ANALYSIS OF INCOMPLETE HIGH-DIMENSIONAL MULTIVARIATE NORMAL DATA USING A COMMON FACTOR MODEL

Juwon Song and Thomas R. Belin, UCLA
Juwon Song, 10920 Wilshire Blvd., Suite 300, Los Angeles, CA, USA, 90024-6505

Key words: multiple imputation, missing data, factor analysis, mental health

Abstract

It is common in applied research to have large numbers of variables measured on a modest number of cases. Even with small rates of missingness on individual variables, such data sets can have a large number of incomplete cases. As a result, complete-case analysis can lead to substantial loss of efficiency even if the missing data are missing completely at random and can lead to substantial bias even if the missing data are missing at random. Here we present a new method for handling missing continuously scaled items in multivariate data sets based on a common factor model for reducing the number of covariance parameters to be estimated in a multivariate normal model. The technique is illustrated using applications from mental health research.

1. Introduction

When the number of variables is large relative to the number of cases, even a small number of missing items on each variable can result in a large number of incomplete cases. For example, with 20 variables on 100 cases, if 10% of the values on each variable are randomly missing, we would expect only about 12 cases ($0.9^{20} \cong 0.12$) to be completely observed on all variables. As is now well known, complete-case analysis can be inefficient when missing data are missing completely at random (MCAR) and potentially biased when missing data are missing at random (MAR) (Little and Rubin 1987; Little 1992).

Multiple imputation offers an attractive alternative to complete-case analysis in that it can represent uncertainty due to missingness. It is valid under MAR when imputations are "proper" as defined by Rubin (1987). When missing data are MAR and the parameters of the data model and missing data mechanism are distinct, the missing data mechanism is said to be "ignorable" (Rubin 1987).

When we apply multiple imputation, it is recommended to include available information to the fullest extent possible because systematic difference between completely and partially observed cases may be reduced by incorporating important covariate information (Meng 1994; Rubin 1996). However, when the sample size is modest, even a simple model can be overparameterized. For example, if we observe 50 variables, $50 \times 49/2 = 1225$ correlation parameters would need to be estimated in a multivariate normal model with a general covariance matrix. Moreover, sometimes several variables are closely related to one another, which can cause problems with parameter estimation. In such cases, analysis often proceeds with an arbitrary choice of variables to include or exclude.

Schafer (1997a) introduced a method to handle possible overparameterization using a ridge prior distribution for a multivariate normal data model. The ridge prior is a limiting case of the normal inverted-Wishart prior. When data Y follows $N(\mu, \Sigma)$, a normal inverted-Wishart prior for the mean μ and variance Σ is implied by the specification that $\mu \sim N(\mu_0, \tau^{-1}\Sigma)$ and $\Sigma \sim W^{-1}(m, \Lambda)$. When $\tau \to 0$, the resulting prior distribution is called ridge prior by analogy with ridge regression. Posterior distributions are then given by

$$\mu \mid \Sigma, Y \sim N\left(\bar{y}, \frac{1}{n}\Sigma\right),$$

and

$$\Sigma \mid Y \sim W^{-1}\left(n+m, \left(\Lambda^{-1}+nS\right)^{-1}\right),$$

where $\bar{y}$ is the sample mean vector of Y and S is the sample covariance matrix of Y. When we standardize the data, a common choice of Λ^{-1} is $\Lambda^{-1} = \varepsilon I$, for $m = \varepsilon > 0$ and an identity matrix I. Then, the prior smoothes the sample correlation matrix toward an identity matrix. Schafer (1997a) showed that small positive values of ε work well to stabilize parameter estimates.

The present paper proposes multiple imputation based on a common factor model to reduce the dimension of the parameters in a multivariate normal model. We used the Gibbs sampler (Geman and Geman 1984) to draw parameter estimates, factor scores, and missing items. Based on the assumed factor structure, it is straightforward to randomly draw the means, factor loadings, uniqueness, factor scores as well as missing items from conditional distributions with other parameters fixed.

In Section 2, we describe multiple imputation based on a common factor model in detail. In Section 3, simulation results demonstrate that multiple imputation with a sufficient number of factors produces little bias under a variety of conditions. The application of this method to an emergency room intervention study in

Section 4 suggests that the proposed method has potential advantages in more complicated settings such as with longitudinally measured data. Finally, we discuss future directions of this research in Section 5.

2. Method

The idea of the factor model is to ignore factors corresponding to small eigenvalues and reduce the dimension of parameters. To be precise, we denote a data set as a matrix Y with n rows and p columns, where n represents the number of observations and p represents the number of variables. Then, Y_i, $i = 1, 2, \ldots, n$, denotes the ith observation of Y representing an iid random draw from an underlying sampling distribution. When we denote observed elements of Y as Y_{obs} and unobserved items as Y_{mis}, the data matrix Y can be written as $Y = (Y_{obs}, Y_{mis})$. The factor model with k underlying factors can be described as

$$Y_i = \alpha + Z_i \beta + \varepsilon_i,$$

where

α is a $1 \times p$ mean vector;
Z_i is a $1 \times k$ factor score vector;
β is a $k \times p$ factor-loading matrix;
$\varepsilon_i \sim N(0, \tau^2)$
where $\tau^2 = diag(\tau_1^2, \tau_2^2, \ldots, \tau_p^2)$;
and Z_i and ε_i are independent.

Rubin and Thayer (1982) discuss the EM algorithm for maximum likelihood factor analysis when there are no missing items. They considered factor scores as missing items and showed that EM algorithm can be used to calculate maximum likelihood estimates. Little and Rubin (1987) hint at an approach to handle missing items in the factor analysis, and Jamshidian (1997) introduced the EM algorithm for factor analysis when the data include missing items.

Markov chain Monte Carlo techniques such as data augmentation (Tanner and Wong 1987) and Gibbs sampling can be applied in multivariate incomplete data problems to multiply impute missing items as well as to estimate parameters. The complete-data likelihood function under the factor model can be expressed as

$$\begin{aligned} & L(\alpha, \beta, \tau \mid Y, Z) \\ &= \prod_{i=1}^{n} f(Y_i, Z_i \mid \alpha, \beta, \tau) \\ &= \prod_{i=1}^{n} f(Y_i \mid Z_i, \alpha, \beta, \tau) \cdot f(Z_i \mid \alpha, \beta, \tau) \\ &\propto \prod_{j=1}^{p} \left(\tau_j^2\right)^{-\frac{n}{2}} \cdot exp\left(-\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{p} \frac{\left(Y_{ij} - \alpha_j - Z_i \beta_j\right)^2}{\tau_j^2} \right) \\ &\quad \cdot exp\left(-\frac{1}{2} \sum_{i=1}^{n} Z_i Z_i' \right) \end{aligned}$$

For the prior distribution of τ_j^2, $j = 1, 2, \ldots, p$, we assume an inverse gamma distribution $IG\left(\frac{v_i}{2}, \frac{b_i}{2}\right)$. Although other prior descriptions would be possible, we prefer inverse-gamma prior distribution due to its convenience as a conjugate prior. We also assume conjugate prior distributions for α_j and β_j, namely:

$$\alpha_j \left| \tau_j^2 \right. \sim N\left(\alpha_0, \frac{1}{n_\alpha} \tau_j^2 \right), \quad \text{for } j = 1, 2, \ldots, p$$

and

$$\beta_j \left| \tau_j^2 \right. \sim N\left(\beta_0, \frac{1}{n_\beta} \tau_j^2 I_k \right) \quad \text{for } j = 1, 2, \ldots, p.$$

When n_α and n_β equal zero, these priors become non-informative priors for α_j and β_j, respectively.

With these specifications, the posterior distribution of model paramters becomes:

$$\begin{aligned} & P(\alpha, \beta, \tau \mid Y, Z) \\ &= L(\alpha, \beta, \tau \mid Y, Z) \cdot \prod_{j=1}^{p} IG(\tau_j^2 \mid \frac{v_i}{2}, \frac{b_i}{2}) \\ &\quad \cdot \prod_{j=1}^{p} N\left(\alpha_0, \frac{\tau_j^2}{n_\alpha} \right) \cdot \prod_{j=1}^{p} N\left(\beta_0, \frac{1}{n_\beta} \tau_j^2 I_k \right) \\ &\propto \prod_{j=1}^{p} \left(\tau_j^2\right)^{-\frac{n}{2}} exp\left(-\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{p} \frac{(Y_{ij} - \alpha_j - Z_i \beta_j)^2}{\tau_j^2} \right) \\ &\quad \cdot exp\left(-\frac{1}{2} \sum_{i=1}^{n} Z_i Z_i' \right) \\ &\quad \cdot \prod_{j=1}^{p} \left(\tau_j^2\right)^{-\frac{v_j+2}{2}} exp\left(-\frac{1}{2} \sum_{j=1}^{p} \frac{b_i}{\tau_j^2} \right) \\ &\quad \cdot \prod_{j=1}^{p} \left(\frac{\tau_j^2}{n_\alpha} \right)^{-\frac{1}{2}} exp\left(-\frac{1}{2} \sum_{j=1}^{p} \frac{n_\alpha \left(\alpha_j - \alpha_0\right)^2}{\tau_j^2} \right) \\ &\quad \cdot \prod_{j=1}^{p} \left(\frac{1}{n_\beta} \tau_j^2 I_k \right)^{-\frac{1}{2}} exp\left(-\frac{1}{2} \sum_{j=1}^{p} \frac{n_\beta \left(\beta_j - \beta_0\right)^2}{\tau_j^2} \right) \end{aligned}$$

The Gibbs sampler can successfully simulate the mean α, factor loadings β, and uniqueness terms τ_j^2 as well as factor scores Z and missing items as follows:

(1) Simulate missing items from

$$\begin{aligned} & Y_{i(mis)} \left| Y_{i(obs)}, \alpha, \beta, \tau^2 \right. \\ &\sim N\left(a_{mis.obs} + b_{mis.obs} Y_{obs}, \Sigma_{mis.obs} \right) \\ &\text{for } i = 1, 2, \ldots, n, \end{aligned}$$

where $a_{mis.obs}$ is an $1\times(p\text{-}p_1)$ intercept vector of the regression of Y_{mis} on Y_{obs} and
$b_{mis.obs}$ is a $p_1\times(p\text{-}p_1)$ slope matrix of the regression of Y_{mis} on Y_{obs} and
$\Sigma_{\text{mis.obs}}$ is a residual matrix of the regression of Y_{mis} on Y_{obs},
when p_1 variables are observed and $p\text{-}p_1$ variables are not observed.

(2) Simulate factor scores from

$$Z_i \mid Y_{i(obs)}, Y_{i(mis)}, \alpha, \beta, \tau^2 \sim N\left(Y_i\left(\left(\tau^2+\beta'\beta\right)^{-1}\beta'\right), I-\beta\left(\tau^2+\beta'\beta\right)^{-1}\beta'\right)$$

for $i = 1, 2, \ldots, n$.

(3) Simulate the uniqueness terms from

$$\tau_j^2 \mid Y_{obs}, Y_{mis}, Z \sim IG\left(\frac{n+v_i}{2}, \frac{1}{2}\sum_{i=1}^{n}\left(\left(Y_{ij}-\bar{Y}_j\right)-\left(Z_i-\bar{Z}\right)\cdot\left(\sum_{i=1}^{n}\left(Z_i-\bar{Z}\right)'\left(Z_i-\bar{Z}\right)\right)^{-1}\cdot\left(Z_i-\bar{Z}\right)'\left(Y_{ij}-\bar{Y}_j\right)\right)^2+\frac{b_i}{2}\right)$$

for $j = 1, 2, \ldots, p$. To avoid slow convergence due to high correlation between α and β (Gilks, Richardson and Spiegelhalter 1996), α and β were transformed to $\alpha' = \alpha + \bar{Z}\beta$ and $\beta' = \beta$.

(4) Simulate the mean estimates from

$$\alpha_j' \mid \tau_j^2, Y_{obs}, Y_{mis}, Z \sim N\left(\frac{1}{n+n_\alpha}\left(n\cdot\bar{Y}_j+n_\alpha\cdot\alpha_0'\right), \frac{1}{n+n_\alpha}\tau_j^2\right)$$

for $j = 1, 2, \ldots, p$,
where $\alpha_0' = \alpha_0 + \bar{Z}\beta$.

(5) Simulate the factor loadings from

$$\beta_j' \mid \tau_j^2, Y_{obs}, Y_{mis}, Z \sim N\left(\left[\sum_{i=1}^{n}\left(Z_i-\bar{Z}\right)'\left(Z_i-\bar{Z}\right)+n_\beta\right]^{-1}\cdot\left(\sum_{i=1}^{n}\left(Z_i-\bar{Z}\right)'\left(Y_{ij}-\bar{Y}_j\right)+n_\beta\cdot\beta_0\right), \left[\sum_{i=1}^{n}\left(Z_i-\bar{Z}\right)'\left(Z_i-\bar{Z}\right)+n_\beta\right]^{-1}\tau_j^2\right)$$

for $j = 1, 2, \ldots, p$.
Then α_j' is transformed back to the original α_j by $\alpha_j = \alpha_j' - \bar{Z}\beta_j$.

This model assumes that the number of factors, k, is known in advance. The effect of various choices of k is considered in the simulation.

The convergence of Gibbs sampler can be monitored by the method of Gelman and Rubin (1992) based on multiple starting values chosen after exploration of the likelihood surface using EM algorithm. Rubin and Thayer (1982, 1983) note the possibility of multiple modes in factor-model likelihoods. When there exist several modes in the likelihood, the Gibbs sampler may not mix values across separate regions of appreciable posterior density. However, posterior distributions can be simulated by drawing values from different chains with probability proportional to the posterior density of values.

Multiple imputation results in $m \geq 2$ complete data sets. The standard complete-case analysis pretending imputed values are observed can be applied to each imputed data set, and the results of these analyses can be combined to obtain an overall inference (Rubin 1987).

3. Simulation

A simulation study was carried out to evaluate the bias and coverage when the factor model is correct, overparameterized, or underparameterized. We chose a simple factor structure only with high loadings (0.8) and zero loadings (0). An example for the case of $k = 5$ is as follows:

$$\beta = \begin{bmatrix} 0.8 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0.8 & 0 & 0 & 0 & 0 \\ 0 & 0.8 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0.8 & 0 & 0 & 0 \\ 0 & 0 & 0.8 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & 0 \\ 0 & 0 & 0.8 & 0 & 0 \\ 0 & 0 & 0 & 0.8 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0.8 & \vdots \\ 0 & 0 & 0 & 0 & 0.8 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0.8 \end{bmatrix}.$$

Data were assumed to follow a multivariate normal distribution with a mean 0, variance 1, and covariances determined from the factor structure.

Based on the assumed factor structure, Table 3.1 shows the combinations of conditions used in the simulation study.

Table 3.1: Combinations of the simulation

# of observations (n)	# of variables (p)	# of true factors	# of assumed factors
100	100	5	5,10
		10	5,10
500	100	5	5,10
		10	5,10

The conditions with 100 observations were chosen to represent a modest sample size, and those with 500 observations were chosen to represent a slightly larger sample size. The number of variables, 100, is a large number to have in a multivariate normal model but a realistic number to measure in applied investigations. When the number of true factors is five, The assumption of five factors represents the correct model and the assumption of ten factors represents an overparameterized model. In the same way, when the number of true factors is ten, an assumption that there are five factors present represents an underparameterized model and an assumption of ten factors denotes the correct model. We replicated each combination of simulation conditions seventy-five times, to produce an error standard deviation for 95% coverage statistics of 2.5%.

We explored three missing-data mechanisms. In the first missing data mechanism M1, the first 99 variables were missing 5% of the time completely at random, and the last variable was missing roughly 25% of the time according to a logistic regression model with normally distributed coefficients. Technically, this is an MAR mechanism, but because all of the correlations we simulated were positive and the coefficients of the logistic regression were around zero, we found that even complete-case analysis performed well with this mechanism. Our second missing data mechanism, M2, was similar except that the logistic regression coefficients were taken to be absolute values of normal variates. We also explored another missing data mechanism, M3, where missingness on each variable depended on the underlying values of two adjacent variables. Because the adjacent values could also be missing, this mechanism is technically nonignorable, but because the adjacent values are not missing very often, we characterized this mechanism as "close" to MAR. We use this term "close to MAR" loosely, only to mean that a procedure developed to handle MAR data might perform reasonably well with this mechanism.

When $n = 500$, multiple imputation based on the multivariate normal model performed well with non-informative priors for model parameters. However, when $n = 100$, informative priors were necessary for Gibbs sampler to work. For the ridge prior, $\varepsilon = 3$ was applied, and for the factor model, the same number of degrees of freedom was chosen for τ^2. When $n = 100$, the multivariate normal model using the ridge prior indicated unstable parameters and resulted in higher variance estimates.

The performances on cross-sectinal mean when $n = 100$ are in Tables 3.2-3.4. The first column shows the true number of factors and the second column shows the models we compare. In the factor model, the row "correct" represents the correct model, and the rows "over" and "under" represent the overparameterized model and the underparameterized model, respectively. The third and fourth columns represent Monte Carlo mean and the average length of 95% confidence intervals, respectively. For both the multivariate normal model and the factor model, the Monte Carlo mean and standard error were calculated based on multiple imputation inference (Rubin 1987). The last column represents the actual 95% coverage rate, measured by the number of data sets whose 95% confidence interval covers the true parameter value.

Table 3.2. The mean of the last variable under the missing data mechanism M1

k	Models	M. C. Mean	Ave. Int. Length	Act. 95% Cov.
5	True	0.0000		
	All Data	0.0141	0.3972	0.960
	Available-case	-0.0168	0.4580	0.933
	Normal	0.0164	0.5110	0.987
	Factor correct	0.0159	0.4491	0.947
	over	0.0141	0.4618	0.933
10	True	0.0000		
	All Data	0.0000	0.3928	0.987
	Available-case	-0.0068	0.4557	0.960
	Normal	0.0028	0.5173	0.933
	Factor correct	-0.0044	0.4469	0.960
	under	-0.0112	0.4789	0.947

Table 3.3. The mean of the last variable under the missing data mechanism M2

k	Models	M. C. Mean	Ave. Int. Length	Act. 95% Cov.
5	True	0.0000		
	All Data	0.0030	0.3966	0.960
	Available-case	-0.3170	0.3938	0.133
	Normal	-0.0083	0.5084	0.960
	Factor correct	-0.0224	0.4585	0.880
	over	-0.0328	0.4481	0.920
10	True	0.0000		
	All Data	-0.0074	0.3989	0.960
	Available-case	-0.2889	0.4114	0.200
	Normal	-0.0354	0.5023	0.973
	Factor correct	-0.0445	0.4492	0.920
	under	-0.2275	0.4347	0.440

Table 3.4. The mean of the last variable under the missing data mechanism M3

k	Models	M. C. Mean	Ave. Int. Length	Act. 95% Cov.
5	True	0.0000		
	All Data	-0.0016	0.4004	0.960
	Available-case	0.1386	0.4027	0.773
	Normal	0.0038	0.4388	0.987
	Factor correct	0.0056	0.4124	0.987
	over	0.0095	0.4131	0.987
10	True	0.0000		
	All Data	0.0005	0.3957	0.947
	Available-case	-0.1388	0.4004	0.693
	Normal	-0.0170	0.4291	0.947
	Factor correct	-0.0192	0.4072	0.947
	under	-0.0784	0.4042	0.827

The factor model performed well when the specified number of factors equaled or exceeded the true number of factors. Coverages ranged from 92.0 - 98.7% with one exception, where the coverage was 88.0% under k = 5 and missing data mechanism M2. However, factor models performed poorly when model is underparameterized, showing coverages from 44.0 - 94.7%. Even though multivariate normal model produced larger variance estimates, its coverages were good, ranging from 93.3 - 98.7%. On the other hand, its average interval lengths were longer than those for the factor model. Available-case analysis performed poorly overall with coverages ranging 13.3 - 96.0%. We did not include the result of complete-case analysis because the number of completely observed cases was always less than five when the sample size is 100. However, it is common that the imputation model is larger than the analysis model, so available-case analysis can be considered as complete-case analysis from the analyst's viewpoint.

Real data usually have a complicated factor structure, and the number of factors is unknown in advance in most cases. Therefore, to depict this situation, next simulation was based on an observed covariance matrix. We generated 200 data sets from a multivariate normal distribution with mean and covariance matrix fixed at published values from a study of 24 psychological tests on 145 school children (Harman 1967). The number of factors was unknown, but earlier analyses for these data suggested four, five, or seven factors (Harman 1967; Velicer 1976). We also considered eleven factors based on the cumulative variance explained (80.2%) and because we desired not to underparameterize the model. Table 3.5 shows the performance of the cross-sectional mean of the last variable under the missing data mechanism M1. All factor models performed well, showing good coverage rates. The multivariate normal model also performed well, but available-case analysis was more biased, showing low coverage.

Table 3.5: The mean of the last variable under the missing data mechanism M1.

Models	M. C. Mean	Ave. Int. Length	Act. 95% Coverage
True	25.8300		
All Data	25.8116	1.5385	0.955
Available-case	25.2779	1.7553	0.775
Normal	25.8293	1.8820	0.935
Factor 4-factor	25.9646	1.8012	0.930
5-factor	25.9220	1.8473	0.940
7-factor	25.8962	1.8420	0.940
11-factor	25.8710	1.8558	0.925

4. Application to emergency room intervention data

In another applied research setting, 140 female adolescents were recruited after a suicide attempt to provide information on psychosocial variables in a longitudinal study that tracked subjects beyond a period of follow-up counseling. During the middle of the study, an intervention was implemented in an effort to improve emergency room procedures. The first 75 adolescents therefore received a standard emergency room treatment, and the next 65 adolescents received a specialized emergency room intervention. The specialized emergency room intervention included education of emergency room staff, meeting with bilingual crisis therapists, and video session showing what patients and their families could expect during the follow-up treatment process. It was hypothesized that the specialized emergency room intervention would

improve subsequent psychological outcomes, perhaps in part by resulting in better attendance by adolescents and their caregivers (generally their parents) in counseling session attendance.

Brief demographic assessments as well as mental status exams were gathered at the emergency room. The first baseline assessment asked questions of both adolescents and their caregivers and was obtained after the hospital discharge of the adolescent. Assessments were repeated after three, six, twelve and eighteen months. Interest focused on the effectiveness of the specialized emergency room intervention and the relationship between baseline psychological impairment and outcomes over time. Analyses were performed on 27 outcome variables with several baseline characteristics used as covariates. Most variables had 5-25% missing values, although a few showed 50-60% of missing items (see Table 4.1).

For multiple imputation based on the factor model, we considered each longitudinally measured variable as potentially representing a separate factor. For example, the Beck depression inventory was measured on adolescents at baseline, three, six, twelve and eighteen months, which were entered into the data set as five separate variables. Similarly, the adolescent impulsiveness scale was not measured at twelve months, so that it was reflected as four separate variables in the data set. The ensemble of 27 outcome variables, baseline characteristics including impairment status, and intervention status amounted to 135 variables overall.

In applying the factor model, we chose a 30-factor model. This model explained roughly 80% of the total variation in the original data set and was general enough to allow each longitudinal variable to represent a separate factor. Another motivation for choosing a model with 30 factors was the insight from the simulation study that overparameterization is not a great concern from the vantage point of bias and coverage, while a model with an insufficient number of factors can result in serious bias.

Rubin and Thayer (1982, 1983) warned that it is possible to have multiple modes in the factor model. However, in the high-dimensional data, it is not easy to find starting values covering the whole posterior density space. Therefore, we chose starting values from a run of Gibbs sampler starting from the mode. After 1500 iterations of the Gibbs sampler, we stored maximum and minimum values of each parameter and chose fifty random combinations of them as starting values. It turned out that 27 starting values ended up at a mode with the highest likelihood. Seven converged to another mode with slightly smaller likelihood. We also found five other local modes with smaller likelihoods.

When there are several modes, the Gibbs sampler may not reflect values across separate regions of posterior density. However, a posterior distribution can be represented by drawing different chains with a probability proportional to the posterior density of the values. Among ten starting values of Gibbs sampler, seven were chosen from the major mode and three were from the mode with the second largest likelihood.

Intervention status and family type (single versus multiple adults in the household) were measured as dichotomous variables. Impairment status was measured as three categories which we have labeled low, moderate, and high impairment. The number of sessions attended, acculturation measures, and many outcome measures had limited ranges. For binary variables, we viewed imputed values as a probability. For example, if the categories are coded 1 or 2 and an imputed value is 1.3, we randomly draw a value from Bernoulli distribution with a probability 0.3 and impute category 1 if the drawn value is 0 and impute category 2 if the drawn one is 1. For the variables with limited ranges, values outside the range were truncated.

After we generated ten multiply imputed data set, the analysis proceeded with longitudinal data analysis using SAS PROC MIXED. Rotheram et. al (1999) applied this data to the multiple imputation based on the linear growth-curve model using a program developed by J. L. Schafer (1997b) and reported that impairment status was significantly related with many outcomes but the effect of intervention was diluted after adjusting for the baseline impairment as well as covariates. No substantial differences were shown in multiple imputation based on the factor model, even though some sensitivity appeared in significance of impairment effect and interaction between intervention and impairments, as seen in Table 4.2.

5. Discussion

When data sets have large numbers of variables measured on modest number of cases, the complete-case analysis is inefficient and often biased. Multiple imputation with the factor model showed little bias with good coverages when we considered a large enough number of factors. As Schafer (1994) describes, there are many cases where an ideal imputation model requires many more parameters than we can estimate in the data, and multiple imputation using a common factor model can be a solution to overcome this difficulty.

We developed the model with an assumption of known number of factors. Simulation warned that the underparameterized model can cause a serious bias, but the overparamterized model does not have any serious problem. However, the overparameterized model requires more parameters to be estimated, requiring more computation time. Therefore, more specific

advice on determining an appropriate number of factors would be useful.

The factor model was developed based on the normality assumptions. Sensitivity analyses for categorical or non-normal continuous data were left for future research. A worthwhile extension would be to consider a factor model tailored to handle longitudinally measured variables. These methods presumably could also be extended to accommodate structural equation models, although the numerous criticisms of any causal interpretation of structual equation models (e.g., Freedman 1987; Rosenbaum 1995) leaves us less enthusiastic about pursuing this direction.

References

Freedman, D. (1987), "As Others See Us: A Case Study in Path Analysis" (with discussion), *Journal of Educational Statistics, 12*, 101-128.

Gelman, A. and Rubin, D. B. (1992), "Inference from Iterative Simulation Using Multiple Sequences" (with discussion), *Statistical Science*, 7, 457-511.

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

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (1996), *Markov Chain Monte Carlo in Practice*, London: Chapman and Hall.

Harman, H. H. (1967), *Modern Factor Analysis*, Chicago: The University of Chicago Press.

Jamshidian, M. (1997), "An EM Algorithm for ML Factor Analysis with Missing Data," In Berkane, M. (Ed.), *Latent Variable Modeling and Applications to Causality* (pp. 247-258), New York: Springer.

Little, R. J. A. (1992), "Regression with Missing X's: A Review," *Journal of the American Statistical Association*, **87**, 1227-1237.

Little, R. J. A., and Rubin, D. B. (1987), *Statistical Analysis with Missing Data*, New York: John Wiley.

Meng, X. L. (1994), "Multiple-Imputation Inferences with Uncongenial Sources of Input," *Statistical Science*, 9, 538-573.

Rosenbaum, P. R. (1995), Discussion of "Causal diagrams for empirical research" (Pearl, J.), *Biometrika*, **82**, 698-699.

Rotheram-Borus, M. J., Piacentini, J. C., Cantwell, C., Belin, T. R., and Song, J. (1999), "The Long-term Impact of an Emergency Room Intervention for Adolescent Suicide Attempters," submitted manuscript, UCLA Neuropsychiatric Institute.

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

Rubin, D. B. (1996), "Multiple Imputation After 18+ Years," *Journal of the American Statistical Association*, **91**, 473-489.

Rubin, D. B. and Thayer, D. T. (1982), "EM Algorithms for ML Factor Analysis," *Psychometrika*, **47**, 69-76.

Rubin, D. B. and Thayer, D. T. (1983), "More on EM for ML Factor Analysis," *Psychometrika*, **48**, 253-257.

Schafer, J. L. (1994), Comment on "Multiple-Imputation Inferences with Uncongenial Sources of Input" (Meng, X. L.), *Statistical Science*, **9**, 560-561.

Schafer, J. L. (1997a), *Analysis of Incomplete Multivariate Data*, New York: Chapman & Hall.

Schafer, J. L. (1997b), "Imputation of Missing Covariates Under a Multivariate Linear Mixed Model," unpublished technical report, Dept. of Statistics, Penn State University.

Tanner, M. A. and Wong, W. H. (1987), "The Calculation of Posterior Distributions by Data Augmentation" (with discussion), *Journal of the American Statistical Association*, **82**, 528-550.

Velicer, W. F. (1976), "The Relation Between Factor Score Estimates, Image Scores, and Principal Component Scores," *Educational and Psychological Measurement*, **36**, 149-159.

Table 4.1: Variables and percentage of missing items in each time point[1)]

		Baseline	3 Months	6 Months	12 Months	18 Months
Baseline Characteristics	Impairment	6.4	-	-	-	-
	Intervention	0.0	-	-	-	-
	Mother's Education	13.6	-	-	-	-
	Family Type	0.0	-	-	-	-
	Accultration : Adolescent	0.7	-	-	-	-
	Accultration : Parent	11.4	-	-	-	-
	Number of Treatment Attendence : Adolescent	0.0	-	-	-	-
	Number of Treatment Attendence : Parent	0.0	-	-	-	-
Adolescent Measure	Beck Depression Inventory	1.4	13.6	12.9	7.9	7.1
	Rosenberg Self-Esteem	6.4	17.1	12.9	7.1	7.1
	Impulsiveness Scale	0.0	12.9	17.1	-	7.1
	Hass Ideation Factor : Suicidal Ideation	0.0	16.4	12.9	7.1	7.1
	Hass Ideation Factor : Substance Use	0.0	16.4	12.9	7.1	7.1
	Frequency of sexual partners in past 3 months	13.6	-	14.3	57.9	52.1
	Number of sex in past 3 months	15.0	-	15.0	57.9	52.9
	Conduct Disorder	-	13.6	13.6	7.9	7.1
	Delinquency	20.0	-	19.3	7.9	7.9
	School Problem	6.4	-	19.3	7.9	7.9
	Maternal Caretaking	0.7	12.1	20.0	-	7.1
	Overprotectiveness	0.7	12.1	20.0	-	7.1
	Family Adaptability	0.7	15.7	12.9	7.1	7.1
	Family Cohesion	0.7	15.7	12.9	7.1	7.1
Parent Measure	Beck Depression Inventory	12.1	16.4	14.3	15.7	17.1
	BSI - General Severity Index	10.0	15.7	14.3	15.0	16.4
	BSI - Somatization	10.0	15.7	14.3	15.0	16.4
	BSI - Obsessive, Compulsive Behavior	10.0	15.7	14.3	15.0	16.4
	BSI - Interpersonal Sensitivity	10.0	15.7	14.3	15.0	16.4
	BSI - Depression	10.0	15.7	14.3	15.0	16.4
	BSI - Anxiety	10.0	15.7	14.3	15.0	16.4
	BSI - Hostility	10.0	15.7	14.3	15.0	16.4
	BSI - Phobic Anxiety	10.0	15.7	14.3	15.0	16.4
	BSI - Paranoid Ideation	10.0	15.7	14.3	15.0	16.4
	BSI - Psychoticism	10.0	15.7	14.3	15.0	16.4
	Family Adaptability	10.0	20.7	14.3	15.0	16.4
	Family Cohesion	10.0	20.7	14.3	15.0	16.4

1) - means that the variable was not measured at that time point.

Table 4.2: The result of linear mixed model regression for emergency room intervention data (parameter estimates and p-values in parenthesis)

		Adolescent	Parent					
		Depression	Depression (BDI)	Gen. Severity (BSI)	Anxiety (BSI)	Depression (BSI)	Adaptability (FACES III)	Cohesion (FACES III)
Intercept		5.389(<0.001)	7.358 (<0.001)	0.627 (0.011)	0.779 (0.003)	0.562 (0.020)	32.189 (<0.001)	44.702 (<0.001)
Time		-0.043 (<0.001)	-0.030 (<0.001)	-0.004 (0.091)$^{-}$	-0.003 (0.238)$^{-}$	-0.003 (0.291)$^{-}$	0.070 (0.121)	-0.037 (0.366)
Impairment	(M vs. L)	0.635 (0.018)$^{+}$	0.268 (0.312)	0.136 (0.199)	0.228 (0.047)$^{+}$	0.187 (0.080)	0.337 (0.834)	-1.106 (0.521)
	(H vs. L)	1.420 (<0.001)	1.250 (<0.001)	0.220 (0.041)	0.300 (0.009)	0.315 (0.004)	-0.405 (0.795)	-3.737 (0.017)$^{+}$
Intervention		-0.546 (<0.001)	-0.153 (0.481)	-0.039 (0.717)	-0.036 (0.749)	0.006 (0.953)	-0.781 (0.622)	-3.875 (0.023)
Mother's Education		-0.007 (0.961)	-0.269 (0.041)$^{+}$	0.016 (0.836)	0.013 (0.873)	-0.011 (0.890)	-0.853 (0.443)	-1.472 (0.208)
Family Type		-0.043 (0.733)	0.178 (0.213)	0.075 (0.235)	0.113 (0.088)$^{-}$	0.054 (0.385)	0.620 (0.535)	0.875 (0.387)
Accultration SA		0.003 (0.798)	-0.021 (0.106)	-0.011 (0.075)	-0.013 (0.038)$^{+}$	-0.010 (0.100)	-0.010 (0.915)	0.149 (0.131)
Accultration Mother		0.006 (0.653)	0.017 (0.246)	0.012 (0.147)	0.006 (0.429)	0.010 (0.238)	-0.274 (0.014)	-0.329 (0.006)
Impairment×Intervention								
	(M vs. L)	0.901 (0.033)$^{+}$	0.302 (0.423)	-0.010 (0.945)	0.002 (0.989)	-0.029 (0.840)	0.444 (0.840)	0.591 (0.797)
	(H vs. L)	0.284 (0.425)	-0.306 (0.266)	-0.199 (0.154)$^{-}$	-0.217 (0.128)$^{-}$	-0.273 (0.037)	3.824 (0.093)	4.591 (0.062)

+ means that it became significant at α=0.05 but it was not significant in Rotheram et. al. (1999).

- means that it became not significant but it was significant at α=0.05 in Rotheram et. al. (1999).