The Pennsylvania State University

The Graduate School

The Eberly College of Science

MULTIPLE IMPUTATION FOR MISSING ITEMS IN MULTI-THEMED QUESTIONNAIRES

A Dissertation in

Statistics

by

Rong Liu

Submitted in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

May 2010

The dissertation of Rong Liu was reviewed and approved* by the following:

Joseph L. Schafer
Associate Professor of Statistics
Dissertation Advisor, Chair of Committee

Murali Haran
Assistant Professor of Statistics

Runze Li
Professor of Statistics

Wayne D. Osgood
Professor of Crime, Law, and Justice

Bruce G. Lindsay
Professor of Statistics
Head of the Department of Statistics

*Signatures are on file in the Graduate School.

Abstract

Questionnaires used in survey-based research are often arranged in multiple sections. Each section contains items that are closely interrelated, serving one or more themes. Even with a modest number of sections, the resulting dataset may have a large number of variables, which poses special analytic challenges for dealing with missing values. Current procedures for multiple imputation may fail because the underlying models do not take into account the thematic nature of the questionnaire and are over-parameterized. Attempts to simplify the model—for example, by assuming that the items within a theme are conditionally independent given a small number of latent factors—may fail to capture special features of the data if the specified model does not fit. In this dissertation, I develop new multiple-imputation procedures for multi-themed questionnaire data based on a flexible class of confirmatory factor models. I present PX-EM algorithms for maximum-likelihood estimation in exploratory and confirmatory factor analysis with incomplete data. The factor model is then relaxed by adding an additional random component which allows the covariance structure to deviate from the assumed model. I present an MCMC algorithm for generating Bayesian multiple imputations under this extended model. These techniques are illustrated using data on emotional distress from a large adolescent health survey.

Table of Contents

List of Figures

List of Tables

Acknowledgments

First and foremost, I would like to sincerely thank my advisor, Dr. Joseph Schafer, one of the nicest professionals that I have ever met. I am deeply grateful for his guidance and support throughout my graduate career and during the completion of this dissertation. When I just started working on missing data, Dr. Schafer once spent a whole afternoon showing me how to efficiently check imputation algorithms. When I was working on the dissertation, he put tremendous efforts into the revision of my drafts. When I was looking for a job, he was kind enough to refer me to available resources and even helped me revise my interview presentation slides. When I started my job at Indianapolis and had to work 9 to 5, he advised me through Skype regularly at 6 AM. I can never thank him enough for what he has done for me. He has demonstrated that not only is he an expert in his research areas, but also a great advisor who is always considerate of his students. There's an old Chinese saying that literally translates into "being strict with oneself and lenient to others". Dr. Schafer embodies the saying perfectly.

I would like to thank Dr. Murali Haran, Dr. Runze Li, and Dr. Wayne Osgood for taking time from their hectic schedules to serve on my committee.

Thanks to all my professors in the Department of Statistics at the Pennsylvania State University. I would thank Dr. Donald Richards for recruiting me as a Ph.D. student to this marvelous department. Special thanks to Dr. Naomi Altman, Dr. Francesca Chiaromonte and Dr. K.B. Boomer (now at Bucknell university), for their constant support and encouragement during my first two years of study. I thank Dr. Bing Li for making his theoretical courses so absorbing to us students. I also want to thank Dr. Steve Arnold whose wisdom has greatly influenced me when I took his course.

Last but not least, I'd like to thank my family. My husband Wei Zhang never forgets to use his own brand of humor to encourage me. My daughter Jasmine does

not fail to use her sweet "Go, Mommy, go!" to cheer me up. I am infinitely indebted to my parents, Jidong Liu and Zhaoyu Sun, whose love has been so consistently observed with no "missing" values at any time point since the day I was born.

Chapter 1

Introduction

1.1 Missing data in multi-themed questionnaires

Questionnaires are used in business, economics, psychology, public opinion research and many other areas to collect data from human respondents. Until recently, the term *questionnaire* almost invariably referred to a paper form. But many surveys are now administered by telephone, computer assisted personal interview devices or web-based electronic forms (Groves et al., 2004), and my use of this term is intended to cover all of these data-collection modes.

Many survey questionnaire items have a limited number of possible responses, and the resulting analytic variables are typically binary or ordinal. But occasionally the measures are nominal, as in categories of race or ethnicity, or continuous, as in body weight. And it is not uncommon to find measures that are a mixture of discrete values and continuously distributed responses. An example of the latter is dollars spent by members of a household in a given year on a category of consumer goods (e.g., refrigerators) that are not frequently purchased.

Whether a survey question is intended to produce a categorical or numeric mea-

sure, the resulting data often contain codes for non-responses that are treated as missing values. Sometimes these codes are for missing values, such as "don't know" or "refused." Sometimes these missing values arise by design, when questions are omitted for random subsets of the sample to keep the questionnaire short. Datasets may also include codes for items that are legitimately skipped because they are not applicable to certain participants. For example, a respondent who indicates that he or she has never used marijuana may be instructed to skip items pertaining to frequency and amount of marijuana use. In those cases, the "missing" values are not really missing, because a negative response to the initial question logically implies values of zero for all frequencies and amounts.

Survey methodologists use the term *item nonresponse* to refer to genuinely missing values that occur on individual items during the process of data collection and capture. This is distinguished from *unit nonresponse* which results from failure of the whole interview process, e.g. when a sampled individual fails to show up or refuses to participate. Item nonresponse and unit nonresponse are handled in different ways. Item nonresponse is often addressed by imputation, whereas unit nonresponse is typically handled by weighting adjustments (Little & Rubin, 2002).

This dissertation is concerned with statistical methods for item nonresponse in multi-theme questionnaires. In a multi-themed questionnaire, the items may be grouped according to the subject matter areas being addressed. Items within a theme are often analyzed together, and responses to multiple items are often aggregated into summary measures or scores. For example, one theme of a health survey may be nutrition. Participants may be asked about the frequency of consumption of different types of foods. The questions may include, "How many glasses of milk do you drink each day?" Followup questions may be asked about whether the milk is whole, reduced-fat or skim, and additional questions may be asked about

other types of dairy products (cheese, ice cream). Depending on the purpose of the analysis, responses to these items may be combined in different ways to produce measures of fat intake, calcium intake, and so on. Milk consumption may also be compared or contrasted with items with other types of liquids (water, 100% juice, soft drinks, etc.) to produce variables that may appear as responses or predictors in regression models or other types of analyses.

In the previous example, questionnaires items pertaining to nutrition may be grouped together in the data-collection process, so that these items may appear in a nutrition section that is distinct from sections containing non-nutritional items. Some multi-themed surveys do have sections that are thematically distinct. In other surveys, however, a lengthy questionnaire may be divided into sections that are not thematic but simply patition the workload of data collection into subunits. If that is the case, then a data analyst may need to cull the items pretaining to a theme from multiple sections of the questionnaire. The themes described in this dissertation are groupings by subject matter, not necessarily by physical position in the questionnaire or by temporal ordering within the interview.

Analyses of data from multi-theme questionnaires are typically multivariate, and the items used in a given analysis may be drawn from a single theme or from multiple themes. As the number of items appearing in the analysis increases, the nuisance created by missing values can accumulate very rapidly, even when the percentage of missing values on any single item is small. For illustration, consider a scenario in which each variable in a set of variables is missing with a fixed probability independently of the other variables. The percentages of complete cases (the participants who have observed values for every variable in the set) for various missingness rates and numbers of items are shown in Table 1.1. In an analysis of 100 variables, a missingness rate of 1% per item will result in complete

	% missing per variable		
# of variables	1%	3%	5%
10	90	74	60
25	78	47	28
50	61	22	8
100	37	5	1

Table 1.1. Percentage of complete cases

data for only 37% of the cases.

The assumption of independent missingness may be unrealistic, because missing items tend to occur together; individuals' propensities to respond often vary. But anecdotal evidence suggests that this does sometimes happen. (One member of my dissertation committee, Dr. Osgood, noted that he has encountered near-independent missingness on items from one large survey, *Monitoring the Future*, which has been extensively analyzed by him and his colleagues.) This artificial example well illustrates the drawbacks of one simple but widespread statistical method for handling missing values: listwise deletion, also known as complete-case analysis (Little & Rubin, 2002; Schafer & Graham, 2002). Listwise deletion causes large proportions of cases to be discarded, making the resulting estimates inefficient. Case-deletion procedures may also introduce bias if the cases that remain are not representative of the population. Based on a meta-analytic study of published results from regression analyses in political science, King et. al. (2001) concluded that listwise deletion of incomplete cases often produced results that were worse than if the incomplete *variables* had been removed from the regression model.

1.2 Multiple imputation: a general framework

1.2.1 Statistical modeling with missing values

Although missing values are ubiquitous, statistical methods and software for data analysis are often not designed to handle them. Data analysts are tempted to edit their datasets to make them appear complete, either by removing incomplete cases or by performing simple imputation procedures, replacing the missing values with means or other values obtained in an ad hoc fashion (Little & Rubin, 2002). The shortcomings of these widely used missing-data adjustments have been well documented (Schafer & Graham, 2002). In more specialized approaches (e.g., censoring or truncation models), the nonresponse is stochastically modeled as part of the data-generating process. Although many examples of these models have appeared in recent years, analysts without special expertise still tend to avoid them because the models are finely tuned to specific applications, and software for fitting these models may be unfamiliar or unavailable (King et al., 2001).

Increasingly, software for multivariate statistical modeling is being extended to accommodate missing values without modeling the processes that lead to nonresponse. For nearly a decade, the program Mplus (Muthén & Muthén, 1998–2007) has been able to fit many types of continuous and discrete-data models with incomplete data under an assumption that the missing values are missing at random (MAR), to be defined later. Multivariate modeling procedures that accommodate missing values are also found in Amos (Arbuckle, 2006), PROC LCA/LTA (Lanza et al., 2008). In these procedures, inferences about model parameters are based on a likelihood function which is maximized by an EM algorithm or some other method, or a posterior distribution, which is usually simulated using Markov chain Monte Carlo (MCMC). Thus the missing values are removed from likelihood

function or posterior distribution by marginalizing the joint distribution of the multivariate data over the unseen values (Schafer & Olsen, 1998). An advantage of these procedures is that is highly efficient, making use of all the observed data, and essentially unbiased if the modeling assumptions are true. A limitation is that these methods provide inferences about the parameters for only one model. Researchers who wish to perform a variety of analyses, or who wish to venture outside of the family of models provided by the software, are likely to pursue other missing-data options.

1.2.2 Imputation

Another approach is to fill in the missing values in a reasonable fashion, taking care to use a method that leads to reliable estimates and measures of uncertainty when the completed data are subsequently analyzed. Imputation is a general term for any method that fills in or replaces missing items. Imputation is not an end in itself; rather, it is a preliminary or intermediate step employed to make a later analysis easier. Imputation allows an analyst to explore the data in a straightforward way, to use statistical methods and software that were designed for complete data, and to focus attention on the scientific questions of substantive interest rather than on the missing-data aspect which is usually a nuisance.

Some of the earliest imputation methods applied to questionnaire data were whole-case substitution, cold-deck and hot-deck imputation (Little & Rubin, 1987). Case substitution and cold-deck imputation replace the missing values with historical data obtained from donor cases from a previous survey or census. Hot-deck imputation proceeds in a similar fashion but uses donor cases from the same dataset. The donors may be complete cases chosen by a matching process that

requires agreement on a set of variables found in the incomplete record. Variants of hot-deck imputation are still being used today by data collectors in government and the private sector. The rules for choosing donors, which may be quite elaborate, are designed to preserve inter-variable relationships. A much simpler technique used by many analysts is mean substitution, where the missing values are replaced by the average of the observed values on a variable-by-variable basis. Mean substitution is sometimes benign, especially if the percentage of missing values for the item is very low. As rates of missingness increase, however, it adversely affects quantities related to variability (variances) and relationships among variables (correlations), distorting inferences in a variety of ways (Schafer & Graham, 2002).

Researchers in diverse fields have become increasingly aware that, if imputation is to be used, it should be done carefully to preserve the integrity of post-imputation analyses. Many now realize that, especially for multivariate analyses, relationships among variables need to be maintained, and the natural variability among observations should also be maintained. For example, Roth and Switzer (1999) replaced missing values by predictions from a regression equation estimated from cases with complete data. Variants of this approach—formerly known as Buck's method—have been used for nearly a century (Little & Rubin, 1987). It is not difficult to see that this method will tend to overstate correlations. The method can be greatly improved by one simple step: adding random residuals to the regression predictions with variance estimated under the model. For more discussion of these imputation methods, refer to Rubin (1987a), Rubin (1987b) and Harel and Zhou (2007).

Regression imputation methods are straightforward to implement when missing values occur on only one variable. They may also be applied in special cases where

the missing values fall into a special (e.g., monotone) pattern (Little & Rubin, 2002). For multivariate data with arbitrary patterns of missingness, principled techniques for imputation become more computationally elaborate (Schafer, 1997).

1.2.3 Multiple imputation

After missing values have been filled in, the procedures and software applied to the completed data are typically "unaware" that some of the data were imputed. That is, the imputed values are treated as if they had actually been observed. If an inferential procedure does not distinguish observed data from imputed data, the procedure will tend to understate the true levels of uncertainty, because imputed values are less reliable than those that were actually seen.

Rubin (1987a; 1996) addressed this issue by the method of multiple imputation (MI). In MI, each missing value is replaced by a modest number of simulated values, producing multiple versions of the completed data. The variability of the results across these multiple versions produces a between-imputation variance component that is necessary to compute intervals and tests with desirable repeated-sampling properties.

The general framework of MI, and actual implementations of MI, are often motivated by Bayesian arguments. The imputations are simulated repeated draws from the posterior joint predictive distribution of the missing values given the observed values. The multivariate model for the complete data that generates this predictive distribution is often called the *imputation model*. A distinctive feature of MI or any imputation procedure is that the imputation model may or may not be compatible with the *analysis model* that is applied later. Implications of discrepancies between these models were investigated by Meng (1994) and by

Collins, Schafer and Kam (2001).

In this dissertation, I derive and implement new methods for multiple imputation for missing items in multi-themed questionnaires. These methods are based on refinements of the multivariate normal model previously explored by Schafer (1997) and others.

1.3 Example: Add Health

One illustrative example of a multi-themed questionnaire comes from the National Longitudinal Study of Adolescent Health (Add Health) (Udry et. al. 2003). Add Health began with a nationally representative sample of students enrolled in grades 7-12 in the United States during the 1994-95 school year. The students were interviewed in that first year (Wave I) and on three subsequent occasions, most recently in 2008 (Wave IV). Add Health researchers also conducted interviews with siblings, friends, romantic partners, parents and school administrators. For purposes of illustration, we will work with a set of variables from the student interview questionnaire at Wave II (1995–96). The Wave II in-home student interview questionnaire had 39 sections, and we will focus our attention on two of these: *Feelings Scale* (Section 10) and *Tobacco, Alcohol and Drugs* (Section 27).

The *Feelings Scale* section contains 19 items designed to measure the participants' levels of emotional well being or psychological distress. Students were asked, "How often was each of the following things true during the past seven days?" and were then presented with 19 statements about their emotional states. For each item, the possible responses fell on a four-point integer scale ranging from 0 (never or rarely) to 3 (most of the time or all of the time). Tabulated responses for two of these items from the Add Health codebook are shown in Figure 1.1. For 15 of

10. You felt fearful.			H2FS10	num 1
10677	0	never or rarely		
3544	1	sometimes		
381	2	a lot of the time		
119	3	most of the time or all of the time		
1	6	refused		
16	8	don't know		
11. You were happy.			H2FS11	num 1
424	0	never or rarely		
2876	1	sometimes		
6216	2	a lot of the time		
5211	3	most of the time or all of the time		
1	6	refused		
10	8	don't know		

Figure 1.1. Two Items in Feelings Scale (2 of 19 items) (Udry *et al.* 2003)

the 19 items, a higher numerical value is associated with increased distress; for the remaining 4 items, greater distress is indicated by a lower value. A standard practice for analyzing these data is to invert the responses for the items that were reverse-coded and then sum or average the items into a composite measure of emotional distress (McNeely et al., 2001; Woods, 2006). Reducing the 19 items to one average score is consistent with the notion that all of these items are measuring a single latent factor. In later chapters, we will see that the relationships among the 19 items cannot be fully explained by one latent factor. A single-factor model can be formally rejected, but one factor does account for a large part of the observed relationships, and the use of a single composite score has great theoretical and practical appeal.

As Figure 1.1 shows, most participants responded to these items in the *Feelings Scale*, but a few refused to answer or said, "I don't know." In this particular

example, the rates of missing values are so low that the effects of a poor missing-data procedure (e.g., mean substitution) may be negligible. But this will not always be the case. Higher rates of missing values will prompt a researcher to apply a more sophisticated procedure such as multiple imputation under a multivariate normal model (Schafer, 1997). One question that is often asked by researchers is this: "Given that we are going to average these items anyway, do we need to impute all of these items and then average them, or can we average the items first and just impute the composite score?" If missingness on these items is highly correlated—i.e., if the missing values arise primarily from a single group of individuals who fail to respond to many of the items—then the two approaches will yield similar results. In that case, averaging prior to imputation will reduce the dimension of the imputation model, streamlining the computations and allowing the researchers to bring a greater number variables from other sections of the questionnaire into the model, which is generally desirable. But if missingness on these items is not highly correlated, then a multivariate procedure that jointly imputes the items has greater appeal (Schafer & Graham, 2002). Another procedure that seems common in these settings is to compute the composite score by averaging the observed items for each individual. Limited experience suggests that this method may perform well when the inter-variable relationships are consistent with a single-factor model (Schafer & Graham, 2002).

A more challenging situation arises in Section 27, *Tobacco, Alcohol and Drugs*. This section contains a much larger set of items (68) and measures multiple dimensions of substance-use behavior. Tabulated responses to two items (#19 and #20) pertaining to alcohol use in the previous 12 months are shown in Figure 1.2. Many legitimate skips occur because these items were preceded by another question (#15) that asked, "Since [month of last interview], have you had a drink

of beer, wine, or liquor—not just a sip or a taste of someone elses drink—more than two or three times?" If the response to #15 was negative, the interviewer was instructed to skip ahead to item #41, resulting in values of "legitimate skip" for items #16 through #40. But missing values due to refusal and "don't know" also occur. A total of 18 alcohol-related items appear in this section pertaining to frequency and amount of drinking, types of beverages consumed, and frequency of risky or antisocial drinking-related behaviors.

Given the complexity and variety of these alcohol measures, it is not intuitively obvious that the relationships among them could be approximated by a model as simple as one that could describe the *Feelings Scale*, a model that assumes one or several continuous latent factors. Patterns of substance use have been described by latent-class analyses that treat the population as a discrete mixture with a small number of homogeneous types or classes. This approach was pioneered by Collins and Wugalter (1992) and can now be found in dozens of published articles and book chapters. From that perspective, one could envision an imputation procedure based on a latent-class model.

Interestingly, Loken and Molenaar (2008) have demonstrated connections between multivariate models that assume discrete latent classes and models that assume continuous latent traits. In particular, they show that a model with $K-1$ latent traits may be approximated by a model with K latent classes, at least up to the first and second moments. For a heuristic explanation of this near-equivalence, imagine an r-dimensional random vector $Y = (Y_1, \ldots, Y_r)^T$ of zero-centered variables distributed as a K-component mixture of multivariate normal distributions with mixing probabilities $\pi_1, \pi_2, \ldots, \pi_K$, where the kth component is $N(\mu_k, \Sigma)$, where $\mu_k = (\mu_{1k}, \mu_{2k}, \ldots, \mu_{rk})^T$ and $\Sigma = \text{Diag}(\sigma^2_{1k}, \sigma^2_{2k}, \ldots, \sigma^2_{rk})$. The marginal mean of Y is $E(Y) = \sum_k \pi_k \mu_k = 0$, because the variables have been centered. The

variances and covariances of the variables are given by

$$\begin{aligned} E(Y_i^2) &= \sum_k \left(\pi_k \mu_{ik} + \pi_k \sigma_{ik}^2\right), \\ E(Y_i Y_{i'}) &= \sum_k \pi_k \mu_{ik} \mu_{i'k}. \end{aligned}$$

This implied covariance structure can thus be written in matrix form as

$$\Sigma = \Gamma^T \Omega \Gamma + \Delta,$$

where Γ is a $K \times r$, Ω is $K \times K$, and Δ is diagonal. This resembles a traditional factor analysis with K latent factors, with Γ corresponding to factor loadings and Δ corresponding to uniquenesses. Unlike the traditional factor model, however, the K latent factors are indicator variables from a multinomial experiment with probabilities $\pi_1, \ldots, \pi_K$. Because these indicators are constrained to sum to one, their covariance matrix has rank $K - 1$ rather than K, so in reality this is a $(K-1)$-dimensional factor representation. This near equivalence of discrete and continuous latent-variable models suggests that a factor-analytic approach may be broadly useful for imputation modeling even when intuition suggests that the response patterns may follow classes rather than continua.

1.4 Goal and scope

Multiple imputation of incomplete multivariate data has often been carried out using methods described by Schafer (1997) which assume a multivariate normal population model with an unstructured covariance matrix. Normal models are frequently applied to binary and ordinal items, with rounding or truncation ap-

19. During the past 12 months, on how many days did you drink alcohol?			H2TO19	num 2
145	1	every day or almost every day		
389	2	3 to 5 days a week		
1048	3	1 or 2 days a week		
1135	4	2 or 3 days a month		
1719	5	once a month or less (3-12 times in the past 12 months)		
1996	6	1 or 2 days in the past 12 months		
462	7	never *[skip to Q.41]*		
22	96	refused *[skip to Q.41]*		
7808	97	legitimate skip		
14	98	don't know *[skip to Q.41]*		
20. Think of all the times you have had a drink during the past 12 months. How many drinks did you usually have each time? A "drink" is a glass of wine, a can of beer, a wine cooler, a shot glass of liquor, or a mixed drink.			H2TO20	num 2
6273		range 1 to 95 drinks		
19	96	refused		
8306	97	legitimate skip		
140	98	don't know		

Figure 1.2. Two Items in Alcohol Use (Udry *et al.* 2003)

plied to the imputed values. Procedures associated with this unstructured normal model tend to work well when the number of variables p is not excessively large. Applications with 100 variables or more are fairly common, but $p > 200$ will strain the computational resources that are presently available to most data analysts. Use of this model also generally requires that the number of sampled cases or observational units n be substantially larger than p; if it is not, then the model tends to be overparameterized, leading to computational and inferential problems.

It is my goal in this dissertation to extend the normal-based methods of Schafer (1997) and others to accommodate larger numbers of items and, in doing so, to create imputation procedures suitable for multi-themed questionnaires. We will do

so by imposing structure on the covariance matrix to reduce the overall complexity of the model. That is, we will first suppose that the relationships among items measuring a single theme can be attributed to a small number of continuous latent factors. Recognizing that such a simple model might not hold, we will formally accommodate lack of fit by allowing the actual covariance matrix to depart from the factor structure by random amounts. In effect, we will impose soft constraints upon the normal model that are consistent with the notion that items measuring a single theme are related, but the pattern of relationships may deviate from the a priori assumptions.

With these new imputation procedures, we will be able to handle questionnaire sections like the Add Health *Feelings Scale*, where the items measure only a few dimensions. We are not yet able to impute all of the items in a complicated section like *Tobacco, Alcohol and Drugs* with multiple constructs, complicated skip patterns and nonstandard (e.g., nominal) items. However, the computational procedures developed here represent an important starting point for future extensions to these more complex situations.

The computational methods we describe will include new EM-type algorithms for parameter estimation and new MCMC procedures for multiple imputation. The EM-type algorithms are not always necessary for MI, but having mode-finding procedures will be helpful as we investigate the relationships among items measuring a single theme, which will help us to select appropriate models. Results from EM will also be used to tune the parameters of the MCMC procedures and to create proposal distributions for Metropolis-Hastings algorithms.

1.5 Assumptions

1.5.1 The imputation model

Missing-data methods inevitably make assumptions about the manner in which data values became missing. Responsible use of MI also requires a judicious choice for the imputation model. In this section, we describe our key assumptions and some tolerable violations.

Let Y denote a set of data. In the present context, it will refer to a data matrix with rows corresponding to sampled cases and columns corresponding to variables (items) that are subject to missing values and may require imputation. We will partition Y into its observed part and its missing part, denoted by $Y = (Y_O, Y_M)$. We will suppose that the rows of Y are randomly sampled from a population that follows a parametric model.

In addition to the variables in Y that may require imputation, we allow there to be other variables that are observed for all cases and do not need to be imputed. These variables will be called *covariates*, and the matrix of covariates will be denoted by X. Note that our assignment of variables to Y or X is only for the purpose imputation modeling and is not intended to describe what the analyst may do after the missing values have been imputed. In post-imputation analyses, variables in Y may appear as predictors, and variables in X may appear as outcomes. Our use of Y and X is for merely for convenience and computational efficiency, because variables that are completely observed may be conditioned upon and treated as fixed in a multivariate imputation model (Schafer, 1997). In Add Health, for example, sex and age are missing so rarely that we don't need to impute them, so sex and age would be assigned to X.

The imputation model, which describes the conditional distribution of Y given

X in the population, will be written as $P(Y|X;\theta)$, where θ is a set of unknown parameters. To simplify the notation, the covariates X will sometimes be omitted from our expressions, but conditioning on them will always be assumed. The absence of X from an expression does not mean that we have averaged over it; rather, we will drop X from our formulas when its presence would make expressions tedious.

1.5.2 The mechanism of missingness

Define M as the missing-data indicator set for Y. M is a matrix with the same dimensions as Y, with an element of M equal to 1 if the corresponding element of Y is missing and 0 if the corresponding element of Y is observed. It is customary to treat M as a set of random variables and to factor the joint distribution of Y and M as

$$P(Y, M|X;\theta,\xi) = P(Y|X;\theta)P(M|Y,X;\xi),$$

where $P(Y|X;\theta)$ denotes the imputation model, $P(M|Y,X;\xi)$ describes the mechanism of missingness, and ξ is a set of parameters that governs the mechanism of missingness. Mechanisms of missingness are commonly classified into four types.

- Missing completely at random (MCAR): Missingness does not depend on covariates or outcomes. Probabilities of response are unrelated to observed or unobserved measurements,

$$P(M|Y,X;\xi) = P(M|\xi).$$

- Covariate-dependent (CD) missingness: Missingness may possibly depend on

covariates but not on outcomes. Probabilities of response are unrelated to outcomes but may be related to covariates which are fully observed,

$$P(M|Y, X; \xi) = P(M|X; \xi).$$

Note that MCAR is a special case of CD.

- Missing at random (MAR): Missingness may possibly depend on covariates and observed outcomes but not on missing outcomes,

$$P(M|Y, X; \xi) = P(M|Y_O, X; \xi).$$

Note that CD is a special case of MAR.

- Missing not at random (MNAR): Missingness depends on missing outcomes,

$$P(M|Y, X; \xi) \neq P(M|Y_O, X; \xi).$$

Any violation of MAR is MNAR.

1.5.3 Ignorability

Missing at random is often described along with the concept of *distinctness*. From a frequentist perspective, two sets of parameters are said to be distinct if their joint parameter space is the Cartesian cross-product of the individual parameter spaces. From a Bayesian perspective, distinct means that any joint prior distribution applied to the two parameter sets can be factored into independent marginal priors distributions.

Rubin (1974) showed that under an assumption of MAR, and when θ and ξ are

distinct, the joint probability distribution of the observed data Y_O and the missingness indicators M can be factored into two pieces, one regarding the parameter of interest θ and the other regarding the nuisance parameter ξ,

$$\begin{aligned} P(M, Y_O|\theta, \xi) &= \int P(M, Y|\theta, \xi)\, dY_M \\ &= \int P(M|Y; \xi)\, P(Y; \theta)\, dY_M \\ &= P(M|Y_O; \xi) \int P(Y; \theta)\, dY_M, \\ &= P(M|Y_O; \xi)\, P(Y_O; \theta), \end{aligned}$$

where the integral changes to summation if Y is discrete. The factor pertaining to θ is the relevant part for likelihood-based inferences about θ. We define the *observed-data likelihood function*, which we denote by $L(\theta|Y_O)$, to be

$$L(\theta|Y_O) \;=\; c\, P(Y_O; \theta),$$

where c is an arbitrary constant of proportionality that does not depend on θ. In particular, any part of $P(M|Y_O; \xi)$ that depends on ξ may be incorporated into c with no effect on likelihood-based inferential procedures for θ. Therefore, under the assumptions of MAR and distinctness—a combination that is known as ignorability (Little & Rubin, 2002)—all information about the parameters θ of the imputation model is carried by the observed-data likelihood function defined by the distribution of Y_O (which, in reality, is the conditional distribution of Y_o given X, because in our notation we have suppressed X).

An assumption of ignorability is a crucial part of most parametric missing-data procedures because it allows us to make direct likelihood or Bayesian inference

about θ without assuming anything more about the model that may have produced M. If we want to find the maximum-likelihood (ML) estimate for θ, we need only maximize the function $L(\theta|Y_O)$ with respect to θ. If we want to conduct Bayesian analysis of θ, we need only impose a prior distribution on θ.

Although we will proceed under this assumption of ignorability, we recognize that in real applications where missing data are happening for uncontrolled reasons, MAR is never going to be precisely true. Statisticians tend to justify the assumption of MAR on the following grounds. First, MAR tends to greatly simplify the analysis. Second, in highly multivariate applications where Y and X contain many variables, departures from MAR may not be too serious (Little & Rubin, 1987; Rubin, 1987a; Schafer, 1997). The key assumption in MAR is that relationships between Y_M and M are completely mediated by (X, Y_O). If X contains a rich set of covariates, or if Y_O is highly predictive of Y_M, then we may expect that the residual dependence of M upon Y_M after controlling for Y_O and X will be relatively minor. In a few specialized examples where the missing values later became known, it has been found that an assumption of MAR, even when it is demonstrably false, performed better than alternatives that were designed to for MNAR (David et al., 1986; Rubin, Stern & Vehovar, 1995). In situations where ignorable procedures do not perform well, they still provide important baseline analyses for comparing and assessing MNAR-based alternatives (Schafer, 1997). If we dispense with an assumption of MAR, other unverifiable assumptions must be made, and the results from MNAR analyses are often highly sensitive to departures from these assumptions (Little & Rubin, 2002).

1.5.4 Choosing the imputation model

When applying multiple imputation to multivariate data, the imputation model $P(Y|X, \theta)$ should be chosen with an eye toward future analyses. In particular, the imputation model should capture important features of the data that are relevant for post-imputation analyses. For example, if the data are multilevel and will be subject to multilevel modeling, then the imputation model should also be multilevel (Schafer & Olsen, 1998; Carpenter & Goldstein, 2004). For a dataset that represents a multi-themed questionnaire, the model should have a covariance structure that accurately describes the relationships among items within the theme. If the items were designed to measure a single underlying trait, as in the Add Health *Feelings Scale*, the assumed form of the covariance matrix should allow the items to be closely related and, if possible, pool the information from these relationships to strengthen measurement of that trait. If the imputation model does not preserve marginal or and conditional associations among variables that will be investigated in a subsequent analysis, results from that analysis may be biased.

Researchers have often been advised to build imputation models that are *inclusive* in the sense that a rich set of variables has been incorporated into Y and X (Schafer, 1997; Collins et al., 2001). Certainly, all of the variables that will be used in post-imputation analyses should be present. Extra variables that will not be used in the analysis may also be included, and these have been called *auxiliary variables* (Collins et al., 2001; Allison, 2002). Good candidates for auxiliary variables are those that are thought to be predictive of the variables requiring imputation, and those that are thought to be related to reasons for missingness. Auxiliary variables of the former type may increase the precision of the imputed values, strengthening the analysis through a phenomenon that has been called

superefficiency (Meng, 1994). Auxiliary variables of the latter type will tend to make the ignorability assumption more plausible, reducing biases that arise when missing values are not missing completely at random.

If auxiliary variables are not strongly related to missingness indicators or to variables being imputed, it is not necessary to include them, but often there is little harm in doing so. Simulations by Collins at al. (2001) have shown that adding unnecessary variables to an imputation model does not appreciably impair post-imputation analyses, so researchers have been told to build imputation models that are as large as possible. But this advice has theoretical and practical limits. Beyond a certain point, the computer software and hardware will be unable to perform the required computations within a reasonable amount of time. And as more variables are introduced in a sample of a given size, the imputation model will eventually become overparameterized, causing post-imputation analyses to become unstable. Beyond a certain point, it will no longer be beneficial to use an imputation model that allows an unstructured covariance matrix, and additional information will need to be introduced by imposing constraints on the parameter space, by applying informative prior distributions, or both.

1.5.5 Multivariate normality: limitations and possibilities

Multiple imputation techniques for normal imputation models have gradually made their way into the statistical mainstream, but the methods for categorical and mixed-data models (Schafer, 1997, Chap. 7–9) have not. The latter require formation of a contingency table that cross-classifies the sampled units by all of the categorical variables in the model. Allocation and manipulation of these arrays becomes computationally very demanding as the number of variables increases,

and the methods become impractical when the number of categorical variables exceeds 20 or 25. By comparison, algorithms and software based on the normal model with an unstructured covariance matrix can routinely handle 100 variables or more. For this reason, MI is most often carried out under an assumption of multivariate normality, even when the variables to be imputed are discrete. When the normal model is applied to binary and ordinal variables, the imputed values are typically rounded off to the nearest category. Bernaards, Belin and Schafer (2007) and Demirtas (2008) explored different rounding rules for binary variables and demonstrated that these methods perform reasonably well. Others have argued that leaving the imputed values alone (i.e. not rounding them) is often better than rounding (Allison, 2005; Horton et al., 2003).

More recently, Boscardin et al. (2006, 2008) have developed a unified approach to joint imputation of continuous, ordinal and nominal data under a multivariate probit model. That model describes categorical variables as coarsened versions of latent normal scores, and relationships among variables are characterized by correlations among these scores. For subject i, suppose the data vector Y_i consists of a continuous portion C_i with length r_c, an ordinal portion O_i with length r_o and a nominal portion N_i with length r_n,

$$
\begin{aligned}
{C_i}^T &= (C_{i1}, \ldots, C_{ijr_c}), \\
{O_i}^T &= (O_{i1}, \ldots, O_{ijr_o}), \\
{N_i}^T &= (N_{i1}, \ldots, N_{ijr_n}).
\end{aligned}
$$

A joint distribution for these three vectors is constructed as follows.

- For each ordinal variable O_{ij} $(i = 1, \ldots, n,\ j = 1, \ldots, r_o)$, define a normal

latent variable O_{ij}^L. The element O_{ij} takes values within the discrete set $0, 1, \ldots, J_j - 1$, and $O_{ij} = l$ if and only if O_{ij}^L lies within the interval $(\zeta_{j,l-1}, \zeta_{j,l}]$, where the $\zeta_{j,l}$s' are cut points ($\zeta_{j,0} = -\infty$ and $\zeta_{j,J_j-1} = +\infty$). This is a standard multivariate probit model.

- For each nominal variable N_{ij} $(i = 1, \ldots, n,, \; j = 1, \ldots, r_n)$, which is assumed to have p_j possible outcomes, define a normally distributed $(p_j - 1)$-dimensional latent utility vector N_{ij}^L whose maximum determines the outcome of N_{ij}, i.e.

$$N_{ij} = \begin{cases} 0 & \text{if } \max_{l=1,\ldots,p_j-1} N_{ijl}^L < N_{ijk}^L, \\ k & \text{if } \max_{l=1,\ldots,p_j-1} N_{ijl}^L = N_{ijk}^L. \end{cases}$$

  This is known as a multivariate multinomial probit (MVMNP) model.

The continuous observed variables C_i, and the latent variables O_i^L and N_i^L, are concatenated into a single vector that is modeled as a multivariate normal distribution. For purposes of identifiability, the mean and variance of each latent variable is fixed at zero and one, respectively.

This model developed by Boscardin et al. (2006, 2008) will not be considered further in this dissertation. However, we have included this explanation to show how assumptions of normality are not as limiting as they may first seem, and the methods we develop here can in in the future be extended to handle discrete variables as well.

1.6 EM-type algorithms for ML estimation

This section provides a review of EM-type algorithms for maximum-likelihood estimation of parameters in parametric models. The algorithms are described in very generic terms; specific applications of EM and its extensions will appear in later chapters.

1.6.1 Standard EM

Under an assumption of ignorability, the maximum-likelihood (ML) estimates for the parameters of an imputation model are found by maximizing the observed-data likelihood, $L(\theta|Y_O)$, or the observed-data log-likelihood, $l(\theta|Y_O) = \log L(\theta|Y_O)$. Closed-form expressions for these estimators usually do not exist, and they must be found by iterative procedures. An EM algorithm (Dempster, Laird & Rubin, 1977) maximizes $l(\theta|Y_O)$ by repeatedly solving an easier complete-data problem that resembles a maximization of $l(\theta|Y) = \log P(Y_O, Y_M; \theta)$.

EM is defined as follows. Given a provisional estimate $\theta^{(t)}$ of the unknown parameter, let

$$Q(\theta|\theta^{(t)}) = \int l(\theta|Y) P(Y_M|Y_O, \theta^{(t)}) dY_M \tag{1.1}$$

denote the average of $l(\theta|Y)$ over the predictive distribution $P(Y_M|Y_O, \theta^{(t)})$. EM maximizes $L(Y_O|\theta)$ by iteratively maximizing $Q(\theta)$. At iteration $t+1$, we perform two steps:

*E-step: C*ompute $Q(\theta|\theta^{(t)})$ by averaging the complete-data log-likelihood, $l(\theta|Y)$, over $P(Y_M|Y_O, \theta^{(t)})$.

M-step: Update θ with $\theta^{(t+1)}$ by maximizing $Q(\theta|\theta^{(t)})$ over the parameter space Θ.

Under regularity conditions clarified by Wu (1983), EM will converge reliably to a stationary point of the observed-data log-likelihood. If this function is well-behaved, the resulting stationary point of the EM algorithm is the unique global ML estimate of θ.

The EM algorithm has been extended in various ways. These extensions include ECM (Meng & Rubin, 1993), ECME (Liu & Rubin, 1994), AECM (Meng & van Dyk, 1997) and PX-EM (Liu, Rubin & Wu, 1998, Little & Rubin, 2002). Some of these extensions will be used in this dissertation, and we briefly review those extensions here,

1.6.2 The ECM algorithm

In some applications, the complete-data maximum likelihood estimation required for the M-step is not analytically tractable and would itself require iteration, making EM less attractive. The Expectation-Conditional Maximization (ECM) algorithm (Meng & Rubin, 1993) uses the same E-step as EM, but replaces a complicated M-step with a sequence of conditional maximization (CM) steps. Each CM step consists of S substeps. For $s = 1, \ldots, S$, the sth substep in the tth iteration of the ECM maximizes $Q(\theta|\theta^{(t)})$ as defined in (1.1) not over the whole parameter space Θ but subject to some constraint $g_s(\theta) = g_s(\theta^{\{t+(s-1)/S\}})$. After doing this for $s = 1, \ldots, S$, the output from the tth iteration, $\theta^{(t+S/S)} = \theta^{(t+1)}$, becomes the starting value for the $(t+1)$th iteration.

The set of ECM constraint functions $G = \{g_s(\theta); s = 1, \ldots, S\}$ must be preselected and must be fulfill certain conditions to guarantee that the ECM algorithm appropriately converges to the maximizer of $L(Y_O|\theta)$. These conditions require G to be "space filling" in a sense defined by Meng and Rubin (1993). One special

example of G that usually satisfies these conditions is $G = (\theta_1, \ldots, \theta_S)$, where $\theta_1, \ldots, \theta_S$ is a partition of the parameter vector θ into subvectors. In this case, the sth substep in the CM-step maximizes the Q function with respect to θ_s, holding the remaining parameters fixed at their current values.

1.6.3 The ECME algorithm

Despite its stability and reliable convergence behavior, EM may converge slowly in some applications. The Expectation-Conditional Maximization Either (ECME) algorithm replaces one or more CM substeps in with steps that maximizes the corresponding observed-data loglikelihood $l(\theta|Y_O)$ (Liu & Rubin, 1994). ECME tends to converge more rapidly than EM or ECME because it operates on the actual loglikelihood in these substeps rather than an approximation to it. The constraint functions $G = \{g_s(\theta); s = 1, \ldots, S\}$ used in ECME need to satisfy the same space-filling conditions defined by Meng and Rubin (1993).

1.6.4 The PX-EM algorithm

Another method for accelerating EM is the Parameter-Expanded EM (PX-EM) algorithm (Liu et al., 1998). PX-EM is an EM algorithm applied to an enlarged complete-data model which appends an additional parameter α to θ, written as $P_X(Y|\Theta = (\theta_\star, \alpha))$, where $\theta_\star$ is to the enlarged model what θ is to the original model. Liu *et al.* (1998) prove that PX-EM dominates EM in global rate of convergence because it scales down the rate of missing information. The more the complete-data model is expanded, the faster the resulting PX-EM algorithm becomes, as long as the extra computational cost is negligible.

The expanded parameter set $\Theta = (\theta_\star, \alpha)$ must meet two conditions. First,

the original observed-data model parameters must be preserved via a many-to-one known transformation T on the expanded parameter set, $\theta = T(\theta_\star, \alpha)$. Second, the original complete-data model parameters must be obtainable by setting α to a null value α_0,

$$P_X(Y|\Theta = (\theta_\star, \alpha = \alpha_0)) = P(Y|\theta).$$

Some examples of EM, ECME and PX-EM algorithms will be applied to factor analysis models in Chapter 3 and Chapter 4.

1.7 Basics of multiple imputation

1.7.1 Step 1: Creating the imputations

Multiple imputation (Rubin, 1987; 1996) involves three distinct steps: (1) M sets of 'complete' data are formed by simulating the missing data M times. (2) Each of these M sets of 'complete' data is analyzed by standard complete-data methods. (3) The results are combined using simple rules developed by Rubin (1987) and others.

In step 1, we create M independent draws from the distribution of $P(Y_M|Y_O)$, which is a Bayesian posterior predictive distribution for the missing data given the observed data. This distribution can be written as

$$P(Y_M|Y_O) = \int P(Y_M|Y_O, \theta)\, P(\theta|Y_O)\, d\theta. \tag{1.2}$$

The distribution $P(\theta|Y_O)$, which is the observed-data posterior density for θ, is proportional to the product of a prior density of θ, say $\pi(\theta)$, and the observed-

data likelihood function,

$$P(\theta|Y_O) \propto \pi(\theta)\, L(\theta|Y_O).$$

In multivariate missing-data problems, directly simulating Y_M from $P(Y_M|Y_O)$ tends to be difficult, and these imputations are usually generated by Markov chain Monte Carlo (MCMC).

One of the most commonly used MCMC procedures for multiple imputation is known as data augmentation (DA) (Tanner & Wong, 1987; Schafer, 1997). DA is an iterative two-step method that bears a superficial resemblance to EM but whose purpose is very different. EM is a deterministic algorithm that converges to a (possibly local) maximizer of $L(\theta|Y_o)$, whereas DA is a simulation method which, after it has achieved stationarity, produces (usually dependent) draws of (θ, Y_M) from $P(\theta, Y_M|Y_O)$. Each cycle of DA consists of an Imputation or I-step followed by a Posterior or P-step. In the I-step, we update the missing values by drawing them from their predictive distribution given the observed data and current values for the parameters $\theta^{(t)}$,

$$Y_M^{(t+1)} \quad \text{from} \quad P(y_M|Y_O, \theta^{(t)}).$$

In the P-step, we update the parameters by drawing them from their posterior distribution given the observed and simulated missing data,

$$\theta^{(t+1)} \quad \text{from} \quad P(\theta|Y_O, Y_M^{(t+1)}).$$

After repeating the procedure many times, the simulated value of Y_M eventually becomes a draw from $P(Y_M|Y_O)$, the posterior predictive distribution from

which multiple imputations are generated. Sucessive values $Y_M^{(t)}$ and $Y_M^{(t+1)}$ are usually correlated, but MI requires M independent draws from $P(Y_M|Y_O)$. These draws are usually obtained by subsampling the chain, retaining every kth draw, $Y_M^{(k)}, Y_M^{(2k)}, Y_M^{(3k)}, \ldots$, where k is large enough to achieve approximate independence (Schafer, 1997). The simulated values of θ, which cane be regarded as dependent draws from $P(\theta|Y_O)$, are not of primary interest when DA is intended for imputation. These draws, however, may be useful for simulation-based Bayesian inference regarding parameters of the imputation model, and they are also typically examined to monitor the convergence behavior of the algorithm.

If the rates of missing information are modest, only a small number of imputations M may be adequate for post-imputation analyses. These rates of missing information, which are defined by Rubin (1987) and Schafer (1997), are determined by the information (second derivatives) of $l(\theta|Y_o)$ and the expected second derivatives of $l(\theta|Y_O, Y_M)$ with respect to $P(Y_M|Y_O, \theta)$. It has frequently been suggested that $M = 5$ will be sufficient for typical applications. Choosing a larger value for M will reduce the degree of Monte Carlo error, but this error often accounts for only a relative small portion of the overall inferential uncertainty associated with the final estimands. The rules for computing standard errors from a multiply imputed analysis—which will be given below—explicitly account for random error due to a finite number of imputations. In situations where of missing information are thought to be large, a more generous number of imputations is recommended (say, $M = 30$, Meng, 1994). The rules for combining results will yield an estimate of the rate of missing information for any estimand, but this estimate rate can be noisy for small values of M. For this reason, many researchers who use MI are now routinely using $M = 25$ or more.

1.7.2 Step 2: Analyzing the imputed data

Many different analysis models may be applied to imputed datasets for addressing different scientific questions. MI was first proposed by Rubin (1987) for complex surveys in which public-use data sets are shared by many users. Statistical proficiency, objectives and research questions vary across users. Database constructors may have extra information about why values are missing, but sharing this information with users may be infeasible or undesirable. MI techniques were originally designed for situations where imputation and analysis are carried out by different persons or organizations. If the imputer possesses more information than the analyst (e.g., variables that are not released to maintain confidentiality) and incorporates this information into the imputation model, it creates a form of discrepancy between the imputation and analysis models that is statistically advantageous (Meng, 1994). Results from this analysis may have greater efficiency than estimates based on the analysis variables alone. Detailed discussion on properties of MI when the imputation and analysis models differ is given by Meng (1994), Rubin (1996), and Collins *et al.* (2001).

1.7.3 Step 3: Consolidating the results

The most common way to consolidate results from post-imputation analyses is to combine the M sets of estimates and standard errors using the rules presented by Rubin (1987). For a scalar estimand Q, the multiple-imputation estimate is

$$\hat{Q} = \frac{1}{M} \sum_{i=1}^{M} Q_{(i)},$$

and the variance estimate is

$$\mathrm{Var}(\hat{Q}) = \frac{1}{M}\sum_{i=1}^{M}\sigma^2_{(i)} + (1+\frac{1}{M})(\frac{1}{M-1})\sum_{i=1}^{M}\left(Q_{(i)}-\hat{Q}\right)^2,$$

where $Q_{(i)}$ and $\hat{V}_{(i)}$ denote the estimate and variance estimate for Q, respectively, from the ith imputed data set, $i = 1, \ldots, M$. If M is sufficiently large, then the overall estimate $\hat{Q}$ will be approximately normally distributed about Q, so normal-theory confidence intervals and tests may be used. More accurate approximations based on a Student's t-distribution are also available, but the degrees of freedom grow quickly as M increases unless the rate of missing information is unusually high.

1.8 Looking ahead

The focus of this dissertation is to develop multiple-imputation procedures that are better suited than existing methods to data from multi-themed survey questionnaires.

In Chapter 2, we first review algorithms for ML estimation and multiple imputation under a general multivariate normal model assuming an unstructured covariance matrix. We then discuss why we often cannot rely on these unstructured-covariance procedures to handle data from multi-themed questionnaires. We also review some alternatives that have been proposed in the literature, including the use of exploratory factor models to address problems of overparameterization.

In Chapter 3, we develop the idea of imputation under factor-analytic covariance structures. We begin with techniques for exploratory factor models proposed by Song and Belin (2004) and extend them to confirmatory models which are bet-

ter suited to *a priori* notions of how the items in a multi-themed questionnaire are interrelated. We present a new MCMC algorithm for generating multiple imputations under a confirmatory factor model. We also present a new PX-EM algorithm to compute maxima of the likelihood function that will help us to explore alternative models and provide reasonable starting values for MCMC sampling.

In Chapter 4, we acknowledge that, however appealing a confirmatory factor model may seem, the relationships among variables from an actual questionnaire might depart substantially from the assumptions of that model. Instead of assuming that the factor model is precisely correct, we relax the model by allowing "soft constraints." That is, we allow the actual covariance matrix to deviate from the factor model by random amounts described by an inverse-Wishart distribution with unknown degrees of freedom. This idea, which was first suggested by Boscardin and Zhang (2004), is developed and implemented in the context of a confirmatory factor analysis. The technique can be regarded as a kind of Bayesian smoothing that pulls the covariances toward a parsimonious structure.

In Chapter 5, we present results from a series of simulations to compare the new procedures with existing approaches.

Extensions of these methods and future directions for this research are discussed in Chapter 6.

Chapter 2

Overview of Current Methods for Multiple Imputation of High-dimensional Data

2.1 The multivariate normal model with unstructured covariance matrix

2.1.1 The model

The imputation model most commonly used in practice is the multivariate normal model with an unstructured covariance matrix. An advantage of this model is its flexibility; the resulting imputed datasets will be compatible with many kinds of post-imputation analyses. The per-iteration cost of the basic EM and DA algorithms—both in terms of memory and floating-point operations—is low. The generality of this model becomes a drawback as the number of variables increases, however, and applications to multi-theme questionnaires may lead to overparam-

eterization.

The model that we describe here is a slight generalization of the model described by Schafer (1997), Little and Rubin (2002) and others that conditions upon covariates that are completely observed. It extends the generic multivariate normal distribution to a multivariate normal linear regression.

We assume that observational units are independent. Incomplete variables will be put into the columns of a response matrix,

$$Y = \begin{bmatrix} {y_1}^T \\ {y_2}^T \\ \vdots \\ {y_n}^T \end{bmatrix} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1r} \\ y_{21} & y_{22} & \cdots & y_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{nr} \end{bmatrix},$$

and completely observed variables (which we call covariates) are placed in the columns of another matrix,

$$X = \begin{bmatrix} {x_1}^T \\ {x_2}^T \\ \vdots \\ {x_n}^T \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}.$$

In most applications of this model, the first column of X will be a constant, $x_{i1} \equiv 1$. Because X is fully observed, it will not be explicitly modeled. We assume that each row of Y given the covariates in X is independently normally distributed as

$$y_i \sim N(\beta^T x_i, \Sigma),$$

where Σ is a $(r \times r)$ positive definite covariance matrix, and β $(p \times r)$ denotes the

matrix of regression coefficients of Y on X. Another way to write this model is

$$Y = X\beta + \epsilon, \tag{2.1}$$

where the error term ϵ is an $(n \times r)$ matrix of residuals distributed as $\text{vec}(\epsilon) \sim N(0, \Sigma \otimes I_n)$, and I_n denotes the $n \times n$ identity matrix. Without further restrictions on the parameter space, the free parameters in this model are the $p \times r$ elements of β and the $r(r+1)/2$ elements in the upper triangle of Σ.

2.1.2 Complete-data log-likelihood

Aggregating over the independent units $i = 1, \ldots, n$, the complete-data loglikelihood function can be written as

$$l(\beta, \Sigma|Y) \; = \; -\frac{n}{2}\log|\Sigma| - \frac{1}{2}\,\text{tr}\,\Sigma^{-1}(Y - X\beta)^T(Y - X\beta). \tag{2.2}$$

This is a regular exponential family, and the loglikelihood is a linear function of the sufficient statistics $T_1 = X^TY$ and $T_2 = Y^TY$. With complete data, ML estimates for β and Σ are obtained by solving the moment equations in which the realized values of T_1 and T_2 are set equal to their expectations (Cox & Hinkley, 1974). The complete-data ML estimates are

$$\hat{\beta} = (X^TX)^{-1}T_1, \tag{2.3}$$

$$\hat{\Sigma} = \frac{1}{n}(T_2 - {T_1}^T(X^TX)^{-1}T_1). \tag{2.4}$$

2.1.3 ML estimation with incomplete data

If elements of the Y matrix are ignorably missing, ML estimates for β and Σ may be computed by an EM algorithm. For the E-step of EM, we must calculate the expectations of the sufficient statistics $T_1 = \sum_{i=1}^{n} x_i y_i^T$ and $T_2 = \sum_{i=1}^{n} y_i y_i^T$ with respect to the predictive distribution $P(Y_M|Y_O, \beta, \Sigma)$ under assumed values for β and Σ. (Again, in this notation, conditioning on X has been assumed.) Given β and Σ, the missing elements in any row y_i have a multivariate normal regression on the observed elements of that row and all of the elements of x_i.

For the ith row, denote the observed and missing parts of y_i by y_{iO} and y_{iM}, respectively. Let β_{iO} and β_{iM} denote the submatrices of β consisting of the columns that correspond to y_{iO} and y_{iM}. Similarly, suppose we partition Σ into four submatrices that correspond to y_{iO} and y_{iM} in the obvious way, and call these submatrices Σ_{iOO}, Σ_{iOM}, Σ_{iMO}, and Σ_{iMM}. The expectations needed for the E-step of EM are

$$
\begin{aligned}
E(y_{iO}|y_{iO}, \beta, \Sigma) &= y_{iO}, \\
E(y_{iM}|y_{iO}, \beta, \Sigma) &= \beta_{iM}^T x_i + \Sigma_{iMO}\Sigma_{iOO}^{-1}(y_{iO} - \beta_{iO}^T x_i), \\
E(y_{iO}y_{iO}^T|y_{iO}, \beta, \Sigma) &= y_{iO}y_{iO}^T, \\
E(y_{iO}y_{iM}^T|y_{iO}, \beta, \Sigma) &= y_{iO}E(y_{iM}|y_{iO}, \beta, \Sigma), \\
E(y_{iM}y_{iM}^T|y_{iO}, \beta, \Sigma) &= E(y_{iM}|y_{iO}, \beta, \Sigma)E(y_{iM}|y_{iM}, \beta, \Sigma)^T \\
&\quad + \Sigma_{iMM} - \Sigma_{iMO}\Sigma_{iOOO}^{-1}\Sigma_{iOM}.
\end{aligned}
$$

The E-step accumulates these expectations over the units $i = 1, \ldots, n$, producing the expected values for T_1 and T_2. The matrices Σ_{iOO}^{-1} and $\Sigma_{iMM} - \Sigma_{iMO}\Sigma_{iOO}^{-1}\Sigma_{iOM}$ may be calculated by a SWEEP operator as described by Little and Rubin (2002) and Schafer (1997), by sweeping Σ on the positions corresponding to the observed

variables.

Notice that the parameters of the multivariate regression of y_{iM} on y_{iO}—the vector of intercepts, matrix of slopes and matrix of residual covariances—are the same for all observational units i having the same missingness pattern. Therefore, it is often helpful to organize the E-step computations by grouping the observational units $i = 1, \ldots, n$ according to the patterns of missingness found in Y, so that the total number of sweeps may be reduced. The rows associated with any missingness pattern have the same subset of variables observed. For example, suppose that $n = 5$ and the data matrix Y has three variables ($r = 3$) which we denote by Y_1, Y_2, and Y_3. And suppose that the matrix of missingness indicators is

$$M = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \end{pmatrix}.$$

In this example, only three missingness patterns are relevant: (1) no missing values for any variable, (2) only Y_3 missing, and (3) Y_2 and Y_3 both missing. The fourth row, which has all three variables missing, does not need to be taken into consideration when computing ML estimates because this row contributes nothing to the observed-data likelihood. Including this row would merely increase the rates of missing information, causing EM to converge more slowly. This row may be included in an MI procedure, however, because the analyst might possibly need imputed values for this row.

For the M-step of EM, we simply update the estimated values for β and Σ

according to (2.3) and (2.4) with T_1 and T_2 replaced by their expectations from the E-step. Because the complete-data loglikelihood is a linear function of these sufficient statistics, the expected value of the loglikelihood function is obtained by replacing T_1 and T_2 by their expected values.

2.1.4 Multiple imputations for missing values

Multiple imputations for the missing data Y_M under this unstructured normal model can be simulated by a straightforward application of the data augmentation (DA) method described in Chapter 1. Given the current simulated versions of the unknown parameters $\theta^{(t)} = (\beta^{(t)}, \Sigma^{(t)})$ and missing data $Y_M^{(t+1)}$, one iteration of DA consists of an Imputation or I-step followed by a Posterior or P-step. In the I-step, we draw

$$y_{iM}^{(t+1)} \quad \text{from} \quad P(y_{iM}|Y_O, \theta^{(t)}) \tag{2.5}$$

independently for $i = 1, \ldots, n$, and in the P-step, we draw

$$\Sigma^{(t+1)} \quad \text{from} \quad P(\Sigma|Y_O, Y_M^{(t+1)}), \tag{2.6}$$

followed by

$$\beta^{(t+1)} \quad \text{from} \quad P(\theta|Y_O, Y_M^{(t+1)}, \Sigma^{(t+1)}). \tag{2.7}$$

Starting with initial values ($\beta^{(0)}$, $\Sigma^{(0)}$, $Y_M^{(0)}$) and executing (2.5)-(2.7) repeatedly creates a sequence $\{(Y_?^{(t)}, \beta^{(t)}, \Sigma^{(t)}), t = 1, 2, 3, \ldots\}$, whose limiting distribution is $P(Y_M, \beta, \Sigma|Y_O)$.

The simulation of y_{iM} in (2.5) is closely related to the E-step computation that

was just described. Under that notation,

$$\begin{bmatrix} y_{iO}^T \\ y_{iM}^T \end{bmatrix} \sim N\left([\beta_{iO}, \beta_{iM}]^T x_i, \begin{bmatrix} \Sigma_{iOO} & \Sigma_{iOM} \\ \Sigma_{iMO} & \Sigma_{iMM} \end{bmatrix} \right).$$

It follow that the conditional distribution of y_{iM} given y_{iO} and the parameters (β, Σ) is multivariate normal with mean vector

$$\beta_{iM}^T x_i + \Sigma_{iMO}\Sigma_{iOO}^{-1}(y_{iO} - \beta_{iO}^T x_i)$$

and covariance matrix

$$\Sigma_{iMM} - \Sigma_{iMO}\Sigma_{iOO}^{-1}\Sigma_{iOM}.$$

Simulating the missing values in each row $i = 1, \ldots, n$ completes the I-step. As before, it may be computationally advantageous to apply the SWEEP operator and to group the rows by their missingness patterns so that the number of sweeps may be reduced.

To implement the P-step defined by (2.6) and (2.7), we need a joint prior distribution for β and Σ. For β, it is customary to apply uniform "density" over the $(r \times p)$-dimensional real space. This is not a proper density function, because its integral is not finite. Nevertheless, it leads to a proper posterior distribution under most circumstances of interest to us. A natural conjugate class of prior distributions for Σ is the inverted-Wishart family. Suppose we apply the prior distribution $\Sigma^{-1} \sim W(\xi, \Lambda)$, where $\Lambda > 0$ and $\xi \geq r$ are user-specified hyperparameters. If Y were fully observed, this would leads to a complete-data posterior distribution

$$\Sigma^{-1} \mid Y \sim W(\xi', \Lambda'),$$

$$\text{vec}(\beta) \mid Y, \Sigma \sim N(\text{vec}(\hat{\beta}), \Sigma \otimes (X^T X)^{-1}),$$

where $\xi' = \xi + n - p$ and $\Lambda' = [\Lambda^{-1} + (Y - X\hat{\beta})^T (Y - X\hat{\beta})]^{-1}$.

2.1.5 Prior distributions

When prior information about Σ is scarce, it is traditional to apply an improper Jeffrey's prior whose density can be regarded as the limit of the inverted Wishart density as $\xi \to 0$ and $\Lambda^{-1} \to 0$. Assuming that the necessary inverses exist, this leads to the complete-data posterior

$$\Sigma^{-1} \mid Y \sim W(n - p, [(Y - X\hat{\beta})^T (Y - X\hat{\beta})]^{-1}),$$
$$\text{vec}(\beta) \mid Y, \Sigma \sim N(\text{vec}(\hat{\beta}), \Sigma \otimes (X^T X)^{-1}).$$

Drawing Σ from this distribution is straightforward. For simulating β, we can apply the Cholesky factorizations $\Sigma = G^T G$ and $(X^T X)^{-1} = H^T H$, where G and H are a lower-triangular square roots of $G^T G$ and $(X^T X)^{-1}$, respectively. It is easy to show that $G \otimes H$ is then a lower-triangular square root of $\Sigma \otimes (X^T X)^{-1}$. A random draw of $\text{vec}(\beta)$ can thus be obtained as $\text{vec}(\hat{\beta}) + (G \otimes H)z$, where z is a vector of independent standard normal variates of length $p \times r$.

Alternative choices for the prior distribution may help to stabilize inferences when rates of missing information are high (Schafer, 1997, 2008) has described four different versions of the prior distribution for Σ: (a) a uniform prior, which can be viewed as a limiting case of the inverted-Wishart density as $\xi \to -(r + 1)$ and $\Lambda^{-1} \to 0$; (b) the Jeffreys prior, $\xi \to 0$ and $\Lambda^{-1} \to 0$; (c) a data-dependent 'ridge' prior with a user-specified smoothing parameter, which smooths the estimated correlations toward zero; and (d) an inverted Wishart prior with user-specified ξ

and Λ. These four choices can handle many situations encountered in practice.

For certain types of applications, however, the normal model with an unstructured covariance matrix is overparameterized. Overparameterization may occur when the number of variables is very large, or when the number of cases is not substantially higher than the number of variables. For example, with $r = 100$ response variables and $p = 50$ covariates, the unstructured-covariance model has 10,050 free parameters — 5,000 regression coefficient parameters and 5,050 covariances — and the the inference becomes ill-conditioned. To overcome difficulties caused by overparameterization, Schafer (1994) suggests two possible solutions. One is to trim the model by omitting less important variables; the other is to apply the mildly informative ridge prior. Neither of these proposed remedies is particularly attractive for applications to multi-themed questionnaires, for reasons to be described later.

2.1.6 Software

Routines for ML estimation and multiple imputation for the normal model with unstructured covariance matrix have been available for some time. Some of these packages do not allow covariates, which is tantamount to setting $X = (1, 1, \ldots, 1)^T$, as in earlier versions of the NORM program (Schafer, 1997). This is not necessarily a limitation, however, because if completely observed covariates are present, they may also be placed into the columns of Y.

The most version of NORM, which is a library for R, implements the EM and DA algorithms described above (Schafer, 2008). Other procedures that have similar capabilities include the SAS macro MISS and COMBINE (Allison, 1999), the SAS procedure PROC MI and PROC MIANALYZE (Yuan, 2000), the missing-data

library in S-Plus (Schimert *et al.*, 2001), LISREL (Jöreskog, *et al.*, 2001), the Stata module INORM (StataCorp, 2007), and the multiple imputation features in the latest version of SPSS. Amelia (Kind et al., 2007) employs the EM algorithm but uses a different computational technique for MI based on importance resampling. EMCOV (Graham & Hofer, 1993) implements an EM algorithm for ML under the normal model. HLM (Raudenbush, 2004) and Mplus (Muthén & Muthén, 1998) do not generate imputations, but they do support the analysis of multiply-imputed datasets.

2.2 Multiple imputation by chained equations

An frequently cited limitation of the multivariate normal imputation model is that each variable is assumed to be continuously distributed, and its relationships to all other variables are assumed to be additive and linear. This simply does not correspond to the variables typically obtained from survey questionnaires which may be binary, ordinal, or nominal, and which may be related to other variables in complicated ways. Creating a model that can plausibly describe the joint distribution of variables like these can be daunting, especially when the number of variables is large. To circumvent the difficulty of specifying one joint model for all of the items, some have proposed to build an implicit 'model' by specifying a regression for each variable on a subset of the others. Procedures for multiple imputation based on a sequence of univariate-response regression models is called chained equations (Van Buuren & Oudshoorn, 1999; Raghunathan *et al.*, 2000). Each regression in the chain may take a different form depending on the type of variable being modeled: logistic regression for a binary variable, polytomous regression for an ordinal or nominal variable, log-linear regression for a count variable, and so on. These mod-

els are intended to reflect relationships actually seen in the data, and may include nonlinear effects and interactions.

The chained equation method for multiple imputation is an iterative simulation procedure that resembles a Gibbs sampler, but is usually not a true MCMC procedure because the stationary distribution is nonexistent. The missing values are first initialized by an ad hoc imputation method. For each regression model in the chain, the parameters are estimated from the outcomes (observed values only) and the predictor variables (observed and imputed values). New values for the regression parameters are sampled from their approximate posterior distribution, and the model with simulated parameters is then used to randomly impute the missing responses. The tth iteration of the algorithm can be expressed as

$$\begin{aligned}
\theta_1^{(t)} &\sim P(\theta_1|Y_1^O, Y_2^{(t-1)}, \ldots, Y_r^{M(t-1)}) \\
Y_1^{M(t)} &\sim P(Y_1^M|Y_1^O, Y_2^{(t-1)}, \ldots, Y_r^{M(t-1)}) \\
&\vdots \\
\theta_r^{(t)} &\sim P(\theta_r|Y_r^O, Y_2^{M(t)}, \ldots, Y_{r-1}^{M(t)}) \\
Y_r^{M(t)} &\sim P(Y_r^M|Y_r^O, Y_2^{M(t)}, \ldots, Y_{r-1}^{M(t)}),
\end{aligned}$$

where θ_1, θ_2, ..., θ_r, denote the parameters of the respective regression models (Van Buuren & Oudshoorn, 1999; Jacobusse, 2005). The process is repeated until dependence on the starting values is thought to have died down.

Chained equations methods are not mathematically rigorous for the following reason: the conditional distributions specified by the sequence of regression models may not define a joint distribution. In many applications, there will be no joint probability distribution that has the specified regression models as its

full conditionals. Proponents of chained-equation methods freely acknowledge this incoherent aspect and do not think it is usually problematic, but the practical implications of inconsistent conditional distributions are not well understood.

Despite this theoretical problem, chained equations are becoming increasingly popular. Imputation by chained equations is seen to be an attractive alternative to more rigorous methods based on joint modeling, because it can accommodate large numbers of variables at once. Software for chained equations can be found in WinMICE and an R package called MICE (Van Buuren & Oudshoorn 1999; Jacobusse, 2005), IVEWARE (Raghunathan *et al.*, 2000), and routines for Stata (StataCorp, 2003).

2.3 Multivariate linear mixed models

A more rigorous way to address overparameterization with large numbers of variables is to construct a joint model for all variables in question but impose constraints on the covariances to reflect a priori notions about how the variables may be related. Although this has not yet been described for classes of applications involving multi-themed questionnaires, it has been tried for longitudinal surveys (panel studies) in which a common set of items is repeated in multiple waves of data collection. Schafer and Yucel (2002) present methods for multiple imputation under a multivariate linear mixed model that is formally equivalent to the normal regression model in Section 2.1, except that the covariance matrix is assumed to have a Kronecker-product form that is consistent with the notion of repeated measurements over time.

To describe the model of Schafer and Yucel (2002), we will use a slightly different notation from that in Section 2.1. Let y_i denote an $n_i \times r$ matrix of response

matrix for sampled unit i, $i = 1, 2, ...m$. The rows of y_i represent occasions, and the columns of y_i represent variables measured at these occasions. Suppose that portions of the response matrices y_1,y_2,...,y_m are ignorably missing in the sense described in Chapter 1. Schafer and Yucel (2002) suppose that

$$y_i = X_i\beta + Z_i b_i + \epsilon_i, \tag{2.8}$$

for $i = 1, \ldots, m$, where $X_i(n_i \times p)$ and $Z_i(n_i \times r)$ are matrices of fully observed covariates, β $(p \times r)$ is a matrix of regression coefficients common to all units, and b_i is a $(q \times r)$ matrix of random regression coefficients specific to unit i. The matrix b_i is assumed to be distributed as $\text{vec}(b_i) \sim N_q\ (0, \Psi)$, and each row of ϵ_i is assumed to be independently normally distributed with mean zero and covariance matrix Σ. The covariance matrices Ψ and Σ are unknown, and the b_i's and ϵ_i are assumed to independent of one another. The unknown parameters of the model are $\theta = (\beta, \Sigma, \Psi)$. Covariates describing the ith unit that do not change over time may be included in X_i, and time-varying covariates may be placed into the columns of X_i and also in Z_i.

Schafer and Yucel (2002) implemented a Gibbs sampling procedure for multiple imputation of missing values in the y_i matrices under model (2.8). At iteration t, the parameters $\theta^{(t)} = (\beta^{(t)}, \Sigma^{(t)}, \Psi^{(t)})$ and missing data $Y_M^{(t)}$ are updated in three steps. First, draw random effects

$$b_i{}^{(t+1)} \sim P(b_i|\theta^{(t)}, Y_O, Y_M^{(t)}) \tag{2.9}$$

independently for $i = 1, 2, ...m$; second, update the parameters,

$$\theta^{(t+1)} \sim P(\theta|Y_O, Y_M^{(t)}, B^{(t+1)}); \tag{2.10}$$

and finally, impute the missing values,

$$y_{iM}^{(t+1)} \sim P(y_{iM}|Y_O, B^{(t+1)}, \theta^{(t+1)}), \tag{2.11}$$

for $i = 1, \ldots, m$. Given starting values $\theta^{(0)}$ and $Y_M^{(0)}$, repeating the cycle (2.9)-(2.11) yields sequences $\{\theta^{(t)}, t = 1, 2, 3, \ldots\}$ and $\{Y_M^{(t)}, t = 1, 2, 3, ...\}$ whose limiting (stationary) distributions are $P(\theta|Y_O)$ and $P(Y_M|Y_O)$, respectively.

In addition to this Gibbs sampling procedure for creating multiple imputations, Schafer and Yucel (2002) also describe EM-type algorithms for ML estimation of the model parameters. Procedures for MI under this multivariate linear mixed model have been made available in a library called PAN (Schafer & Yucel, 2001) which has been converted to an R package (Junhua Zhao, 2009).

2.4 Factor models

2.4.1 Exploratory and confirmatory factor analysis

Under the multivariate linear mixed model just described, the covariance matrix of the responses is assumed to be

$$V(\text{vec}(y_i)) \; = \; (I_r \otimes Z_i)\Psi(I_r \otimes Z_i)^T \; + \; (\Sigma \otimes I_{n_i}).$$

This covariance structure may be appropriate for a set of questionnaire items that is repeated over time, but it is not well suited to handle items from multi-theme questionnaires. A more reasonable way to begin to construct models for multi-theme questionnaires is to suppose that the items for each theme are conditionally independent given a small number of latent factors.

Let us now return to the notation of Section 2.1 where Y denotes a data matrix with rows corresponding to independent sampled units and columns corresponding to variables. One simple form for a factor model would assume that rows $i = 1, \ldots, n$ of Y are independently distributed as

$$y_i \mid z_i \sim N_r(\beta^T x_i + \gamma^T z_i, \tau^2),$$

where z_i $(k \times 1)$ denotes a vector of unseen factor scores that is normally distributed with mean zero, variances constrained to be equal to one, and correlation matrix R $(k \times k)$. This generalizes the classical common-factors model of Thurstone (1947) to include regressors x_i in the mean structure. Note that this is not the usual way that covariates are incorporated into a factor model. In structural-equations modeling, it is the latent variables z_i, not the manifest variables y_i, that are typically regressed on the exogeneous covariates x_i, and that is done for reasons that are theoretical and substantive. Our version of the model is akin to regressing y_i on x_i and fitting a common-factors model to the residuals. We have chosen to arrange the model in this fashion because the implied unconditional distribution for y_i becomes $N_r(\beta^T x_i, \Sigma)$ with $\Sigma = \gamma^T R \gamma + \tau^2$, which is a special case of the model presented in Section 2.1 and facilitates comparisons between these models. This model is designed not for scientific description but for imputation.

The assumption of unit variances for z_i, or some other set of restrictions, for

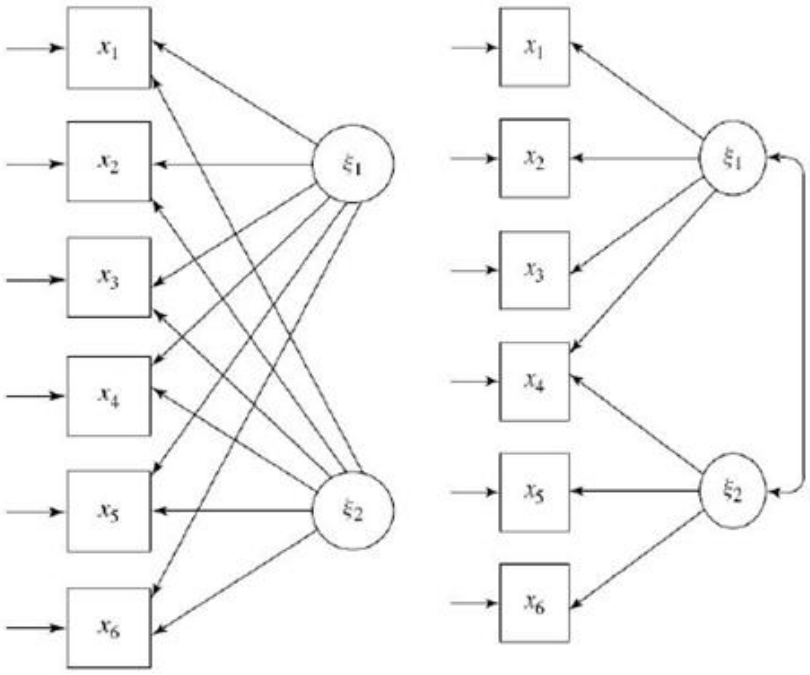

Figure 2.1. EFA and CFA (Joreskog, 2007)

identification. The elements of the matrix γ $(k \times r)$ are called the factor loadings, and $\tau^2 = \text{diag}\ (\tau_1^2, \ldots, \tau_r^2)$ is the matrix with elements $\tau_1^2, \ldots, \tau_r^2$ on the diagonal which are called the uniquenesses. In a common-factors model, the items may be allowed to have factor loadings on the entire set of factors, or some of the loadings may be fixed at zero. Three special cases are defined by restrictions on R and γ (Rubin & Thayer, 1982).

Case 1: $R = I_k$ and no restrictions on γ;

Case 2: $R = I_k$ and *a priori* zeroes in γ;

Case 3: *a priori* zeroes in γ but no extra restrictions on R.

Case 1 is usually called an exploratory factor analysis (EFA) model. In exploratory factor analysis (EFA), all of the factor loadings are free to vary. An example of

EFA is depicted by the path diagram on the left-hand side of Figure (2.1); all variables load on both factors, and the factors are assumed to be uncorrelated.

Cases 2 and 3 are usually referred to as confirmatory factor analysis (CFA) models. With CFA, the researcher is seeking to fit a model where certain elements in γ are assumed to be zero, usually after they were seen to be small when the results from an EFA were examined. CFA represents a refinement of EFA based on empirical findings and theory. In CFA, the factors are usually allowed to be correlated. An example of CFA is shown on the right-hand side of Figure (2.1). In that model, factor ξ_1 does not affect variables x_1, x_2 or x_3, whereas factor ξ_2 does not affect variables x_4, x_5, or x_6. The factor loadings of x_1, x_2, and x_3 on ξ_1, and the factor loadings of x_4, x_5, and x_6 on ξ_2, are assumed to be zero, and the two factors are allowed to be correlated.

2.4.2 Previous work on EFA

Algorithms for multiple imputation under the EFA model were originally developed by Song and Belin (2004). They proposed their method for imputing missing values in high-dimensional multivariate datasets where the number of variables is comparable to the number of sampled cases. Their prototypical example was a psychological test with 100 items administered to 100 subjects. They assumed no prior knowledge about how the underlying factors were related to the items. In their model, they allowed elements of Y to be ignorably missing, but they did not allow covariates, i.e. they assumed that $X = (1, 1, \ldots, 1)^T$. Compared to the model with unstructured covariance matrix described in Section 2.1, the number of covariance parameters has been reduced from $r(r+1)/2$ to $rk + r$.

The methods of Song and Belin (2004) require that the number of factors

be pre-specified. Because this is an exploratory analysis, the number of factors can be chosen empirically based on quality of fit. Using a maximum likelihood approach, they tested the null hypothesis that $k = k_0$ versus the unstructured alternative for various values of k_0. The number of factors can also be chosen by ad hoc rules based on eigenvalues of the correlation matrix. In that approach, one would apply principle components analysis (PCA) to the correlation matrix estimated under the unrestricted model of Section 2.1. The eigenvalues represent the variance explained by each underlying factor. The well known Kaiser-Guttman rule suggests the number of factors should be equal to the number of eigenvalues exceeding 1. The scree-plot rule is based on a plot of eigenvalues ordered by diminishing size; the analyst examines the plot and looks for a notable drop in the size of the eigenvalues. Song and Belin (2008) also discussed the use of penalized likelihood criteria AIC and BIC, suggesting rules for interpreting these criteria with incomplete data.

2.4.3 Multiple imputation under the EFA model

The algorithm of Song and Belin (2004) for creating multiple imputations is a combination of Gibbs sampling and data augmentation. The algorithm proceeds as follows. At the $(t + 1)$th iteration, given the parameters at the tth step $(\beta^{(t)},\ \gamma^{(t)},\ \tau^{2(t)})$, draw

- missing items from $y_{iM} \mid y_{iO}, \beta^{(t)},\ \gamma^{(t)},\ \tau^{2(t)}$,
- factor scores from $z_i \mid y_{iO},\ y_{iM}^{(t)}, \beta^{(t)},\ \gamma^{(t)},\ \tau^{2(t)}$ independently for $i = 1, \ldots, n$,
- uniquenesses from $\tau_j^2 \mid Y_O,\ Y_M^{(t)},\ Z^{(t)}, \beta^{(t)},\ \gamma^{(t)}$,
- mean parameters from $\beta_j \mid Y_O,\ Y_M^{(t)},\ Z^{(t)},\ \gamma^{(t)},\ \tau^{2(t)}$, and

- factor loadings from $\gamma_j \mid Y_O,\ Y_M^{(t)},\ Z^{(t)},\ \beta^{(t)},\ \tau^{2(t)}$, independently for $j = 1, \ldots, r$.

In this notation, β_j, γ_j and τ_j^2 represent the jth element of β, the j column of γ and the j column of τ^2, respectively.

To avoid degenerate variance estimates, Song and Belin (2004) applied weakly informative prior distributions to the uniquenesses in τ^2. They also applied noninformative or weakly informative prior distributions to β and γ. In some cases, the algorithm was seen to converge slowly due to high correlations between elements of β and elements of γ, and a transformation of these parameters was suggested to speed convergence. Because multiple local modes are not uncommon in these exploratory factor models (Rubin & Thayer, 1982), they suggested running the MCMC procedure in multiple chains from overdispersed starting values, and they monitored convergence using diagnostic methods suggested by Gelman and Rubin (1992).

2.4.4 EM-type algorithms for the EFA model

The imputation procedure of Song and Belin (2004) is closly related to earlier published work on ML estimation for factor models. Rubin and Thayer (1982) developed an EM algorithm for the model

$$y_i \mid z_i \sim N_r(\beta^T x_i + \gamma^T z_i, \tau^2),$$

where z_i $(k \times 1)$ denotes a latent factor score vector normally distributed with mean zero and covariance matrix I_k, the $k \times k$ identity matrix. The algorithms of Rubin and Thayer (1982) included procedures for handling missing values in y_i. The slow convergence of these EM algorithms prompted additional work to speed

convergence. Liu and Rubin (1998) discussed the use of ECME, and Liu, Rubin and Wu (1998) proposed PX-EM algorithms.

The complete-data log-likelihood function for this model, which is based on the complete data (Y_O, Y_M, Z), can be written as

$$\begin{aligned} & l_A(\theta|Y_O, Y_M, Z) \\ & \propto -\frac{1}{2}\sum_{i=1}^{n} z_i {z_i}^T - \frac{n}{2}\sum_{j=1}^{r}\log(\tau_j^2) \\ & \quad -\frac{1}{2}\sum_{j=1}^{r}\tau_j^{-2}\sum_{i=1}^{n}\left(y_{ij} - \beta_j^T x_i - \gamma_j^T z_i\right)^2, \end{aligned}$$

and the actual log-likelihood based on Y_O can be written as

$$l_0(\theta) = -\frac{1}{2}\sum_{i=1}^{n}\log|\Phi_{iO}| - \frac{1}{2}\sum_{i=1}^{n}(y_{iO} - \beta_{iO})^T(\Phi_{iO})^{-1}(y_{iO} - \beta_{iO}),$$

where $Z = \{z_i, i = 1, \ldots, n\}$, where β_{iO}, γ_{iO} and τ_{iO}^2 denote the corresponding elements of β, the submatrix of γ and τ^2, respectively, for predicting y_{iO}, and where $\Phi_{iO} = \gamma_{iO}^T\gamma_{iO} + \tau_{iO}^2$.

An EM algorithm adapted from Liu and Rubin (1998) can be described as follows.

- *E-step*: Calculate the expected values of the sufficient statistics

$$\begin{aligned} S_{yy} &= \sum_{i=1} {y_i}^T y_i, \\ S_{Z^\star y} &= \sum_{i=1}^{n}\begin{pmatrix} 1 \\ z_i \end{pmatrix} y_i, \end{aligned}$$

$$S_{Z^\star Z^\star} = \sum_{i=1}^{n} \begin{pmatrix} 1 \\ z_i \end{pmatrix} \begin{pmatrix} 1 & z_i^T \end{pmatrix}.$$

- *M-step*: Given the current estimate of θ, replace the complete-data sufficient statistics with their expected values, and update the parameters with the complete-data maximum likelihood estimates. The computations required for the M-step are a straightforward application of ordinary least-squares regression, akin to regressing each items on the latent factors.

Liu, Rubin and Wu (1998) created a PX-EM algorithm for the exploratory factor model only. They embedded the EFA model into a larger model with an additional parameter Γ,

$$y_i \mid \theta_\star \sim N_r(\beta_\star, \gamma_\star^T \Gamma \gamma_\star + \tau_\star^2), \tag{2.12}$$

where $\theta_\star = (\beta_\star,\ \gamma_\star,\ \tau_\star^2,\ \Gamma)$. The expanded model reduces to the desired model when Γ is the identity. The additional parameter Γ is not identifiable from the observed data; including this parameter is merely a computational device to speed convergence. Setting $\theta = (\beta,\ \gamma,\ \tau^2) = (\beta_\star,\ \gamma_\star \text{Chol}(\Gamma),\ \tau_\star^2)$ recovers the original parameters, where Chol(Γ) denotes the Cholesky factor of Γ. Define $l_A(Y, Z|\theta_\star)$ to be the complete-data log-likelihood of $\theta_\star$ under model (2.12) based on (Y_O, Y_M, Z) and let $E(l_A(Y, Z|\theta_\star))$ be the expected value of $l_A(Y, Z|\theta_\star)$ with respect to the distributions of $Y_M|Y_O, \theta$ and $Z|Y_O, \theta$. One iteration of the PX-EM algorithm can be described as follows.

Step 1. Update $\theta_\star$ by applying the EM algorithm to the expanded model:

PX-E step: Compute $\text{E}(l_A(Y, Z|\theta_\star^{(t)}))$.

PX-M step: Update $\theta_\star$ with $\theta_\star^{(t+1)} = \arg\max_{\theta^\star} \mathrm{E}(l_A(Y|\theta_\star^{(t)}))$.

Step 2. Reduce $\theta_\star$ to the original parameter θ by the reduction formulas

$$\beta^{(t+1)} = \beta_\star^{(t+1)},$$

$$\gamma^{(t+1)} = \gamma_\star^{(t+1)}\mathrm{Chol}(\Gamma^{(t+1)}),$$

$$\tau^{2(t+1)} = \tau_\star^{2(t+1)}.$$

2.4.5 Limitations of existing work

The literature that we have cited is the starting point for our research. To our knowledge, no software package has yet been made available for multiple imputation under EFA models; analysts have no way to apply them, and their properties are not well understood. And to our knowledge, no methods have yet been published on the related problem of multiple imputation under CFA models. The restrictions that are introduced when moving from EFA to CFA are a necessary step in creating procedures appropriate for multi-themed questionnaires. Models for multi-theme questionnaires should to reflect the fact that items within different themes are measuring different sets of underlying characteristics. It is reasonable to think that relationships between items addressing different themes are partially or fully explainable by relationships among the latent characteristics they are measuring. In the chapters ahead, we extend these methods to CFA models in a variety of ways.

Chapter 3

Parameter Estimation and Imputation of Missing Values under Confirmatory Factor Models

3.1 Initial exploration

In the previous chapter, we reviewed missing-data methods for multivariate data under various assumptions about covariances among items. We now return to one of our motivating examples—the *Feelings Scale* from Add Health—to gather empirical evidence on how the items within this questionnaire theme are interrelated.

Figure 3.1 shows a scree plot for these 19 items. A scree plot displays the eigenvalues of the correlation matrix arranged in descending order. These correlations were estimated by applying the EM algorithm for the unstructured normal model described in Section 2.1. The scree plot shows a steep drop between the first and second eigenvalues, indicating that a large portion of the relationships is accounted for by one factor. This finding is consistent with the purpose of the *Feelings Scale*,

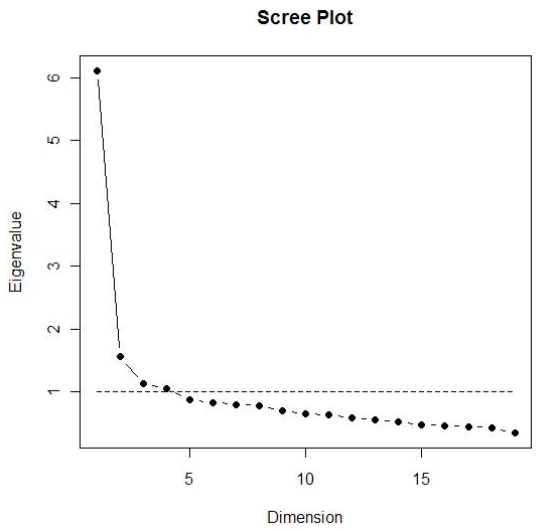

Figure 3.1. Scree plot of $\hat{\Sigma}$, the MLE of Σ

which was designed to measure a single construct (psychological distress). This finding is also consistent with how the data are commonly used; analysts often aggregate the items by summing or averaging them into a composite score. But this scree plot also suggests that a one factor model may not exactly fit the data. The well known Kaiser rule suggests that a four-factor model ought to be considered, because four of the eigenvalues are greater than one.

Applying the EM algorithm for the EFA models described in the previous chapter (Rubin & Thayer, 1982), we examined results from models with varying numbers of factors. An EFA model is identified only up to an arbitrary rotation of the factors. Using varimax rotations, we observed that the four items that were reverse-coded (lower values indicate higher distress) have large loadings for the second factor. Loadings for the third and fourth factors were difficult to interpret. Because each of these models is a special case of the unstructured normal model, we may compare the fit of each model to the unstructured one by a standard likelihood ratio test. In this example, due to the large sample size, the one, two, three

and four-factor models can all be rejected in favor of the unstructured alternative. However, the one-factor model and perhaps the two-factor model have considerable intuitive appeal. This finding is not atypical for data from a multi-themed questionnaires. Items from one thematic section are usually intended to measure a small number of constructs, but the correspondence between observed data and a low-dimensional factor structure may be imperfect. In this chapter, we will proceed as if the relationships among items can indeed be described by factor model with a known number of items; in the next chapter, we will introduce methods that formally account for this lack of fit.

If we were to build an imputation model that integrates items from the *Feelings Scale* with items for additional themes, it is reasonable to think that the relationships between the *Feelings Scale* items and items for other themes would be mediated by the factors describing the *Feelings Scale*, especially the first one. This would be consistent with post-imputation analyses in which an aggregate measure of psychological distress is related to measures for other themes by a correlation analysis or a regression model. For an imputation procedure whose items span multiple themes, this notion could be formalized by moving from an EFA to a CFA model that assumed certain factor loadings are zero. Consider a *one-factor per theme model* in which any one theme of a questionnaire could be well described by one underlying factor.

- Any item for theme j will have a nonzero factor loading on the factor describing theme j.
- Any item for theme j will have a zero loading on the factor describing theme k when $j \neq k$.

In the notation for CFA models used in the last chapter, each column of the factor

loading matrix γ in a one-factor per theme model will have only one nonzero element, and the only nonzero element of column j of γ will occur in row c_j which is known in advance. With this model, it makes sense to allow the factors to be non-orthogonal ($R \neq I$), because post-imputation analyses will often involve hypothesized relationships between aggregate measures from different themes.

In the remaining sections of this chapter, we present new algorithms for parameter estimation and imputation for CFA models with missing items. For simplicity, we describe the methods for the one-factor per theme model, but the extension to models where some themes have two or more factors will be immediate.

3.2 A new PX-EM algorithm for the CFA model

EM algorithms for parameter estimation in factor models were previously described by Rubin and Thayer (1982). In that article, they considered a variety of EFA and CFA models with and without missing items. When reproducing their algorithms and results, we stumbled across an apparently little known fact about the multivariate normal distribution. Suppose that the rows of a data matrix are independently sampled from a multivariate normal distribution with means fixed at zero and variances fixed at one. The ML estimate of the covariance (and correlation) matrix is *not* the sample correlation matrix. With two variables, this model has only a single parameter, the correlation between the two items. In that bivariate case, the sample correlation coefficient does not maximize the one-parameter loglikelihood function. In retrospect, this result is not surprising; the multivariate normal model with constrained variances is not a regular exponential family, so the ML estimates should not correspond to the sample correlations. Closed-form expressions for the ML estimators are not easily found even in the bivariate

case, but numerical estimates can be calculated by iterative procedures such as Newton-Raphson and Fisher scoring.

When Rubin and Thayer (1982) described EM algorithms for CFA models with non-orthogonal factors ($R \neq I$), they erroneously assumed that the part of the Q-function (the expected complete-data loglikelihood) corresponding to R would be maximized by a sample correlation matrix. As a result, the algorithm they described for models with $R \neq I$ does not yield a true ML estimate. Results for the data examples presented in that article were not affected by this mistake, because all of the models applied in the examples assumed that $R = I$. A later article by Liu and Rubin (1998) about EM-type algorithms for estimation in factor models was not affected by this issue either, because throughout that article they assumed that $R = I$. Because non-orthogonal factors are a crucial part of our CFA modeling, we needed to develop a reliable procedure for computing ML estimates when $R \neq I$. Moreover, we did not want to use a conventional EM algorithm that treats the vector of factor scores z_i as missing data, because the part of the M-step pertaining to R would itself require an iterative solution.

This issue can be sidestepped by the principle of parameter extension used in PX-EM. We may enlarge the parameter space by supposing that the factor scores are distributed as $z_i \sim N(0, R_\star)$ with $R_\star = D_\star^{1/2} R D_\star^{1/2}$, where $D_\star$ is a diagonal matrix of variances. The extra parameters in $D_\star$ are not estimable from the observed items. But expanding the model for the complete data in this manner—and here, the term "complete data" includes the latent variables z_i—simplifies the maximization of the Q-function, because the ML estimate for the newly unstructured covariance matrix $R_\star$ may be calculated in the usual way. Another benefit of this parameter extension is that the resulting PX-EM algorithm will converge faster than conventional EM. We have not quantified or illustrated the improve-

ment in the rate of convergence, because we have not implemented a conventional EM algorithm for this problem. But theoretical results regarding the convergence of PX-EM, and the empirical demonstrations by Liu and Rubin (1998) in related parameter-expansion problems, suggest that the computational savings over EM are substantial.

The expanded CFA model for the new PX-EM algorithm is

$$\begin{aligned} y_i \mid z_i, \theta_\star &\sim N_r(\beta_\star^T x_i + \gamma_\star^T z_i, \tau_\star^2), \\ z_i &\sim N_k(0, R_\star), \end{aligned} \tag{3.1}$$

where $R_\star = D_\star^{1/2} R D_\star^{1/2}$ is unstructured. Integrating out the latent variable z_i yields

$$y_i \mid \theta_\star \sim N_r(\beta_\star^T x_i, \gamma_\star^T R_\star \gamma_\star + \tau_\star^2), \tag{3.2}$$

and the parameters of the expanded model are $\theta_\star = (\beta_\star,\, \gamma_\star,\, R_\star,\, \tau_\star^2)$.

This expanded model satisfies two conditions necessary for PX-EM described by Liu et al. (1998). First, the parameters of original model can be obtained from the expanded parameters by the following reduction function,

$$\theta = (\beta, \gamma, R, \tau^2) = (\beta_\star, D_\star^{\frac{1}{2}} \gamma_\star, D_\star^{-\frac{1}{2}} R_\star D_\star^{-\frac{1}{2}}, \tau_\star^2).$$

Second, $\theta_\star$ reduces to θ when $D_\star$ is equal to the identity matrix.

When a factor model is applied to incomplete multivariate data, it is not necessary to treat the missing items as "missing data" when formulating the Q-function. The reason for this is quite intuitive: if the factor scores z_i were actually observed, the measurement parameters (i.e., the factor loadings and uniquenesses) could be estimated by a sequence of independent univariate linear regression models. Miss-

ing items in the data matrix would become missing responses in these univariate regression models, and ML estimates for any one of these regressions could be computed by eliminating the cases with missing values for that response (Little & Rubin, 2002). In the E-step of our PX-EM algorithm, we treat the factor scores z_i as missing, but the missing items Y_M are analytically removed from the likelihood by integration. (In that respect, this PX-EM algorithm could also by regarded as PX-ECME, a parameter-extended version of the ECME algorithm.) Integrating Y_M out of the likelihood should also increase the rate of convergence (Liu & Rubin, 1994), although we have not attempted to quantify this increase, because we believe that iin most cases it will be slight.

Let y_{ij} denote the jth element of y_i. Let β_{iO}, γ_{iO} and τ^2_{iO} denote the columns of β, the columns of γ and the elements of τ^2 for predicting the observed elements, respectively. Define $\gamma_{\star j}$ and $\beta_{\star j}$ as the jth columns of $\gamma_\star$ and $\beta_\star$ respectively, and let $\mathbf{1}_{ij}$ equal to one if y_{ij} is observed and zero otherwise. The "complete-data" loglikelihood based on Y_O and $Z = (z_1, \ldots, z_n)$ becomes

$$\begin{aligned} l_A(\theta_\star | Y_O, Z) \propto & -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{r} \log \tau^2_{\star j}\mathbf{1}_{ij} \\ & -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{r} \frac{1}{\tau^2_{\star j}}(y_{ij} - \beta^T_{\star j}x_i - \gamma^T_{\star j}z_i)^2\mathbf{1}_{ij} \\ & -\frac{n}{2}\log|R_\star| - \frac{1}{2}\sum_{i=1}^{n} z_i^T R_\star^{-1} z_i. \end{aligned}$$

Our PX-EM algorithm can be described as follows. Given the estimated parameters at iteration t, the estimates at iteration $t+1$ are computed by these steps.

- E-step: Compute the expectation of l_A with respect to the distribution

$$z_i | y_{iO}, \theta^{(t)}_\star \sim N(\mu^{(t)}_i, \Sigma^{(t)}_i)$$

for $i = 1, \ldots, n$, where

$$\begin{aligned}
\mu_i^{(t)} &= R_\star^{(t)} \gamma_{\star_{iO}}^{(t)} (\gamma_{\star_{iO}}^{(t)^T} R_\star^{(t)} \gamma_{\star_{iO}}^{(t)} + \tau_{\star iO}^{2^{(t)}})^{-1} (y_{iO} - \beta_{iO}^{(t)^T} x_i), \\
\Sigma_i^{(t)} &= R_\star^{(t)} - R_\star^{(t)} \gamma_{\star_{iO}}^{(t)} (\gamma_{\star_{iO}}^{(t)^T} R_\star^{(t)} \gamma_{\star_{iO}}^{(t)} + \tau_{\star iO}^{2^{(t)}})^{-1} \gamma_{\star_{iO}}^{(t)^T} R_\star^{(t)},
\end{aligned}$$

and where $\gamma_{\star iO}$ and $\tau_{\star iO}^2$ denote the columns of $\gamma_\star$ and the elements of $\tau_\star^2$ for predicting y_{iO}, respectively. For notational simplicity, define

$$E_{ZZ,i}^{(t)} = \mathrm{E}(z_i {z_i}^T | y_{iO}, \theta_\star^{(t)}) = \mu_i^{(t)} \mu_i^{(t)^T} + \Sigma_i^{(t)}.$$

- M-step: Compute the ML estimate of $\theta_\star$ by maximizing the expanded log-likelihood El_A found in the E-step. The maximization is accomplished by

$$\begin{aligned}
R_\star^{(t+1)} &= \frac{1}{n} \sum_{i=1}^{n} \mathrm{E}_{ZZ,i}^{(t)}, \\
\tau_{\star j}^{2^{(t+1)}} &= \frac{\sum_{i=1}^{n} A_{ij}^{(t)} \mathbf{1}_{ij}}{\sum_{i=1}^{n} \mathbf{1}_{ij}}, \\
\begin{bmatrix} \beta_{\star j}^{(t+1)} \\ \gamma_{\star c_j, j}^{(t+1)} \end{bmatrix} &= [\sum_{i=1}^{n} \mathrm{E}(x_{\star,i,j} x_{\star,i,j}^T) \mathbf{1}_{ij}]^{-1} \times [\sum_{i=1}^{n} \mathrm{E}(x_{\star,i,j}) y_{ij} \mathbf{1}_{ij}], \\
A_{ij}^{(t)} &= E \left(y_{ij} - \begin{bmatrix} \beta_{\star j}^{(t+1)} \\ \gamma_{\star c_j, j}^{(t+1)} \end{bmatrix}^T x_{\star,i,j} \mid Y_O, \theta_\star^{(t)} \right)^2, \\
x_{\star,i,j} &= \begin{bmatrix} x_i \\ z_{i,c_j} \end{bmatrix}, \mathrm{E}(x_{\star,i,j}) = \begin{bmatrix} x_i \\ \mu_{i,c_j} \end{bmatrix}, \\
\mathrm{E}(x_{\star,i} x_{\star,i,j}^T) &= \begin{bmatrix} x_i x_i^T & x_i \mu_{i,c_j}^{(t)} \\ \mu_{i,c_j}^{(t)} x_i^T & \mathrm{E}_{ZZ,i,c_j,c_j}^{(t)} \end{bmatrix},
\end{aligned}$$

where z_{i,c_j} and μ_{i,c_j} denote the c_jth elements of z_i and μ_i, respectively, and $\mathrm{E}_{ZZ,i,c_j,c_j}$ denotes the element of $\mathrm{E}_{ZZ,i}$ in the c_jth row and c_jth column.

- Parameter reduction: Reduce the expanded parameters to the original parameters,

$$\begin{aligned}
\beta^{(t+1)} &= \beta_{\star}^{(t+1)}, \\
\gamma_{c_j,j} &= D_{\star,c_j,c_j}^{\frac{1}{2}(t+1)} \gamma_{\star c_j,j}, \\
R^{(t+1)} &= D_{\star}^{-\frac{1}{2}(t+1)} R_{\star}^{(t+1)} D_{\star}^{-\frac{1}{2}(t+1)}, \\
\tau^{2(t+1)} &= \tau_{\star}^{2(t+1)}.
\end{aligned}$$

We implemented this PX-EM algorithm and applied it to real and simulated data examples. We have verified that it converges to a local maximum by numerically perturbing each free parameter from the solution and have seen in each case that the observed-data loglikelihood drops. EM-type algorithms for EFA and CFA models are known to be very stable, but they may converge to local maxima of the observed-data loglikelihood. As noted by Rubin and Thayer (1982), multiple random starting values may be necessary to give us confidence that we have indeed found an ML estimate. Convergence to a stationary value does not imply that the solution is locally identified. To check for local identification, we compute the Hessian (second derivative) matrix for the loglikelihood at the stationary value to verify that it is nonsingular. (Expressions for these derivatives are given in Appendix A). If the starting values are poor, iterations of PX-EM may progress toward a boundary (e.g., a solution where some uniquenesses are zero) where the loglikelihood is not concave. If the second derivative matrix is not negative definite, we rerun the PX-EM algorithm from alternative starting values

until it arrives at a reasonable solution.

3.3 A PX-DA algorithm for multiple imputation

3.3.1 Prior distributions

For multiple imputation under the CFA model, we will again work with an expanded parameter set. We now denote the covariance matrix for z_i by $W = D^{\frac{1}{2}} R D^{\frac{1}{2}}$, where R is a correlation matrix and D is a diagonal matrix containing variances. The same parameter extension for multiple imputation was used by Boscardin and Zhang (2004), but in a different modeling context. This extension is helpful for the following reason: Specifying a sensible prior distribution for a correlation matrix can be difficult (Barnard et al., 2000), but convenient priors for a covariance matrix are readily available. The Jacobian of the transformation from W to (R, D) is

$$J_{W \to R,D} = (|D|)^{\frac{k-1}{2}}. \tag{3.3}$$

To create an algorithm for multiple imputation, we apply the following prior distributions to the model parameters $\theta = (\beta, \gamma, R, \tau^2)$.

- For the regression coefficients β, we use an improper uniform density over $\mathcal{R}^{pr}$.
- For the factor loadings in γ, we apply improper independent uniform densities to all the nonzero factor loadings.
- For the factor correlation matrix R and the extra variance parameters D, we apply a standard noninformative prior to W, $p(W) \propto |W|^{-\frac{k+1}{2}}$, which leads

to

$$\begin{aligned} p(R,\ D) &\propto p(W) J_{W \to R,D} \\ &\propto |D|^{-1} |R|^{-\frac{k+1}{2}}. \end{aligned}$$

- For the uniquenesses in τ^2, we use diffuse conjugate inverse-gamma priors, $\tau_j^2 \sim \text{Inv-Gamma}(\alpha_{\tau^2}, \beta_{\tau^2})$ for $j = 1, \ldots, r$, where α_{τ^2} and β_{τ^2} are both small values, e.g., 0.002.

3.3.2 Proposal densities

Our algorithm for multiple imputation can be viewed as a parameter-extended version of data augmentation (PX-DA) (Liu & Wu, 1999; Meng & van Dyk, 1999). It can also be viewed as a Gibbs sampler that partitions the unknown quantities (parameters and latent factors) into convenient groups and draws each group from its conditional posterior distribution given the other groups. Drawing from some of these conditional distributions is not tractable, so we replace them with Metropolis-Hastings steps. We will draw the uniquenesses from a proposal density, and then accept the drawn value with a probability derived from a Metropolis-Hastings density ratio. A similar technique is applied to the correlation matrix R. We draw a covariance matrix W from its inverse Wishart posterior distribution, translate it back to a correlation matrix through a reduction function, and accept the simulated correlation matrix based on a Metropolis-Hastings ratio.

First, we apply the following notation.

- Let l_0 denote the actual observed-data loglikelihood function of the CFA model.

- Let $(\beta_0, \gamma_0, R_0, \tau_0^2)$ denote the local maximizer of l_0 derived from the PX-EM algorithm previously discussed.

Our proposal distribution for τ^2 is a multivariate t distribution with α degrees of freedom for $\log(\tau^2)$,

$$q_{\tau^2}(\tau^{2\star}|\tau^2) = q_{mvt}(\log \tau^{2\star}|\log \tau^2) \times \frac{1}{\prod_{j=1}^{r} \tau_j^{2\star}}.$$

We center this proposal density at the current value of $\log(\tau^2)$, choose $\alpha = 4$ to make it a heavy-tailed distribution (Gelman *et.al.*, 1997) and set the scale matrix to be $S_{\log\tau^2} = c \times \frac{\alpha+2r}{\alpha} \times (-\frac{\partial^2 l_0}{\partial \log\tau^2 \partial \log\tau^{2T}})^{-1}|_{\tau^2=\tau_0^2}$, where c is a constant and

$$\frac{\partial^2 l_0}{\partial \log\tau^2 \partial \log\tau^{2T}} = \mathrm{diag}(\tau^2)\frac{\partial^2 l_0}{\partial \tau^2 \partial \tau^{2T}} + \mathrm{diag}(\frac{\partial l_0}{\partial \tau^2}).$$

The proposal density for $(R,\ D)$ is a jumping kernel $q_R(R^\star, D^\star|R, D)$,

$$q_R(R^\star, D^\star|R, D) \propto \mathrm{Wishart}(W^\star|d_w, \frac{W^{(t)}}{d_w}) \times J_{W^\star \to R^\star, D^\star}.$$

where $\mathrm{Wishart}(W^\star|d_w, \frac{W}{d_w})$ denotes a Wishart distribution with degrees of freedom parameter d_w and scale matrix equal to $\frac{1}{d_w}W$, centered at the current value of W. Boscardin and Zhang (2004) chose the scale matrix to be the current value of W, i.e., $W^\star$ is drawn from a Wishart distribution centered at $d_w W$, which according to our simulation, may lead to very slow convergence rates even with a small value of d_w.

3.3.3 The PX-DA algorithm for CFA

The posterior density of the expanded parameters given the complete data Y and factor scores Z is

$$
\begin{aligned}
& P(\beta, \gamma, R, D, \tau^2 | Y, Z) \\
& \propto (\prod_{j=1}^{r} \tau_j^2)^{-\frac{n}{2}} \exp(-\sum_{j=1}^{r} \frac{\sum_{i=1}^{n} (y_{ij} - \beta_j^T x_i - \gamma_j^T z_i)^2}{2\tau_j^2}) \\
& \quad \times |R|^{-\frac{n}{2}} \exp(-\frac{1}{2} \sum_{i=1}^{n} z_i^T R^{-1} z_i)\ p(\beta)\ p(\gamma)\ p(R, D)\ p(\tau^2).
\end{aligned}
$$

Simulating draws from this posterior distribution forms the posterior or P-step of our data augmentation procedure. Because this joint distribution is difficult to handle, we partition the parameters and simulate their conditional distributions directly or indirectly by Metropolis-Hastings steps. The P-step is accompanied by a imputation or I-steps which simulate the missing elements of the Y matrix and the unknown factor scores Z given assumed values for the parameters. Each cycle of the algorithm proceeds as follows.

- Draw the missing elements of Y. For this step, we group the rows of Y according to their missingness patterns as described in Section 2.1. For missingness pattern s, we partition Σ into submatrices Σ_{sOO},Σ_{sMO},Σ_{sOM} and Σ_{sOO} corresponding to the observed and missing variables. We then simulate the missing elements for the rows within each missingness pattern from $y_{iM} \mid y_{iO}, \Sigma, \gamma, R, D, \tau^2 \sim \mathrm{N}(\mu_i,\ \Sigma_i)$ independently for $i = 1,\ldots,$ n, where

$$\mu_i = \beta_{iM}^T x_i + \Sigma_{sMO} \Sigma_{sOO}^{-1} (y_{iO} - \beta_{iO}^T x_i),$$

$$\Sigma_i = \Sigma_{sMM} - \Sigma_{sMO}\Sigma_{sOO}^{-1}\Sigma_{sOM}.$$

- Draw the regression coefficients and factor loadings. These are distributed as

$$\begin{bmatrix} \beta_j \\ \gamma_{c_j,j} \end{bmatrix} \mid Y_M, Y_O, Z, \Sigma, \gamma, R, D, \tau^2 \sim N(\mu_{\beta_{\star j}}, \Sigma_{\beta_{\star j}})$$

  independently for $j = 1, \ldots, r$, where

$$\mu_{\beta_{\star j}} = (\sum_{i=1}^{n} x_{\star,i,j} x_{\star,i,j}^T)^{-1} (\sum_{i=1}^{n} y_{ij} x_{\star,i,j}),$$
$$\Sigma_{\beta_{\star j}} = \tau_j^2 (\sum_{i=1}^{n} x_{\star,i,j} x_{\star,i,j}^T)^{-1}.$$

- Draw the factor scores. These are distributed as

$$z_i \mid Y_M, Y_O, \beta, \gamma, R, D, \tau^2 \sim N(\mu_{z_i}, \Sigma_{z_i})$$

  for $i = 1, \ldots, n$, where

$$\mu_{z_i} = (\sum_{j=1}^{r} \gamma_j \tau_j^{-2} \gamma_j^T + R)^{-1} \sum_{j=1}^{r} \tau_j^{-2} \gamma_j (y_{ij} - \beta_j^T x_i),$$
$$\Sigma_{z_i} = (\sum_{j=1}^{r} \gamma_j \tau_j^{-2} \gamma_j^T + R)^{-1}.$$

- Draw the uniquenesses $P(\tau_j^2 \mid Y_M, Y_O, \beta, \gamma)$, which are distributed as

$$\text{Inv-Gamma}(0.5n + \alpha_{\tau^2}, [0.5 \sum_{i=1}^{n} (y_{ij} - \beta_j^T x_i - \gamma_j^T z_i)^2] + \beta_{\tau^2})$$

  independently for $j = 1, \ldots, r$.

- Draw the factor correlation matrix and the extra variance parameters. Because direct simulation of (R, D) from $P(R, D \mid Y_M, Y_O, Z, \beta, \gamma, \tau^2)$ is not straightforward, we generate a candidate $(R^\star, D^\star)$ from the inverse Wishart proposal $q_R(R^\star, D^\star|R, D)$. We take $(R,D) = (R^\star,D^\star)$ with probability $\min(1, \alpha_R)$, where

$$\alpha_R = \frac{p(R^\star, D^\star)q_R(R,\ D|R^\star, D^\star)}{p(R,\ D)q_R(R^\star, D^\star|R,\ D)} \times \frac{G_R(R^\star, D^\star)}{G_R(R, D)},$$

  and where

$$G_R(R, D) = |D|^{-1}|R|^{-\frac{n+k+1}{2}} \exp(-\frac{1}{2}\sum_{i=1}^{n} z_i^T R^{-1} z_i).$$

3.3.4 The PX-DA algorithm for EFA

In addition to the PX-DA algorithm just described, we desire a method for creating multiple imputations under an EFA model, so that we can contrast our results with those from EFA. The algorithm of Song and Belin (2004) was designed for a CFA model without covariates. Here we extend their method to EFA with covariates.

In exploratory factor modeling, it is customary to assume that the latent factors are uncorrelated ($R = I$). In EFA, the factor loadings are identified only up to an orthogonal rotation. For the applications that we envision, the loadings themselves are not of interest; this model is being used only as a device to impute the missing questionnaire items. Setting $R = I$, and eliminating *a priori* zeros from the factor loadings, the posterior density of the parameters given the complete data and factor scores is

$$P(\beta, \gamma, \tau^2|Y, Z)$$

$$\propto (\prod_{j=1}^{r} \tau_j^2)^{-\frac{n}{2}} \exp(-\sum_{j=1}^{r} \frac{\sum_{i=1}^{n} (y_{ij} - \beta_j^T x_i - \gamma_j^T z_i)^2}{2\tau_j^2})$$
$$\times \exp(-\frac{1}{2} \sum_{i=1}^{n} z_i^T z_i)\ p(\beta)\ p(\gamma)\ p(\tau^2).$$

For the model parameters $\theta = (\beta, \gamma, \tau^2)$, we specify the same priors as those described in Section 3.3.1. The algorithm is implemented as follows.

- Group the rows of Y according to their missingness patterns. For missingness pattern s, partition Σ into submatrices Σ_{sOO},Σ_{sMO},Σ_{sOM} and Σ_{sOO} corresponding to the observed and missing variables, and simulate the missing elements for the rows within each missingness pattern from $y_{iM} \mid y_{iO}, \Sigma, \gamma, \tau^2 \sim \mathrm{N}(\mu_i,\ \Sigma_i)$ independently for $i = 1, \ldots, n$, where

$$\mu_i = \beta_{iM}^T x_i + \Sigma_{sMO} \Sigma_{sOO}^{-1} (y_{iO} - \beta_{iO}^T x_i),$$
$$\Sigma_i = \Sigma_{sMM} - \Sigma_{sMO} \Sigma_{sOO}^{-1} \Sigma_{sOM}.$$

(2) Draw the factor loadings and regression coefficients from

$$\begin{bmatrix} \beta_j \\ \gamma_j \end{bmatrix} \mid Y_M, Y_O, Z, \Sigma, \gamma, \tau^2 \ \sim\ N(\mu_{\beta_{\star j}}, \Sigma_{\beta_{\star j}})$$

independently for $j = 1, \ldots, r$, where

$$\mu_{\beta_{\star j}} = (\sum_{i=1}^{n} x_{\star,i} x_{\star,i}^T)^{-1} (\sum_{i=1}^{n} y_{ij} x_{\star,i}),$$
$$\Sigma_{\beta_{\star j}} = \tau_j^2 (\sum_{i=1}^{n} x_{\star,i} x_{\star,i}^T)^{-1},$$

and where

$$x_{\star,i} = \begin{bmatrix} x_i \\ z_i \end{bmatrix}.$$

- Draw the factor scores from

$$z_i \mid Y_M, Y_O, \beta, \gamma, \tau^2 \sim N(\mu_{z_i}, \Sigma_{z_i})$$

for $i = 1, \ldots, n$, where

$$\mu_{z_i} = (\sum_{j=1}^{r} \gamma_j \tau_j^{-2} \gamma_j^T + I_k)^{-1} \sum_{j=1}^{r} \tau_j^{-2} \gamma_j (y_{ij} - \beta_j^T x_i),$$
$$\Sigma_{z_i} = (\sum_{j=1}^{r} \gamma_j \tau_j^{-2} \gamma_j^T + I_k)^{-1}.$$

- Draw the uniquenesses $P(\tau_j^2 | Y_M, Y_O, \beta, \gamma)$ from

$$\text{Inv-Gamma}(0.5n + \alpha_{\tau^2}, [0.5 \sum_{i=1}^{n} (y_{ij} - \beta_j^T x_i - \gamma_j^T z_i)^2] + \beta_{\tau^2})$$

independently for $j = 1, \ldots, r$.

As noted by Song and Belin (2004), if we apply familiar MCMC convergence diagnostics to the elements of the factor loadings matrix γ — time-series plots, autocorrelation functions, etc. — the procedure will appear to never converge because the elements of γ are not identified. But the elements of $\gamma^T \gamma$ are identified, so in practice we apply convergence diagnostics to the upper triangle of $\gamma^T \gamma$.

Chapter 4

A Softly Constrained CFA Model

4.1 Formulating the model

In the previous chapter, we constrained the multivariate normal imputation model for a multi-themed questionnnaire by supposing that the items within each theme were conditionally independent given a small number of latent factors. Items from different themes were assumed to be related only through their respective factors. Although those models are intuitive appealing, real questionnaire data are likely to depart from these assumptions. In our example from Add Health, items in the *Feelings Scale* were designed to measure a single construct (emotional distress), and and one would hope that a single factor could describe these items well. However, the one factor-model was strongly rejected by a goodness-of-fit test. Two- and three-factor models were not adequate, and even a four-factor solution did not fit. Belin and Song (2004) have demonstrated that understating the number of factors in an imputation model may lead to bias in post-imputation analyses. But in this example, additional factors beyond the first (or perhaps the second) have no theoretical justification. When extra factors are applied solely to accommodate

lack of fit, the intuitive appeal of the model and its original rationale are lost.

In this chapter, we develop strategies for imputation under a CFA models with small numbers of factors when the posited factor structure does not fit. Using an idea presented by Boscardin and Zhang (2004), we relax the constraints on the covariances by introducing an additional random component that allows the covariance matrix to deviate from the ideal form. That is, we apply an informative prior distribution to the the actual covariance matrix for the questionnaire items that is centered at a CFA structure, and we introduce a dispersion parameter ν that governs the lack of fit. The model is

$$\begin{aligned} y_i \mid x_i \ &\sim N(\beta^T x_i, \Sigma) \quad \text{for} \quad i = 1, \ldots, n \\ \Sigma^{-1} \mid \theta \ &\sim \text{Wishart}(\nu, \ \nu\,\Omega(\gamma, R, \tau^2)^{-1}), \end{aligned} \tag{4.1}$$

where $\Omega(\gamma, R, \tau^2) = \gamma^T R \gamma + \tau^2$ denotes the covariance matrix under a CFA model. In one-factor per theme model, any column j of γ has only one nonzero element in row c_j, and

$$\gamma_{ij} = 0 \quad \text{if} \quad i \neq c_j, \quad \text{for} \quad j = 1, \ldots, r.$$

The number of degrees of freedom, ν, describes the fidelity of our model (4.1) to the one-factor per theme model. As $\nu \to \infty$ it reduces to CFA, and as $\nu \to 0$ it becomes an unstructured covariance model.

4.2 Prior distributions

The prior distributions imposed on $(\beta, \gamma, \ R, \ \tau^2)$ are identical to those we described in Section (3.3.1). We apply an improper uniform density on the matrix of regression coefficients β and improper uniform densities on the nonzero loadings

in γ. To avoid placing an awkward prior on the correlation matrix R, we expand it to a covariance matrix $W = D^{\frac{1}{2}}RD^{1/2}$ and apply a standard Jeffreys prior to W. The uniquenesses in τ^2 are described by diffuse inverse-gamma distributions.

The important but difficult question now is how to handle the lack-of-fit parameter ν. Boscardin and Zhang (2004) suggested two different approaches. First, we may fix this parameter at a value that is chosen *a priori* or estimated from the data. Second, we may treat ν in a fully Bayesian fashion, apply a prior distribution, and sample values of ν from its posterior distribution. For this dissertation, we use the second approach. We apply a lognormal prior, assuming that $\log \nu$ is normally distributed with mean μ_ν and variance σ_ν^2. In the simulations to be presented in Chapter 5, we select different values for the hyperparameters μ_ν and σ_ν^2 to assess the sensitivity of our results to changes in the prior. Opting to regard ν as random rather than fixed does increase the complexity of the imputation algorithm, and the practical implications of this choice on the computational and inferential performance of these procedures is still largely unknown. Gaining a better understanding of strategies for handling ν is one important topic of ongoing research which we discuss in Chapter 6.

4.3 Proposal densities

Our algorithm for multiple imputation is an expanded version of the Gibbs sampler with embedded Metropolis-Hastings steps that we described in the last chapter. As before, we will need proposal densities for the Metropolis-Hastings steps that mimic the shape of the posterior distributions for (R, D), τ^2 and γ in local areas of the parameter space. For (R, D) and τ^2, we apply the same proposal densities as those described in Section 3.3.2, a Wishart kernel for $W = D^{\frac{1}{2}}RD^{\frac{1}{2}}$ and a multivariate t-

distribution for $\log \tau^2$. For γ, we apply multivariate t distribution with parameters based on the results from our PX-EM procedure defined in Chapter 3. That is, we center our proposal density at the current value of γ, and we find a scale matrix by equating the final value of $\frac{\partial^2 l_0}{\partial\gamma\partial\gamma^T}$ from scoring to the second derivative of the logarithm of the t density. The scale matrix is $S_\gamma = c \times \frac{\alpha+2r}{\alpha}(-\frac{\partial^2 l_0}{\partial\gamma\partial\gamma^T})^{-1} \mid_{\gamma=\gamma_0}$, where γ_0 denotes the local maximizer of l_0, the observed-data loglikelihood of the one-factor per theme model derived from the PX-EM algorithm as described in Section 3.3.2, and c is a tuning parameter. When ν is treated as a random parameter, we use a log-normal proposal density to ν centered at the current value with variance $\sigma^2_{p,\nu}$ which is selected and tuned by trial and error to achieve a reasonable acceptance rate.

4.4 PX-DA procedure for multiple imputation

The complete-data posterior density of $(\Sigma, \beta, \gamma, R, D, \tau^2, \nu)$ is equal to the product of the likelihood in equation (4.1) and the prior densities given in equation (4.1) and Section (4.2),

$$
\begin{aligned}
& P(\Sigma, \beta, \gamma, R, D, \tau^2, \nu \mid Y) \qquad (4.2)\\
\propto\ & (2^{\nu r/2}\ \Gamma_r(\frac{\nu}{2}))^{-1}\ |\Sigma|^{\ -n/2} \exp(-\frac{1}{2}\sum_{i=1}^{n}(y_i - \beta^T x_i)\Sigma^{-1}(y_i - \beta^T x_i)^T)\\
& \times\ \mid \nu\Omega(\gamma, R, \tau^2) \mid^{\ \nu/2}\ \ |\Sigma|^{\ -(\nu+r+1)/2}\ \exp(\mathrm{tr}((-\frac{1}{2}\Sigma^{-1}\nu\Omega(\gamma, R, \tau^2)))\\
& \times p(R, D)\ p(\gamma)\ p(\tau^2)\ p(\nu)\ p(\beta).
\end{aligned}
$$

The actual covariance matrix Σ plays a similar role to that of a random ef-

fect in a mixed-effects regression model. Because the inverted Wishart density is conjugate to a multivariate normal likelihood, we can analytically integrate Σ out from this expression, which gives

$$
\begin{aligned}
& P(\beta,\ \gamma, R, D, \tau^2, \nu \mid Y) \qquad (4.3)\\
\propto\ & \nu^{\frac{\nu}{2}} (2^{\frac{\nu r}{2}}\ \Gamma_r(\frac{\nu}{2}))^{-1} 2^{\frac{(n+\nu)r}{2}}\ \Gamma_r(\frac{n+\nu}{2})\\
& \times\ |\Omega(\gamma, R, \tau^2)|^{\ \nu/2}\ |\ \sum_{i=1}^{n} (y_i - x_i\beta)(y_i - x_i\beta)^T + \nu\Omega(\gamma, R, \tau^2)\ |^{\ -\frac{n+\nu}{2}}\\
& \times p(R, D)\ p(\gamma)\ p(\tau^2)\ p(\nu)\ p(\beta).
\end{aligned}
$$

For certain parameters, this marginalized density is more difficult to handle than (4.2), but for others the marginalization does not increase the complexity. In each cycle of our algorithm, we condition on a simulated value of Σ draw from its full conditional posterior distribution given the other parameters, so this integration is performed stochastically rather than analytically.

For notational convenience, we define $G(\beta, \gamma, R, D, \tau^2, \nu)$ as the function

$$
|\gamma^T R\gamma + \tau^2|^{\ \frac{\nu}{2}} \times\ |\nu(\gamma^T R\gamma + \tau^2) + \sum_{i=1}^{n} (y_i - \beta^T x_i)(y_i - \beta^T x_i)^T|^{\ -\frac{n+\nu}{2}}.
$$

One cycle of our MCMC procedure, which can be viewed as an PX-DA algorithm with embedded Metropolis-Hastings steps, can be described as follows.

(1) Conditioning on Y_O, β and Σ, impute the missing elements of Y in the same way as under the normal model with unstructured covariance matrix described in Section 2.1. First, group the rows of Y according to their missingness patterns. If y_i bears missingness pattern s, we partition Σ into submatrices Σ_{sOO},Σ_{sMO},Σ_{sOM} and Σ_{sOO} corresponding to the observed and missing

variables. We then simulate the missing elements in y_i given the observed elements, drawing $y_{iM} \mid y_{iO}, \beta, \Sigma, \gamma, R, D, \tau^2, \nu \sim \mathrm{N}(\mu_i,\ \Sigma_i)$ independently for i=1,..., n, where

$$\mu_i = \beta_{iM}^T x_i + \Sigma_{sMO}\Sigma_{sOO}^{-1}(y_{iO} - \beta_{iO}^T x_i),$$
$$\Sigma_i = \Sigma_{sMM} - \Sigma_{sMO}\Sigma_{sOO}^{-1}\Sigma_{sOM}.$$

(2) Draw $\beta \mid Y_M, Y_O, \Sigma, \gamma, R, \tau^2, \nu$ from a multivariate normal distribution,

$$\mathrm{vec}(\beta|Y_M, Y_O, \Sigma) \quad \sim \quad N(\mathrm{vec}(\hat{\beta}),\ \Sigma \otimes (X^T X)^{-1}),$$

where $\hat{\beta} = (X^T X)^{-1} X^T Y$.

(3) Simulate (Σ, θ) from $\mathrm{P}(\Sigma, \theta|Y_M, Y_O, \beta, \nu)$ using the following four steps.

(a) Generate $\Sigma \mid Y_O, Y_M, \beta, \theta, \nu$ from

$$W^{-1}(\nu + n, \sum_{i=1}^{n}(y_i - \beta^T x_i)(y_i - \beta^T x_i)^T + \nu\Omega(\gamma, R, \tau^2)).$$

(b) To simulate $\gamma \mid Y_O, Y_M, \beta, \tau^2, R, D, \nu$, generate a candidate $\gamma^\star$ from the jumping kernel $q_1(\gamma^\star \mid \gamma)$ which was described earlier. Accept $\gamma^\star$ as γ with probability min(1, α_1), where

$$\alpha_1 = \frac{p_1(\gamma^\star) q_1(\gamma \mid \gamma^\star)}{p_1(\gamma) q_1(\gamma^\star \mid \gamma)} \times \frac{G(\beta, \gamma^\star,\ R,\ D,\ \tau^2, \nu)}{G(\beta, \gamma,\ R,\ D,\ \tau^2, \nu)},$$

where $p_1(\cdot)$ denotes the prior density for γ.

(c) To simulate $\tau^2 \mid Y_M, Y_O, \beta, \gamma, R, D, \nu$, generate a candidate $\tau^{2\star}$ according to a jumping kernel $\mathrm{q}_2(\tau^{2\star} \mid \tau^2)$ described earlier. Accept $\tau^{2\star}$ as τ^2 with

probability min(1, α_2), where

$$\alpha_2 = \frac{p_2(\tau^{2\star})q_2(\tau^{2(t)} \mid \tau^{2\star})}{p_2(\tau^{2(t)})q_2(\tau^{2(\star)} \mid \tau^2)} \times \frac{G(\beta, \gamma,\ R,\ D,\ \tau^{2\star}, \nu)}{G(\beta, \gamma,\ R,\ D,\ \tau^2, \nu)},$$

and where $p_2(\cdot)$ is the prior density for τ^2.

(d) To simulate $R, D \mid Y_M, Y_O, \beta, \gamma, \tau^2, \nu$, generate a candidate $(R^\star,\ D^\star)$ according to $q_3(R^\star,D^\star \mid R,D)$. Accept $(R^\star,D^\star)$ as (R,D) with probability min(1, α_3),

$$\alpha_3 = \frac{p_3(R^\star, D^\star)q_3(R,\ D \mid R^\star, D^\star)}{p_3(R,\ D)q_3(R^\star, D^\star \mid R,\ D)} \times \frac{G(\beta, \gamma,\ R^\star,\ D^\star,\ \tau^2, \nu)}{G(\beta, \gamma,\ R,\ D,\ \tau^2, \nu)},$$

where denotes the prior density of (R, D).

(4) If the parameter ν is treated as random, simulate $\nu \mid Y_M, Y_O, \beta, \gamma, R, D, \tau^2, \Sigma$ by generating a candidate $\nu_\star$ from the lognormal jumping kernel, and accept $\nu_\star$ with probability min(1, α_4), where

$$\begin{aligned}\alpha_\nu =& \frac{p_4(\nu^\star)q_4(\nu \mid \nu^\star)}{p_4(\nu)q_4(\nu^{(\star)} \mid \nu)} \times \frac{2^{\frac{\nu r}{2}}\Gamma_r(\frac{\nu}{2})}{2^{\frac{\nu^\star r}{2}}\Gamma_r(\frac{\nu^\star}{2})} \times \frac{\text{etr}(-\frac{1}{2}\nu^\star\Omega(\gamma, R, \tau^2)\Sigma^{-1})}{\text{etr}(-\frac{1}{2}\nu\Omega(\gamma, R, \tau^2)\Sigma^{-1})} \\ & \times \frac{\mid \nu^\star\Omega(\gamma, R, \tau^2)\Sigma^{-1} \mid^{\nu^\star/2}}{\mid \nu\Omega(\gamma, R, \tau^2)\Sigma^{(t+1)^{-1}} \mid^{\nu/2}},\end{aligned}$$

where $p_4(\cdot)$ and $q_4(\cdot)$ denote the prior and jumping kernel, respectively.

The performance of this imputation procedure will be evaluated in the next chapter when we apply it to simulated data.

Chapter 5

A Simulated Application

5.1 Purpose of the simulation study

The 19-item *Feelings Scale* section from Add Health has a simple yet realistic structure that makes it an appealing prototype application for these new procedures. As with any real dataset, however, values of the true population parameters are unknown, and a single sample from that population does not allow us to assess repeated-sampling properties.

The purpose of this simulation experiment is to evaluate the performance of the new softly constrained CFA imputation model and compare it to available and proposed alternatives. With respect to this new model, we have two main concerns. The first concern is bias. If a CFA model does not describe the population well, then how harmful can it be to apply this incorrect model? If the CFA model does not fit, does the inclusion of the random component in Σ to account for lack of fit effectively mitigate that bias? The second main concern is efficiency. If the CFA model does indeed describe the population well, then does the inclusion of the random component in Σ add unnecessary noise to the imputed values and

decrease efficiency of the resulting inferences?

5.2 Data generation

For this simulation, we created independent samples from a multivariate normal population whose means and covariances were estimated from the actual items in the *Feelings Scale*. We began with data from the Add Health Wave II student interview public use dataset, which is essentially a random half-sample of the participants made available by the Add Health investigators with minimal restrictions on its use. This sample yielded 4,595 observations on 19 items describing the participants' emotional states. We also included a single covariate, sex, which is mildly predictive of the *Feelings Scale* items. From these data, we estimated regression coefficients and covariances under the unstructured multivariate normal model described in Chapter 2, using the maximum-likelihood procedures in Schafer's (2008) NORM library for R. Regarding these parameters as population values, we then drew two hundred independent samples of $n = 300$ observations each and imposed missing values on the samples in various ways.

The sample size of $n = 300$ can be justified as follows. Based on personal experience, we know that it is not uncommon for a researcher with access to thousands of cases to attempt analyses involving hundreds of items. The largest applications of the unstructured normal model to date have involved approximately 150 variables; beyond that, the current procedures tend to break down or become computationally infeasible. Significant advantages would be realized if the maximum number of variables could be doubled to about 300, which is approximately 7% of the number of cases in our Add Health sample. Applying a 7% item-to-case ratio to the 20 items in question (19 *Feelings Scale* items plus sex), we obtained a target

sample size of roughly $n = 300$.

5.3 Missing data mechanisms

After drawing 200 random samples from the population of $n = 300$ cases each, we imposed patterns of missing values on each sample by two different random procedures.

The first procedure is a missing completely at random (MCAR) mechanism in which item nonresponse is clustered within individuals. With probability 0.7, an individual provided complete responses to all 19 items, and with probability 0.3, the individual answered each item with probability 0.9. This mechanism produces a low missingness rate per item of 3% and a 74% average rate of complete cases, which seems realistic for a survey like Add Health.

The second procedure is missing at random (MAR) in which missingness varies by covariates. For each of the 19 items, we set the logit-probability of missingness equal to $|\delta_0| + |\delta_1|X_{i,1} + |\delta_2|X_{i,2}$, where X_{i1} and X_{i2} are the regressors in the X matrix (a constant and sex), and $\delta_l \sim N(0.0.5)$ independently for $i = 0, 1, 2$. The regression coefficients were forced to be positive because, as pointed out by Song and Belin (2004), coefficients symmetrically distributed about zero may lead to prediction errors in both directions that tend to cancel each other out. Under those conditions, even naive procedures such as case deletion will have very little discernible bias.

5.4 Imputation models

Each incomplete data set was imputed using the following eight models: the normal model with unstructured covariance matrix, the normal model with a ridge prior with $\epsilon = 0.02$, a 1-factor EFA model, a 2-factor EFA model, a 3-factor EFA model, a 2-factor CFA model, and our new softly constrained 1-factor CFA models with prior distributions for $\log \nu$ centered at $\log 30$ and $\log 20$. Those two prior guesses were thought to be reasonable intermediate points between extremely small values of ν, which would approximate the unstructured normal model, and extremely high values for ν, which would approximate the 1-factor EFA solution.

Our 2-factor CFA model was motivated by the the exploratory factor analysis on the actual data (Section 3.1) which revealed that the 15 positively-worded items heavily loaded on the first factor and the 4 negatively-worded items mainly loaded on the second factor. Therefore, we assumed that the factor loadings γ have the following form, where '1' denotes a parameter that is freely estimated, and '0' denotes a parameter that is constrained to zero:

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix}$$

In the 1-factor CFA model with soft contraints, all items are assumed to load on the single factor.

The population covariance matrix for this example, which was empirically determined from Add Health, does not precisely follow a k-factor structure for k=1, 2, 3 or 4. Much of the inter-item correlations can be explained by the first principal component, but additional components beyond the first are not negligible. The unstructured covariance model with a ridge prior, and the EFA models with 1-3

latent factors, are conditions similar to those investigated by Song and Belin (2004) in their simulation work. The CFA model with 2 factors and the softly constrained CFA models are new conditions that have not yet been tried. Comparing 2-factor CFA with softly constrained 1-factor CFA is especially interesting, because these represent alternative strategies for accommodating lack of fit. One strategy, which was suggested by Song and Belin (2004), is to increase the number of components; the other strategy, which motivates our work, is to smoothly mix the inadequate factor model with an unstructured model.

5.5 Estimands

Following Song and Belin (2004), we selected several mean and correlation parameters that we believed would characterize the typical behavior and performance of our methods over repeated samples. We randomly chose three marginal means, μ_1, μ_7 and μ_{19}. We also chose pairs of items exhibiting some of the strongest and weakest correlations in the population, including items that were positively worded and negatively worded. The correlations we selected were $\rho_{2,11}$, $\rho_{11,17}$, and $\rho_{16,18}$.

5.6 Evaluation criteria

For each sample and each imputation model, we generated $M = 25$ imputations for the missing items by the algorithms described in Chapters 2–4. We computed complete-data estimates and standard errors for each target estimand Q and combined them using the well known rules developed by Rubin (1987). After obtaining the point estimate and confidence interval for Q from each sample and method, we computed four performance statistics suggested by Collins, Schafer and Kam

(2001): standardized bias, root-mean-square error (RMSE), average 95% width of confidence interval, and actual 95% coverage rate. Standardized bias is defined to be 100 $\times$ (average estimate minus true value)/SE, where SE is the standard deviation of Q. Standardized bias may be considered practically significant if its absolute value is greater than 50%, because that is the approximate point at which the actual coverage of a nominal 95% confidence interval drops below 90%, doubling the rate of Type 1 errors. RMSE is the average squared difference between the estimate and the true value. If two interval estimates have similar rates of coverage, the one that yields narrower confidence intervals should be preferred.

Because of the extensive computations required by these imputation procedures, we were able to perform only 200 replications for this dissertation. That number is not sufficient to accurately measure the coverage rates of nominal 95% intervals; ideally we would have liked to perform 1,000 or even 5,000 runs. The simulated coverage rates reported in the following tables have a margin of error of roughly $\pm 2\sqrt{.95 \times .05/200} = 3\%$. Despite that inaccuracy, comparisons among the coverage rates within each table may still be meaningful. Note that this is a *blocked* experiment; the same 200 samples were treated by all eight imputation methods. Moreover, the rates of missing information here are modest; with no missing values at all, the results from any method applied to the same sample would be identical. Under these conditions, a small difference in simulated coverage rates — say, 2.5%, which means that one method captured the true parameter in 5 samples when the other method did not — does provide some evidence that the actual coverage rates are different.

5.7 Simulation results

The performance of the eight methods for the six parameters is summarized in Tables 5.1–5.6. The results for μ_1 (Table 5.1) show that all imputation methods performed well for this parameter. Biases are neglible, and there is little variation in bias, RMSE or interval width among any of the methods. Estimates are less precise and intervals are wider for MAR than for MCAR, because this particular MAR mechanism produces slightly higher rates of missingness.

For the mean parameters μ_7 and μ_{19} (Tables 5.2–5.3), all methods performed well except the 2-factor CFA model, which seriously underestimated these means. At this moment, we are not entirely sure if those results are trustworthy. Although we have repeatedly checked our the imputation routines for CFA, a discrepancy of this size leads us to suspect that the program may still contain a bug. If this result is not due to a programming error, then it serves as a stark warning about the dangers of applying inappropriate constraints to an imputation model. And if the result is trustworthy, it bodes well for our strategy of applying soft constraints, because both versions of the softly constrained model perform much better than the inappropriate hard constraints of CFA.

Examining the results for $\rho_{2,11}$, $\rho_{11,17}$ and $\rho_{16,18}$ (Tables 5.4–5.6), we see some meaningful differences among the methods. Practically significant biases appear in the EFA methods for the MAR condition. These biases are especially pronounced for $\rho_{16,18}$. Interestingly, the softly constrained models, which represent a compromise between one-factor EFA and the unstructured model, show none of this tendency for bias. The allowance for lack of fit does appears to correct this difficulty in the EFA models without any noticeable inflation of variance.

Another encouraging sign is the nearly equivalent good performance of the

softly constrained model and the unstructured model in each of the Tables 5.1–5.6. With $n = 300$ cases and 20 variables, all parameters of the unstructured model are well estimated, and that model should perform well. The model with soft constraints is more parsimonious and makes more assumptions, but those assumptions do not seem to bias the results in any noticeable way. The generality of the unstructured model is an asset in this example, but in situations with many more variables and/or higher variable-to-case ratios, the unstructured model becomes unstable or unusable. As variables are added to the softly constrained CFA model, it too grows in complexity, but much more slowly than the unstructured model. As more questionnaire items and themes are included in an imputation procedure, we must eventually reach a tipping point where the generality of the unstructured model becomes a liability, and every additional variable will cause the performance to deteriorate relative to CFA with soft constraints.

Table 5.1. Simulation results of the unstructured normal model, the normal model with a ridge prior, the 1-factor EFA model, the 2-factor EFA model, the 3-factor EFA model, the ordinary 2-factor CFA model, and the "soft" constraint model, $Q = \mu_1$

Missingness	Models	Standardized Bias	RMSE	Actual 95% Coverage	Average Interval Width
MCAR	normal(unstructured)	-0.0606	0.0017	0.95	0.1608
	normal(ridge)	-0.0652	0.0017	0.94	0.1606
	1-factor EFA	-0.0631	0.0017	0.945	0.1612
	2-factor EFA	-0.0706	0.0017	0.95	0.1611
	3-factor EFA	-0.0628	0.0018	0.94	0.1611
	2-factor CFA	-0.0776	0.0017	0.95	0.1606
	normal(soft)[1]	-0.0622	0.0017	0.94	0.1606
	normal(soft)[2]	-0.0594	0.0018	0.945	0.1606
MAR	normal(unstructured)	-0.0223	0.0024	0.925	0.1858
	normal(ridge)	-0.0215	0.0024	0.92	0.1818
	1-factor EFA	-0.032	0.0025	0.935	0.1893
	2-factor EFA	-0.0135	0.0026	0.915	0.188
	3-factor EFA	-0.0287	0.0026	0.915	0.1891
	2-factor CFA	-0.0519	0.0024	0.92	0.181
	normal(soft)[1]	-0.0247	0.0024	0.93	0.1833
	normal(soft)[2]	-0.039	0.0024	0.92	0.1828

Table 5.2. Simulation results of the unstructured normal model, the normal model with ridge prior, the 1-factor EFA model, the 2-factor EFA model, the 3-factor EFA model, the ordinary 2-factor CFA model, and the "soft" constraint model, $Q = \mu_7$

Missingness	Models	Standardized Bias	RMSE	Actual 95% Coverage	Average Interval Width
MCAR	normal(unstructured)	-0.1364	0.0009	0.955	0.1177
	normal(ridge)	-0.1357	0.0009	0.955	0.1175
	1-factor EFA	-0.1461	0.0009	0.955	0.1181
	2-factor EFA	-0.1503	0.0009	0.945	0.118
	3-factor EFA	-0.1506	0.0009	0.95	0.1181
	2-factor CFA	-1.4138	0.0025	0.73	0.1176
	normal(soft)[1]	-0.1342	0.0009	0.95	0.1175
	normal(soft)[2]	-0.1363	0.0009	0.955	0.1175
MAR	normal(unstructured)	-0.1316	0.001	0.96	0.13505
	normal(ridge)	-0.1273	0.001	0.96	0.1327
	1-factor EFA	-0.1429	0.0012	0.97	0.1386
	2-factor EFA	-0.1598	0.0011	0.975	0.1389
	3-factor EFA	-0.1509	0.0012	0.955	0.1387
	2-factor CFA	-1.3053	0.0027	0.775	0.1328
	normal(soft)[1]	-0.141	0.001	0.96	0.1326
	normal(soft)[2]	-0.1354	0.001	0.965	0.1326

Table 5.3. Simulation results of the unstructured normal model, the normal model with ridge prior, the 1-factor EFA model, the 2-factor EFA model, the 3-factor EFA model, the ordinary 2-factor CFA model, and the "soft" constraint model, $Q = \mu_{19}$

Missingness	Models	Standardized Bias	RMSE	Actual 95% Coverage	Average Interval Width
MCAR	normal(unstructured)	-0.1234	0.0023	0.95	0.1932
	normal(ridge)	-0.1227	0.0023	0.95	0.1931
	1-factor EFA	-0.1327	0.0023	0.95	0.194
	2-factor EFA	-0.1244	0.0024	0.95	0.1939
	3-factor EFA	-0.1225	0.0024	0.95	0.1939
	2-factor CFA	-2.0013	0.0115	0.51	0.1933
	normal(soft)[1]	-0.1277	0.0023	0.955	0.193
	normal(soft)[2]	-0.1256	0.0023	0.955	0.193
MAR	normal(unstructured)	-0.0712	0.003	0.95	0.2209
	normal(ridge)	-0.0731	0.003	0.95	0.2172
	1-factor EFA	-0.0228	0.0032	0.945	0.2273
	2-factor EFA	-0.0434	0.0032	0.945	0.2288
	3-factor EFA	-0.0206	0.0033	0.95	0.2269
	2-factor CFA	-1.6661	0.0116	0.62	0.2204
	normal(soft)[1]	-0.0762	0.0029	0.945	0.2177
	normal(soft)[2]	-0.0785	0.0029	0.95	0.218

Table 5.4. Simulation results of the unstructured normal model, the normal model with ridge prior, the 1-factor EFA model, the 2-factor EFA model, the 3-factor EFA model, the ordinary 2-factor CFA model, and the "soft" constraint model, $Q = \rho_{2,11}$

	Models	Standardized Bias	RMSE	Actual 95% Coverage	Average Interval Width
MCAR	normal(unstructured)	-0.0595	0.0036	0.96	0.2335
	normal(ridge)	-0.0571	0.0036	0.96	0.2331
	1-factor EFA	-0.199	0.0035	0.97	0.2344
	2-factor EFA	-0.1997	0.0035	0.965	0.2345
	3-factor EFA	-0.1982	0.0035	0.965	0.2344
	2-factor CFA	-0.0204	0.0034	0.97	0.2334
	normal(soft)[1]	-0.0514	0.0036	0.97	0.233
	normal(soft)[2]	-0.0548	0.0036	0.97	0.2332
MAR	normal(unstructured)	-0.1899	0.0056	0.94	0.2996
	normal(ridge)	-0.1824	0.0061	0.925	0.295
	1-factor EFA	-1.7293	0.0069	0.94	0.2785
	2-factor EFA	-1.7212	0.0069	0.95	0.2801
	3-factor EFA	-1.6959	0.007	0.93	0.2801
	2-factor CFA	0.1107	0.0031	0.975	0.2772
	normal(soft)[1]	-0.1343	0.0056	0.92	0.2939
	normal(soft)[2]	-0.1303	0.0057	0.945	0.2962

Table 5.5. Simulation results of the unstructured normal model, the normal model with ridge prior, the 1-factor EFA model, the 2-factor EFA model, the 3-factor EFA model, the ordinary 2-factor CFA model, and the "soft" constraint model, $Q = \rho_{11,17}$

	Models	Standardized Bias	RMSE	Actual 95% Coverage	Average Interval Width
MCAR	normal(unstructured)	-0.0361	0.0033	0.95	0.2328
	normal(ridge)	-0.0342	0.0033	0.95	0.2325
	1-factor EFA	-0.134	0.003	0.965	0.2342
	2-factor EFA	-0.1257	0.003	0.97	0.2341
	3-factor EFA	-0.1285	0.003	0.97	0.2339
	2-factor CFA	-0.0072	0.0031	0.965	0.2326
	normal(soft)[1]	-0.0281	0.0033	0.95	0.2326
	normal(soft)[2]	-0.0291	0.0033	0.95	0.2327
MAR	normal(unstructured)	-0.1311	0.0047	0.965	0.3036
	normal(ridge)	-0.1032	0.0048	0.96	0.2972
	1-factor EFA	-1.1675	0.0038	0.99	0.2794
	2-factor EFA	-1.1399	0.0038	0.99	0.2804
	3-factor EFA	-1.0975	0.0038	0.995	0.2802
	2-factor CFA	-0.1822	0.0022	1	0.2741
	normal(soft)[1]	-0.0764	0.0045	0.965	0.2945
	normal(soft)[2]	-0.074	0.0045	0.97	0.2971

Table 5.6. Simulation results of the unstructured normal model, the normal model with ridge prior, the 1-factor EFA model, the 2-factor EFA model, the 3-factor EFA model, the ordinary 2-factor CFA model, and the "soft" constraint model, $Q = \rho_{16,18}$

Missingness	Models	Standardized Bias	RMSE	Actual 95% Coverage	Average Interval Width
MCAR	normal(unstructured)	0.1679	0.004	0.935	0.2335
	normal(ridge)	0.1744	0.0041	0.935	0.2329
	1-factor EFA	-0.1852	0.0037	0.945	0.2363
	2-factor EFA	-0.1765	0.0037	0.95	0.2363
	3-factor EFA	-0.1763	0.0037	0.945	0.2361
	2-factor CFA	0.2305	0.0041	0.945	0.2329
	normal(soft)[1]	0.1714	0.0041	0.93	0.2328
	normal(soft)[2]	0.1771	0.0041	0.925	0.233
MAR	normal(unstructured)	-0.0203	0.0068	0.955	0.3022
	normal(ridge)	0.0851	0.0072	0.93	0.2957
	1-factor EFA	-3.668	0.0311	0.28	0.2841
	2-factor EFA	-3.7099	0.0311	0.29	0.2851
	3-factor EFA	-3.5736	0.0307	0.3	0.2843
	2-factor CFA	-0.5682	0.008	0.905	0.3014
	normal(soft)[1]	0.0516	0.0062	0.94	0.2967
	normal(soft)[2]	0.0725	0.0065	0.94	0.2974

Chapter 6

Discussion

6.1 What has been accomplished

Despite two decades of rapid growth in multiple imputation methods, data analysts who work with large, multi-themed questionnaires still face daunting challenges. The imputation models and software that are currently available, most of which are based on an unstructured normal model, still cannot handle the large numbers of variables necessary for many research projects.

Building upon the exploratory factor imputation model of Song and Belin (2004), we proposed a confirmatory version specifically designed for multi-themed questionnaires. We developed algorithms for parameter estimation and multiple imputation under this confirmatory model, and then we extended the model to accommodate lack of fit by allowing the population covariance matrix to randomly deviate from the posited factor structure.

Missing data are usually a nuisance, not the main focus of scientific inquiry, and many researchers cannot afford to spend a great deal of effort and resources to develop and fine-tune an imputation model for any specific application. Ideally,

we would like to have a flexible class of procedures that would enable the imputer quickly specify and build a model that may not be perfect, but is good enough for the task at hand, so that he or she may impute the missing values quickly and move on. If the number of variables is large, the procedure may require the imputer to make intelligent decisions about which relationships are of primary importance and need to be preserved, and which ones are less crucial and may be omitted. Decisions made during the imputation-modeling phase will inform data users regarding the possible impact of the imputation on subsequent analyses. An ideal imputation procedure would also be self-correcting in the sense that if some relationships posited by the imputer are strongly contradicted by the data, the model would automatically relax those assumptions to accommodate the data. The methods that we have proposed and developed in this dissertation are a meaningful step toward this goal, but much work remains to be done before these procedures are ready for routine use.

6.2 Work that remains

The simulation study presented in Chapter 5 involved 20 variables in samples of 300 cases, a situation where the unstructured covariance model works well and specialized techniques are not really needed. That simulation was primarily a test to see whether this proposed imputation scheme is feasible and well behaved. Additional simulations are needed see how the new method works when the unstructured model performs poorly (e.g., 400 cases and 100 variables) or fails without significant prior input (e.g., 100 cases and 100 variables).

More investigations are needed to understand and fine-tune the soft constraints that allow Σ to randomly deviate from the factor model. Is it better to treat the

degrees of freedom ν as an unknown parameter and impose a prior distribution on it, as we have done, or to fix it at a point estimate? To do the latter, we would need a reliable method to estimate ν. It may be possible to expand the PX-EM procedure developed in Chapter 4 to estimate this parameter along with the others. Researchers who use factor analysis and structural-equations models are familiar with the concept of a fit index. A fit index is essentially a goodness-of-fit statistic that compares the given model to the unstructured alternative, but the statistic is scaled by the sample size so that the fit does not appear to worsen merely because n has increased. The parameter ν is obviously related to a fit index, but the nature of that relationship has not yet been described. If rules of thumb could be developed that show correspondence between ν and the popular fit indices, then researchers would be able to guess a reasonable value for ν and either fix ν at that guess or propose a prior distribution for ν centered at that guess.

By using an inverted-Wishart distribution to characterize the discrepancy between Σ and the proposed factor structure, we have described the discrepancy by a single parameter ν. Yet we can imagine a situation where a factor model describes the relationships among items well, except that one pair of items has a much higher correlation than the model predicts. An experienced data analyst could introduce a residual covariance between those two items, effectively allowing an off-diagonal element of the uniqueness matrix τ^2 to be nonzero. Our inverted-Wishart approach might not react well to that situation; it may smooth the aberrant correlation coefficient too little or too much. It is worthwhile to investigate alternatives to the inverted Wishart that could apply different degrees of smoothing to different portions of Σ if warranted by the data.

6.3 Extensions to discrete items

Our assumption of normality is not well suited to survey items that are binary, ordinal or nominal. Aside from computational convenience, the main reason why we chose to work within the multivariate normal framework is that it is feasible to extend this model to accommodate discrete variables. A multivariate probit model describes binary items as coarsened versions of correlated latent normal variates. Chib and Greenberg (1998) presented MCMC methods for Bayesian posterior simulation in multivariate probit models. The main difficulty in these methods, as Chib and Greenberg (1998) noted, is that the variances of the latent normal variates must be fixed to identify the model parameters. These variances are usually fixed at one, so that the covariance matrix decribing these relationships becomes a correlation matrix. Bayesian inference for correlation matrices is awkward because, as we noted in Chapter 3, convenient prior distributions are not available and posterior distributions are difficult to simulate. However, the parameter extension method that we applied in Chapters 3–4, which augments the correlations by a vector of inestimable variances, has been demonstrated to be effective in simplifying the computations for multivariate probit models (Liu, 2000).

Boscardin et al. (2006, 2008) applied this parameter-extension method to an extended class of multivariate probit models that describe binary, ordinal and nominal items, and this extension can be incorporated into our framework without much difficulty. The extension would add two steps to the MCMC procedures described in Chapters 3 and 4. One step would simulate posterior draws of the unknown threshold values that relate the latent normal variates to the observed items. Drawing these thresholds can be done one variable at a time using straightforward procedures described by Boscardin et al. (2008). The second additional

step requires conditional simulation of the latent normal variates given their correlation matrix and the observed discrete items. With r binary responses, this could be done by simulating an r-dimensional normal candidate with given correlations and accepting the candidate if it falls within the correct orthant. That procedure is very inefficient for large values of r, and it may be replaced by a Gibbs sampler that cycles through the r variates and simulates each one given the others from a truncated normal distribution (Chib & Greenberg, 1998). The extension of this procedure to items with three or more levels is immediate (Boscardin et. al, 2008).

6.4 Extensions to multilevel data

In the Add Health study, participants were selected by a two-stage sampling procedure in which the participating schools were sampled from a master list of schools, and then students were selected within the schools. The clustering of students within the schools is a crucial element of multilevel regression analyses which seek to explain inter- and intra-school variability. Even when the clustering is a nuisance, e.g., when the data are analyzed by conventional regression methods, acknowledging the two-stage sampling procedure can be crucial for computing correct standard errors, and the clustering ought to be reflected in the imputation model as well.

The imputation models that we have developed can be expanded to account for clustering in the following way. Let y_{ic} denote the vector of variables to be imputed for participant i nested within cluster c. A multilevel factor model can be formulated as

$$y_{ic} \,|\, \delta_c \;\sim\; N(\,\beta^T x_{ic} + \delta_c^T w_{ic},\, \gamma^T R \gamma + \tau^2),$$

where x_{ic} and w_{ic} are vector of covariates, and δ_c is a matrix of random coefficients distributed as $\text{vec}(\delta) \sim N(0, \Delta)$. The vector w_{ic} will usually include a constant and possibly additional variables from x_{ic}, and Δ may be assumed to have a block-diagonal structure with r independent blocks corresponding to the variables in y_{ic}. This model combines factor analysis with the multilevel features of the imputation models described by Schafer and Yucel (2002). If the cluster-level random effects δ_c were known, then the model for $y_{ic} - \delta_{ic}^T w_{ic}$ would reduce to EFA or CFA. Therefore, we can accommodate this extension by adding two more steps to the MCMC procedures previously described. One step would sample the random effects δ_c for each cluster from their posterior distribution given the parameters and complete data. The other step would sample Δ from its posterior distribution given the random effects. The block-diagonal structure of Δ could be accommodated by applying independent prior distributions to each block.

Appendix A

The Gradient and the Hessian Matrix under the One-factor Per Section Model

A.1 The actual log-likelihood

Let $\Phi_{iO} = \gamma_{iO}^T R \gamma_{iO} + \tau_{iO}^2$. The actual log-likelihood based on Y_O is

$$\begin{aligned} l_0(\theta) =& \sum_{i=1}^{n} l_{0,i}(\theta) \\ \propto & -\frac{1}{2}\sum_{i=1}^{n} \log|\Phi_{iO}| - \frac{1}{2}\sum_{i=1}^{n} \left(y_{iO} - \beta_{iO}^T x_i\right)^T (\Phi_{iO})^{-1} (y_{iO} - \beta_{iO}^T x_i) \\ \propto & -\frac{1}{2}\sum_{i=1}^{n} \log|\Phi_{iO}| - \frac{1}{2}\sum_{i=1}^{n} y_{iO}^T (\Phi_{iO})^{-1} y_{iO} \\ & -\frac{1}{2}\sum_{i=1}^{n} x_i^T \beta_{iO} (\Phi_{iO})^{-1} \beta_{iO}^T x_i + \sum_{i=1}^{n} y_{iO}^T (\Phi_{iO})^{-1} \beta_{iO}^T x_i. \end{aligned}$$

Let r_i denote the total number of the observed variables for the ith unit, $\beta_{iO,l^\star}$ denote the $l^\star$th column of β_{iO} which is indexed as the lth column in β_i and $(\Omega_{iO}^{-1})_{,u^\star}$

denote the $u^\star$th column of Ω_{iO}^{-1} which is indexed as the uth column in Ω_i^{-1}.

A.2 The first derivative vector

For $i = 1, \ldots, n$, $j = 1, \ldots, r$, $l = 2, \ldots, k$, and $m = 1, \ldots, l-1$,

$$
\begin{aligned}
\frac{\partial l_0}{\partial \beta_j} &\propto \sum_{i=1}^{n} x_i (y_{iO} - \beta_{iO}^T x_i)^T (\Omega_{iO}^{-1})_{,j^\star} \mathbf{1}_{ij} \\
\frac{\partial l_0}{\partial \tau_j^2} &\propto -\frac{1}{2} \sum_{i=1}^{n} (\Omega_{iO}^{-1})_{j^\star,j^\star} \mathbf{1}_{ij} + \frac{1}{2} \sum_{i=1}^{n} [((y_{iO} - \beta_{iO}^T x_i)^T (\Omega_{iO}^{-1})_{,j^\star})^2 \mathbf{1}_{ij}] \\
\frac{\partial l_0}{\partial \gamma_{c_j,j}} &\propto -\sum_{i=1}^{n} \sum_{u^\star=1}^{r_i} (\Omega_{iO}^{-1})_{j^\star,u^\star} R_{c_j,c_u} \gamma_{iO,c_u,u^\star} \mathbf{1}_{ij} \\
&\quad + \sum_{i=1}^{n} \sum_{u^\star=1}^{r_i} (\Omega_{iO}^{-1})_{j^\star,} (y_{iO} - \beta_{iO}^T x_i) (\Omega_{iO}^{-1})_{u^\star,} (y_{iO} - \beta_{iO}^T x_i) R_{c_j,c_u} \gamma_{iO,c_u,u^\star} \mathbf{1}_{ij} \\
\frac{\partial l_0}{\partial R_{l,m}} &\propto -\sum_{i=1}^{n} \sum_{1 \le v^\star < u^\star \le r_i} \frac{\partial \Omega_{iO,u^\star,v^\star}}{\partial R_{lm}} (\Omega_{iO}^{-1})_{u^\star,v^\star} \\
&\quad + \sum_{i=1}^{n} \sum_{1 \le v^\star < u^\star \le r_i} [\frac{\partial \Omega_{iO,u^\star,v^\star}}{\partial R_{lm}} (\Omega_{iO}^{-1})_{u^\star,} (y_{iO} - \beta_{iO} x_i) (\Omega_{iO}^{-1})_{v^\star,} (y_{iO} - \beta_{iO} x_i)],
\end{aligned}
$$

where

$$
\frac{\partial \Omega_{iO,u^\star,v^\star}}{\partial R_{l,m}} = \begin{cases} \gamma_{iO,l,u^\star} \gamma_{iO,m,v^\star} & \text{if } c_u = l, c_v = m \\ 0 & \text{otherwise.} \end{cases}
$$

A.3 The second derivative matrix

For $i = 1, \ldots, n$, $j = 1, \ldots, r$, $q = 1, \ldots, r$, $l, l_2 = 2, \ldots, k$, $m = 1, \ldots, l-1$ and

$m_2 = 1, \ldots, l_2 - 1,$

$$
\begin{aligned}
\frac{\partial^2 l_0}{\partial \beta_j^T \partial \beta_q} \propto & -\sum_{i=1}^{n} x_i x_i^T (\Omega_{iO}^{-1})_{j^\star,q^\star} \mathbf{1}_{ij} \mathbf{1}_{iq} \\
\frac{\partial^2 l_0}{\partial \beta_j \partial a_1} \propto & \sum_{i=1}^{n} x_i (y_{iO} - \beta_{iO}^T x_i)^T \frac{\partial (\Omega_{iO}^{-1})_{,j^\star}}{\partial a_1} \mathbf{1}_{ij} \mathbf{1}_{iq} \\
\frac{\partial^2 l_0}{\partial \tau_j^2 \partial a_2} \propto & -\frac{1}{2} \sum_{i=1}^{n} \frac{\partial (\Omega_{iO}^{-1})_{j^\star,j^\star}}{\partial a_2} \mathbf{1}_{ij} \mathbf{1}_{iq} \\
& + \sum_{i=1}^{n} (y_{iO} - \beta_{iO}^T x_i)^T (\Omega_{iO}^{-1})_{,j^\star} (y_{iO} - \beta_{iO}^T x_i)^T \frac{\partial (\Omega_{iO}^{-1})_{,j^\star}}{\partial a_2} \mathbf{1}_{ij} \mathbf{1}_{iq}
\end{aligned}
$$

where a_1, a_2 could be $\gamma_{c_q,q}$ or τ_q^2;

$$
\begin{aligned}
\frac{\partial^2 l_0}{\partial \beta_j \partial R_{l,m}} \propto & \sum_{i=1}^{n} x_i (y_{iO} - \beta_{iO}^T x_i)^T \frac{\partial (\Omega_{iO}^{-1})_{,j^\star}}{\partial R_{l,m}} \mathbf{1}_{ij} \\
\frac{\partial^2 l_0}{\partial \tau_j^2 \partial R_{l,m}} \propto & -\frac{1}{2} \sum_{i=1}^{n} \frac{\partial (\Omega_{iO}^{-1})_{j^\star,j^\star}}{\partial R_{l,m}} \mathbf{1}_{ij} \\
& + \sum_{i=1}^{n} (y_{iO} - \beta_{iO}^T x_i)^T (\Omega_{iO}^{-1})_{,j^\star} (y_{iO} - \beta_{iO}^T x_i)^T \frac{\partial (\Omega_{iO}^{-1})_{,j^\star}}{\partial R_{l,m}} \mathbf{1}_{ij} \\
\frac{\partial^2 l_0}{\partial \gamma_{c_j,j} \partial \gamma_{c_q,q}} \propto & -\sum_{i=1}^{n} (\Omega_{iO}^{-1})_{j^\star,q^\star} R_{c_j,c_q} \mathbf{1}_{ij} \mathbf{1}_{iq} \\
& - \sum_{i=1}^{n} \sum_{u^\star=1}^{r_i} \frac{\partial (\Omega_{iO}^{-1})_{j^\star,u^\star}}{\partial \gamma_{c_q,q}} R_{c_j,c_u} \gamma_{iO,c_u,u^\star} \mathbf{1}_{ij} \mathbf{1}_{iq} \\
& + \sum_{i=1}^{n} \sum_{u^\star=1}^{r_i} [(y_{iO} - \beta_{iO}^T x_i)^T \frac{\partial (\Omega_{iO}^{-1})_{,j^\star}}{\partial \gamma_{c_q,q}} (y_{iO} - \beta_{iO}^T x_i)^T (\Omega_{iO}^{-1})_{,u^\star} \\
& \times R_{c_j,c_u} \gamma_{iO,c_u,u^\star} \mathbf{1}_{ij} \mathbf{1}_{iq}] \\
& + \sum_{i=1}^{n} \sum_{u^\star=1}^{r_i} [(y_{iO} - \beta_{iO}^T x_i)^T (\Omega_{iO}^{-1})_{,j^\star} (y_{iO} - \beta_{iO}^T x_i)^T \frac{\partial (\Omega_{iO}^{-1})_{,u^\star}}{\partial \gamma_{c_q,q}} \\
& \times R_{c_j,c_u} \gamma_{iO,c_u,u^\star} \mathbf{1}_{ij} \mathbf{1}_{iq}] \\
& + \sum_{i=1}^{n} [(y_{iO} - \beta_{iO}^T x_i)^T (\Omega_{iO}^{-1})_{,j^\star} (y_{iO} - \beta_{iO}^T x_i)^T (\Omega_{iO}^{-1})_{,q^\star}
\end{aligned}
$$

$$
\begin{aligned}
&\times R_{c_j,c_q}\mathbf{1}_{ij}\mathbf{1}_{iq}] \\
\frac{\partial^2 l_0}{\partial\gamma_{c_j,j}\partial R_{l_2,m_2}} \propto & -\sum_{i=1}^{n}\sum_{u^\star=1}^{r_i}\frac{\partial(\Omega_{iO}^{-1})_{j^\star,u^\star}}{\partial R_{l_2,m_2}}R_{c_j,c_u}\gamma_{iO,c_u,u^\star}\mathbf{1}_{ij} \\
& -\sum_{i=1}^{n}\sum_{u^\star=1}^{r_i}(\Omega_{iO}^{-1})_{j^\star,u^\star}\gamma_{iO,c_u,u^\star}\mathbf{1}_{(max(c_u,c_j)=l_2,min(c_u,c_j)=m_2)}\mathbf{1}_{ij} \\
& +\sum_{i=1}^{n}\sum_{u^\star=1}^{r_i}[(y_{iO}-\beta_{iO}x_i)^T\frac{\partial(\Omega_{iO}^{-1})_{,j^\star}}{\partial R_{l_2,m_2}} \\
& (y_{iO}-\beta_{iO}x_i)^T(\Omega_{iO}^{-1})_{,u^\star}\times R_{c_j,c_u}\gamma_{iO,c_u,u^\star}\mathbf{1}_{ij}] \\
& +\sum_{i=1}^{n}\sum_{u^\star=1}^{r_i}[(y_{iO}-\beta_{iO}x_i)^T(\Omega_{iO}^{-1})_{,j^\star} \\
& (y_{iO}-\beta_{iO}x_i)^T\frac{\partial(\Omega_{iO}^{-1})_{,u^\star}}{\partial R_{l_2,m_2}}\times R_{c_j,c_u}\gamma_{iO,c_u,u^\star}\mathbf{1}_{ij}] \\
& +\sum_{i=1}^{n}\sum_{u^\star=1}^{r_i}[(y_{iO}-\beta_{iO}^Tx_i)^T(\Omega_{iO}^{-1})_{,j^\star}(y_{iO}-\beta_{iO}^Tx_i)^T(\Omega_{iO}^{-1})_{,u^\star} \\
& \times\mathbf{1}_{(max(c_j,c_u)=l_2,min(c_j,c_u)=m_2)}\gamma_{iO,c_u,u^\star}\mathbf{1}_{ij}] \\
\frac{\partial^2 l_0}{\partial R_{l,m}\partial R_{l_2,m_2}} \propto & -\sum_{i=1}^{n}\sum_{u^\star=1}^{r_i}\sum_{v^\star=1}^{r_i}\mathbf{1}_{c_u=l}\gamma_{iO,c_u,u^\star}\mathbf{1}_{c_v=m}\gamma_{iO,c_v,v^\star}\frac{\partial(\Omega_{iO}^{-1})_{u^\star,v^\star}}{\partial R_{l_2,m_2}} \\
& +\sum_{i=1}^{n}\sum_{u^\star=1}^{r_i}\sum_{v^\star=1}^{r_i}[\mathbf{1}_{c_u=l}\gamma_{iO,c_u,u^\star}\mathbf{1}_{c_v=m}\gamma_{iO,c_v,v^\star}\frac{\partial(\Omega_{iO}^{-1})_{u^\star,}}{\partial R_{l_2,m_2}} \\
& (y_{iO}-\beta_{iO}x_i)(\Omega_{iO}^{-1})_{v^\star,}(y_{iO}-\beta_{iO}x_i) \\
& +\sum_{i=1}^{n}\sum_{u^\star=1}^{r_i}\sum_{v^\star=1}^{r_i}[\mathbf{1}_{c_u=l}\gamma_{iO,c_u,u^\star}\mathbf{1}_{c_v=m}\gamma_{iO,c_v,v^\star}(\Omega_{iO}^{-1})_{u^\star,} \\
& (y_{iO}-\beta_{iO}x_i)\frac{\partial(\Omega_{iO}^{-1})_{v^\star,}}{\partial R_{l_2,m_2}}(y_{iO}-\beta_{iO}x_i)],
\end{aligned}
$$

where the derivatives of $(\Omega_{iO}^{-1})_{j^\star,u^\star}$ can be derived using the formula

$$
\frac{\partial(\Omega_{iO})^{-1}}{\partial\theta}=-(\Omega_{iO})^{-1}(\frac{\partial\Omega_{iO}}{\partial\theta})(\Omega_{iO})^{-1},
$$

and the derivatives of Ω_{iO} w.r.t the parameters are

$$\begin{aligned}
\frac{\partial \Omega_{iO,j^\star,u^\star}}{\partial \tau_q^2} \propto & \mathbf{1}_{ij}\mathbf{1}_{iq}\mathbf{1}_{j^\star=u^\star=q^\star} \\
\frac{\partial \Omega_{iO,j^\star,u^\star}}{\partial \gamma_{c_q,q}} \propto & \ \mathbf{1}_{ij}\mathbf{1}_{iq}[\mathbf{1}_{j^\star=q^\star}R_{c_q,c_u}\gamma_{iO,c_u,u^\star} + \mathbf{1}_{u^\star=q^\star}R_{c_q,c_j}\gamma_{iO,c_j,j^\star}] \\
\frac{\partial \Omega_{iO,j^\star,u^\star}}{\partial R_{l_2,m_2}} \propto & \ \mathbf{1}_{ij}\gamma_{iO,l_2,j^\star}\gamma_{iO,m_2,u^\star}\mathbf{1}_{c_j=l_2,c_u=m_2},
\end{aligned}$$

for $1 \leq j^\star \leq u^\star \leq r_i$. The other elements of the derivative matrix of Ω_{iO} and the Hessian matrix can be derived by symmetry.

Bibliography

[1] Agresti, A. (1989), A survey of models for repeated ordered categorical response data, *Statistics Medicine*, 8, 1209-1224.

[2] Allison, P.D., http://www.ssc.upenn.edu/ allison.

[3] Allison, P.D. (2005). Imputation of categorical variables with PROC MI. SAS Users Group International Conference, Philadelphia, PA, April 10-13.

[4] Barnard, J., McCulloch, R. & Meng, X.L. (2000), Modeling covariance matrices in terms of standard deviations and correlations with application to shrinkage. *Statistica Sinica*, 10, 1281- 1311.

[5] Basilevsky, A., Statistical Factor Analysis and Related Methods: Theory and Applications. John Wiley & Sons, New York: Wiley-Interscience.

[6] Bernaards, C.A., Belin, T.R. & Schafer, J.L (2007). Robustness of a multivariate normal approximation for imputation of binary incomplete data. *Statistics in Medicine*, 26, 1368-82.

[7] Bryk, A. S. & Raudenbush, S. W. (1992), Hierarchical Linear Models: Applications and Data Analysis Methods. Sage Publications Inc.

[8] Boscardin, W. J. & Zhang, X. (2004), Modeling the covariance and correlation matrix of repeated measures. *Applied Baysian modeling and casual inference from incomeplete-data perspectives*, E.d. Gelman, A. and Meng, X-L., John Wiley & Sons, Ltd.

[9] Boscardin,W.J, Zhang, X, & Belin, T.R. (2008), Modeling a mixture of ordinal and continuous repeated measures. *Journal of Statistical Computation and Simulation*. DOI: 10.1080/00949650701480259.

[10] Boscardin, W.J. (2008), A Joint modeling for incomplete repeated measures data of mixed data types. *presented at JSM*.

[11] Carpenter J.R & Goldstein H. (2004), Multiple imputation in MLwiN. *Multilevel Modelling Newsletters*, www.missingdata.org.uk.

[12] Channon, A. R. (2008), Multilevel multiple imputation of missing birth weights in developing countries: Analysing neonatal and post-neonatal mortality, Southampton Statistical Sciences Research Institute and Division of Social Statistics, University of Southampton, http://paa2008.princeton.edu/download.aspx?submissionId=80288.

[13] Centre for Multilevel Modelling Team, University of Bristol. http://www.cmm.bristol.ac.uk.

[14] Chib, S. & Greenberg, E. (1995), Understanding the Metropolis-Hastings algorithm. *The American Statistician*, 49, 327-335.

[15] Chib, S., & Greenberg, E. (1998), Bayesian Analysis of Multivariate Probit Models. *Biometrika*, 85, 347-361.

[16] Collins, L. M., Schafer, J. L., & Kam, C. M. (2001), A comparison of inclusive and restrictive strategies in modern missing-data procedures. *Psychological Methods*, 6, 330351.

[17] Collins, L. M., Wugalter S. E. (1992), Latent class models for stage-sequential dynamic latent variables. *Multivariate Behavioral Research*, 27, 131157.

[18] Cox, D.R. & Hinkley, D.V. (1974), Theoretical Statistics, Chapman & Hall, London.

[19] Daniels, M.J. & Kass, R.E. (1999), Nonconjugate Bayesian estimation of covariance matrices and its use in hierarchical models. *Journal of the American Statistical Association*, 94, 1254-1263.

[20] David, M., Little, R.J.A., Samuhel, M.E. & Triest, R.K. (1986), Alternative methods for CPS income imputation, *Journal of the American Statistical Association*, 81, 29-41.

[21] Dempster, A.P., Laird, N.M. & Rubin, D.B.(1977), Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society Series B*, 39, 1-38.

[22] Drasgow, F. (1986), Polychoric and polyserial correlations, *Encyclopedia of Statistical Science*, New York: Wiley.

[23] Fitzmaurice, G. & Laird N.M, Ware, J. (2004), Applied longitudinal analysis, John Wiley & Sons. bibitem Gary, K., Honaker, J. , Joseph, A. & Scheve, K. (2001), Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation. *American Political Science Review*, 95, 49-69,

[24] Gelman A., Carlin J. B., Stern H., and Rubin D. B. (1997), Bayesian data analysis. Chapman & Hall/ CRC, Boca Raton, 1st edition.

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

[26] Gibson, N.M. & Olejnik, S. (2003), Treatment Of Missing Data At The Second Level Of Hierarchical Linear Models. *Educational and Psychological Measurement*, 63, 204-238.

[27] Gilks, W.R., Richardson, S. & Spiegelhalter, D.J., eds (1996), *Markov Chain Monte Carlo in Practice*, Chapman & Hall, London.

[28] Golden, H. (1986), Multilevel Mixed Linear Model Analysis Using Iterative Generalized Least Squares, *Biometrika*, 73, 43-56

[29] Golden, H. & Rashbash, J. (1994), Efficient analysis of mixed hierarchical and crossed-classified random structures using a multilevel model, *Journal of Educational and Behavioral Statistics*, 19, 337-350.

[30] Graham, J.W. & Hofer, S.M. (1993), EMCOV.EXE users' guide[Computer software manual]. Department of Prevention Research, University of Southern California, Los Angeles.

[31] Groves, R.M., Fowler, F.J., Couper M.P., Lepkowski, J.M., Singer E., & Tourangeau, R. (2004), Survey Methodology. New York: Wiley.

[32] Hartley, H.O. & Rao, J.N.K. (1967), Maximum likelihood estimation for the mixed analysis of variance model. *Biometrika*, 54, 93-108.

[33] Harel, O. & Zhou, X. H. (2007), Multiple imputation: review of theory, implementation and software. *Statiscs in Medicine*, 26, 30573077.

[34] Honaker, J., King, G, Blackwell, M. (2007), Amelia II: A Program for Missing Data. http://gking.harvard.edu/amelia/docs/amelia.pdf.

[35] Howell, D.C. (2007), Treatment of missing data. http://www.uvm.edu/ dhowell.

[36] Huttenlocher, J.E., Haight, W., Bryk, A.S., & Seltzer, M. (1991), Early vocabulary growth: Relation to language input and gender. *Developmental Psychology*, 27(2), 236-249.

[37] Institute of Education (2005), MLwiN Version 2.02.

[38] Jacobusse, G. (2005), WinMICE Users Manual for WinMICE prototype (Version 0.1), Leiden, The Netherlands: Netherlands Organization for Applied Scientific Research (TNO), Retrieved Sep. 10, 2008 from http://web.inter.nl.net/users/S.van.Buuren/mi/hmtl/mice.htm.

[39] Jennrich, R.I. & Schluchter, M.D. (1986), Unbalanced repeated-measures models with structured covariance matrices. *Biometrics*, 38, 963-974.

[40] Jöreskog, K. G., & Sörbom, D. (2001), LISREL (Version 8.5) [Computer software]. Chicago: Scientific Software International.

[41] Laird. N.M., Lange, N. & Stram, D. (1987), Maximum likelihood computations with repeated measures: application of the EM algorithm. *Journal of the American Statistical Association*, 82, 97-105.

[42] Laird. N.M. (1982), Random-Effects Models for Longitudinal Data. *Biometrics*, 38, 97-10.

[43] Loken, E., & Molenaar, P. (2008). Categories or continua? The correspondence between mixture models and factor models. In G. R. Hancock & K. M. Samuelsen (Eds.), Advances in latent variable mixture models (pp. 277-297). Charlotte, NC: Information Age Publishing.

[44] Li, J, Alterman. T & Deddens, J.A. (2006), Analysis of Large Hierarchical Data with Multilevel Logistic Modeling Using PROC GLIMMIX. http://www2.sas.com/proceedings/sugi31/151-31.pdf.

[45] Lidstrom, M.J. and Bates, D.M. (1988), Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association*, 83, 1014-1022.

[46] Liechty, J., Liechty, M., & Muller, P. (2004), Bayesian Correlation Estimation, *Biometrika*, 91, 1-14.

[47] Littell, R.C., Milliken, G.A., Stroup, W.W. & Wolfinger, R.D. (1996), SAS System for Mixed Models, Cary, NC: SAS Institute, Inc.

[48] Little, R.J.A. & Rubin, D.B. (1987), Statistical analysis with missing Data, J. Wiley & Sons, New York.

[49] Little, R.J.A. & Yau, L. (1996), Intent-to-treat analysis for longitudinal studies with drop-outs. *Biometrics*, 52, 142-147.

[50] Liu, C. & Rubin, D. (1994), The ECME algorithm: A simple extension of EM and ECM with faster monotone convergence. *Biometrika*, 81, 633-648.

[51] Liu, C., Rubin, D. B. (1998), Maximum likelihood estimation of factor analysis using the ECME algorithm with complete and incomplete data. *Statistics Sinica*, 8, 729-747.

[52] Liu, C., Rubin, D. B., & Wu, Y. N. (1998), Parameter expansion to accelerate EM: the PX-EM algorithm. *Biometrika*, 85, 755-770.

[53] Liu, M., Taylor J.M.G. & Belin, T.R. (2000), Multiple imputation and posterior simulation for multivariate missing data in longitudinal studies. *Biometrics*, 56, 1157-1163.

[54] Liu,X., & Daniels M.J. (2006), A New algorithm for simulating a correlation matrix based on parameter expansion and re-parameterization. *Journal of Computational and Graphical Statistics*, 1, 897-914(18),

[55] Mazumdar, S., Liu, K.S., Houck, P.R. & Reynolds, III C.F. (1999), Intent-to-treat analysis for longitudinal clinical trials: copint with the challenge of missing values. *Journal of Psychiatric Research*, 33, 87-95.

[56] McLachlan, G. & Krishnan, T. (1997), The EM algorithm and extensions. *Wiley series in probability and statistics*. John Wiley & Sons.

[57] McNeeley, C. A., Nonnemaker, J. M., & Blum, R. W. (2002), Promoting school connectedness: Evidence from the national longitudinal study of adolescent health. *Journal of School Health*, 72(4), 138146.

[58] Meng, X.(1994), Multiple-imputation inference with uncongenial sources of input. *Statistical Science*, 9, 538-557.

[59] Meng, X. & Rubin, D.B. (1993), Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, 80, 267-278.

[60] Meng, X. & van Dyk, D.A. (1997), The EM Algorithm–An Old Folk-Song Sung to a Fast New Tune. *Journal of the Royal Statistical Society. Series B (Methodological)*, 59, 511-567.

[61] Meng, X. (1998), Fast EM-type implementations for mixed effects models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 60, 559-578.

[62] Molenberghs G & Kenward M (2007), Missing Data in Clinical Studies. John Wiley & Sons Ltd.

[63] Multilevel Data Analysis. http://www.nur.utexas.edu/researchlab/papers.

[64] Muthén, L.K. & Muthén, B.O. (1998), *Mplus user's guide* [Computer software manual]. Los Angeles: Muth'e n, L.K. & Muthén.

[65] Olkin, I. & Tate, R.F. (1961), Multivariate correlation models with mixed discrete and continuous variables, *Annals of Mathematical Statistics*, 32, 448465.

[66] Paschall M. J., Freisthler B., & Lipton R. I. (2005), Moderate alcohol use and depression in young adults: findings from a national longitudinal study, *American Journal of Public Health*, 95, 453-457.

[67] Pearson, K. (1990), Mathematical Contributions to the Theory of Evolution. VII. On the Correlation of Characters not Quantitatively Measurable. *Philosophical Transactions of the Royal Society of London*, Series A, 195, 1-47.

[68] R development Core Team (2005), R: A language and environment for statistical computing, R Foundation for statistial computing, ISBN3-900051-07-0, http://www.R-project.org.

[69] Raghunathan, T.E., Lepkowski, J. M., Hoewyk J. V. & Solenberger, P. (2001), A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models. *Survey Methodology*, 27, 85-95.

[70] Raghunathan T.E., Solenberger PW & Van Hoewyk J. (2000) IVEware: Imputation and Variance Estimation Software Installation instruction and User Guide. University of Michigan, Ann Arbor, MI. Software.

[71] Raghunathan, T.E. (2006), Combining information from multiple surveys for assessing health disparities. *Allgemeines Statistisches Archiv*. 90, 515526.

[72] Raudenbush, S.W. & Bryk, A.S. (2002), Hierarchical Linear Models: Applications and Data Analysis Methods (2nd ed.), Sage Publications Inc.

[73] Rasbash, J., Steele, F., Browne, W. & Prosser, B. (2004), A users guide to MLwiN (version 2.0), London: Institute of Eductation, 20 Bedford Way, http://www.cmm.bristol.ac.uk/MLwiN/download.

[74] Raudenbush, S., Bryk, A., Cheong, Y.F., & Congdon, R. (2004), HLM 6: Hierarchical Linear and Nonlinear Modeling. Lincolnwood, IL: Scientific Software International.

[75] Ritter, C. & Tanner, M.A. (1992), Facilitating the Gibbs sampler: The Gibbs stopper and the Giddy-Gibbs sampler. *Journal of the American Statistical Association*, 87, 861-868.

[76] Roth, P.L., & Switzer, F.S. (1999), Missing data: Instrument-level heffalumps and item-level woozles. Retrieved April 26, 2004, available at http://division.aomonline.org/rm/1999_RMD_Forum_Missing_Data.htm.

[77] Royston, P. (2005), Multiple imputation of missing values: update. *The Stata Journal*, 5, 188201.

[78] Rubin, D.B. (1974), Characterizing the estimation of parameters in incomplete data problems. *Journal of the American Statistical Association*, 69, 467-474.

[79] Rubin, D.B. (1982), EM algorithm for ML factor analysis. *Psycholmetrika*, 47, 69-76.

[80] Rubin, D.B. (1987a), Multiple imputation for nonresponse in surveys, J. Wiley & Sons, New York.

[81] Rubin, D. B. (1987b), Commenta noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fraction of missing information are modest: the SIR algorithm. *Journal of the American Statistical Association*, 82,543546.

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

[83] Schafer, J.L., Khare, M. & Ezzati-Rice, T.M. (1993), Multiple imputation of missing data in NHANES III. *Proceedings of the Annual Research Conference*, Bureau of the Census, Washington, DC, 459487.

[84] Schafer, J. L. (1994), Comment on Multiple-imputation inferences with uncongenial sources of input, by X.L. Meng, *Statistical Science*, 9, 560561.

[85] Schafer, J. L. (1997), Analysis of Incomplete Multivariate Data, Chapman & Hall, New York.

[86] Schafer, J.L. (1998a), Some improved procedures for linear mixed models. *Technical Report*, University Park: Pennsylvania State University, the Methodology Center.

[87] Schafer, J. L., and J. W. Graham. (2002), Missing data: Our view of the state of the art. *Psychological Methods*, 7:2, 147177.

[88] Schafer, J.L. & Olsen, M.K. (1998b), Multiple imputation for multivariate missing-data problems: a data analyst's perspective, *Multivariate Behavioral Research*, 33(4), 545-571. University Park: Pennsylvania State University.

[89] Schafer, J.L. (1999a), Multiple imputation: A primer. *Statistical Methods in Medical Research*, 8, 3-15.

[90] Schafer, J.L. (1999b), NORM: Multiple imputation of incomplete multivariate data under a normal model [Computer software]. University Park: Pennsylvania State University, Department of Statistics.

[91] Schafer, J.L. (2001), Multiple imputation with PAN. In A. G. Sayer & L. M. Collins (Eds.), New methods for the analysis of change (pp. 355377), Washington, DC: American Psychological Association.

[92] Schafer, J.L. (2005), Missing data in longitudinal studies: A review. Presented at the annual meeting of the American Association of Pharmaceutical Scientists, Nashville.

[93] Schafer, J.L. (2008) NORM: Analysis of incomplete multivariate data under a normal model, Version 3. Software package for R. University Park, PA: The Methodology Center, The Pennsylvania State University.

[94] Schafer, J. L., and Yucel, R. M. (2001), PAN: Multiple imputation for multivariate panel data, software for Windows 95/98/NT. Available at http://www.stat.psu.edu/jls/misoftwa.html.

[95] Schafer, J. L. & Yucel, R.M. (2002), Computational strategies for multivariate linear mixed-effects models with missing values. *Journal of Computational and Graphical Statistics*, 11, 437-457.

[96] Schimert, J., Schafer, J. L., Hesterberg, T., Fraley, C., & Clarkson, D. (2001), Analyzing missing values in SPLUS. Seattle, WA: Insightful.

[97] Smith, A. F. M. & G. O. Roberts (1993), Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods, *Journal of the Royal Statistical Society*, B, 55, 3-24.

[98] Song, J. & Belin, T. (2004), Imputation for incomplete high-dimensional multivariate normal data using a common factor model, *STATISTICS IN MEDICINE*, 23, 28272843.

[99] Song, J. & Belin, T. (2008), Choosing an appropriate number of factors in factor analysis with incomplete data, *Computational Statistics & Data Analysis*, Vol. 52, No. 7, 3560-3569.

[100] SPSS Advanced Models, Version 13.0 [software package], SPSS, Inc. Chicago, IL, 2004.

[101] SPSS, Inc., Linear Mixed-Effects Modelling in SPSS: An introduction to the MIXED procedure. *SPSS Technical Report LMEMWP-1002*, Chicago, IL, 2002.

[102] StataCorp LP, Stata Statistical Software: Release 10 [software package], College Station, TX: StataCorp LP, 2007.

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

[104] Tierney, L. (1994), Markov chains for exploring posterior distributions (with discussion), *Annals of Statistics*. 22, 1701-1762.

[105] Thomas, N., Raghunathan, T.E., Schenker, N., Katzoff, M.J. & Johnson, C.L. (2006), An Evaluation of Matrix Sampling Methods Using Data from the National Health and Nutrition Examination Survey. *Survey Methodology*, 32, 217-231.

[106] Thurstone, L.L. (1947), Multiple-factor analysis. Chicago, IL: University of Chicago Press.

[107] Van Buuren, S. & Oudshoorn, C.G.M. (2000), Multivariate Imputation by Chained Equations: MICE V1.0 User's Manual. *TNO Report* PG/VGZ/00.038.

[108] Walls, T. A., Jung, H. & Schwartz, J. E. (2006), Multilevel models for intensive longitudinal data, *Models for Intensive Longitudinal Data*, E.d. Theodore A. Walls, Joseph Schafer(Eds.), New York: Oxford University Press Inc, 3-37.

[109] West, B., Welch, K., Galecki, A.T. & Gillespie, B.W. (2007), Linear mixed models: A practical guide using statistical software. *CRC Press*.

[110] Wright, S. P.. Multivariate analysis using the MIXED procedure. http://ssc.utexas.edu/docs/sashelp/sugi/23/Stats/p229.pdf.

[111] Yuan, Y. C. (2000), Multiple imputation for missing data: Concepts and new development. In Proceedings of the Twenty-Fifth Annual SAS Users Group International Conference (Paper No. 267), Cary, NC: SAS Institute.

[112] Yucel, R. M. (2008), Multiple imputation inference for multivariate multilevel continuous data with ignorable non-response. *Philosophical Trasactions of the royal society A* 366, 23892403. (doi:10.1098/rsta.2008.0038),

[113] Zhang, X., Boscardin, W. J., Belin T.R. (2006), Sampling Correlation Matrices in Bayesian Models With Correlation Latent Variables, *Journal of Computational and Graphical Statistics*, Vol. 15, No.4, pp.880-896.

[114] Zhang, X., Boscardin, W.J. & Belin, T.R. (2008), Bayesian analysis of multivariate nominal measures using multivariate multinomial probit models. *Computational Statistics and Data Analysis*, DOI: 10.1016/j.csda.2007.12.012.

[115] Udry, J. R. (2003), The National Longitudinal Study of Adolescent Health (Add Health), Waves I & II, 1994-1996; Wave III, 2001-2002 [machine-readable data file and documentation]. Chapel Hill, NC: Carolina Population Center, University of North Carolina at Chapel Hill.

[116] Woods, C. (2006), Careless Responding to Reverse-Worded Items: Implications for Confirmatory Factor Analysis. *Journal of Psychopathology and Behavioral Assessment*, Vol. 28, No.3, pp. 189-194.

[117] Zhao, J. (2009) PAN: Multiple imputation for multivariate panel or clustered data. Software package for R. available at: http://cran.r-project.org/web/packages/pan/pan.pdf.

Vita

Rong Liu

Rong Liu is a native of Jinan, Shandong Province of China. In 2002, after graduating from Hohai University, she came to the United States to continue pursuing her advanced degrees. At the University of Toledo she earned a Master of Statistics. In 2004, she went to the Georgia Institute of Technology for her Ph.D in bioinformatics. In 2005, she transferred to the Pennsylvania State University and became a Ph.D student in Statistics.