal and continuous repeated measures.* Journal of Statistical Computation and Simulation Vol.78, 873-886.

[5] Carpenter, J., Kenward, M. and White, I. (2007). *Sensitivity analysis after multiple imputation under missing at random: a weighting approach.* Statistical Methods in Medical Research 2007, 16, 259-275.

[6] Chib, S. and Greenberg, E. (1998). *Analysis of multivariate probit models.* Biometrika, 85, 347-361.

[7] Dunson, D. (2005). *Bayesian latent variable models for mixed discrete outcomes.* Biostatistics, 6, 1, 11-25.

[8] Galton, F. (1888). *Co-relations and their measurement, chiefly from anthropometric data.* Proceedings of the Royal Society, 45, 135-140.

[9] Gelfand, A. E. and Smith, A. F. M. (1990). *Sampling-based approaches to calculating marginal densities.* Journal of American Statistical Association, 85, 398-409.

[10] Gelman, A. and Rubin, D. B. (1992). *Inference from iterative simulation using multiple sequences.* Statistical Science, 7, 457-511.

[11] Geman, D. and Geman, S. (1984). *Stochastic relaxation, Gibbs distributions, and the Bayesian reconstruction of images.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 6, 721-741.

[12] Horel, R. W. and Kennard, R. W. (1970). *Ridge regression: biased estimationfor nonorthogonal problems.* Technometrics, 12, 55-67.

[13] Jamshidian, M. (1997) *An EM algorithm for ML factor analysis with missing data, In Berkane, M, (ED.).* Latent Variable Modeling and Applications to Causality, New York: Springer 247-258.

[14] Little, R. J. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data, 2nd edition, New York: John Wiley series*

[15] Little, R. J. and Schluchter, M. D. (1985). *Maximum likelihood estimation for mixed continuous and categorical data with missing values.* Biometrika 72, 497-512.

[16] Lopes, H. and West, M. (2004). *Bayesian model assessment in factor analysis.* Statistica Sinica, 14, 41-67

[17] Martin, J. K. and McDonald, R. P. (1975). *Bayesian estimation in unrestricted factor analysis: a treatment for Heywood cases.* Psychometrika, 40, 505-517.

[18] Olkin, I. and Tate, R.F. (1961). *Multivariate correlation models with mixed discrete and continuous variables.* Ann. Math. Statist. 32, 448-465.

[19] Quinn, M. K. (2004). *Bayesian factor analysis for mixed ordinal and continuous responses.* Political Analysis 12, 338-353.

[20] Raghunathan et al. (2001). *A multivariate technique for multiply imputing missing values using a sequence of regression models.* Survey Methodology Vol.27, 85-95.

[21] Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods.* 2nd edition Springer.

[22] Rubin, D. B. (1976). *Inference and missing data.* Biometrica 63, 581-592.

[23] Rubin, D. B. and Thayer, D. T. (1982). *EM algorithm for ML factor analysis.* Psychometrika, 47, 69-76.

[24] Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys.* New York: John Wiley.

[25] Rubin, D. B. (1996). *Multiple imputation after 18+ years.* Journal of American Statistical Association 91, 473-489.

[26] Schafer, J. (1997). *Analysis of Incomplete Multivariate Data. London: Chapman & Hall*

[27] Schenker, N. and Taylor, J. M. (1996). *Partially parametric techniques for multiple imputation.* Computational Statistics and Data Analysis 22 (4), 425-446.

[28] Siddique, J. and Belin, T. R.(2008). *Multiple imputation using an iterative hot-deck with distance-based donor selection.* Statistics in Medicine 27 (1), 83-102.

[29] Siddique, J. and Harel, O. (2009). *A SAS macro for Multiple Imputation using distance-Aided selection of donors.* Journal of Statistical Software 2009 Feb; 29(9).

[30] Siddique, J., Harel, O. and Crespi, K. (2012). *Generating multiple imputations from multiple models to incorporate model uncertainty in nonignorable missing data problems.* Unpublished technical report

[31] Song, J. and Belin, T. R. (2004). *Imputation for incomplete high-dimensional multivariate normal data using a common factor model.* Statistics in Medicine 23, 2827-2843.

[32] Song, J. and Belin, T. R. (2008). *Choosing an appropriate number of factors in factor analysis with incomplete data.* Computational Statistics and Data Analysis 52, 3560-3569.

[33] Spearman, C. (1904). *General intelligence objectively determined and measured.* American Journal of Psychology, 15,201-293.

[34] Tanner, M. A. and Wong, W. H. (1987). *The calculation of posterior distributions by data augmentation.* Journal of American Statistical Association 82, 528-550.

[35] Van Buuren, S., Boshuizen, H. C. and Knook, D. L. (1999) *Multiple imputation of missing blood pressure covariates in survival analysis.* Statistics in Medcine, 18, 681-694.

[36] Zhang, X., Boscardin, W. J., Belin, T. R. (2006). *Sampling correlation matrices in Bayesian models with correlated latent variables.* Journal of Computational Graphics and Statistics 15, 880-896.

[37] Zhang, X., Boscardin, W. J. and Belin, T. R. (2008). *Bayesian analysis of multivariate nominal measures using multivariate multinomial probit models.* Computational Statistics and Data Analysis 52, 3697-3708.