Multiple Imputation for High-Dimensional Mixed Incomplete Data Using a Factor Model

Ren He Thomas Belin
Department of Biostatistics
University of California Los Angeles

September 11, 2012

Abstract

One strategy for producing imputations for high-dimensional incomplete data is to model associations among variables using a factor-analysis framework, thereby avoiding concerns with a more general association structure where some parameters are poorly estimated. Song and Belin (2004) pursued such a strategy for continuous outcomes; here we propose a similar strategy allowing for mixed data types (continuous, binary, ordinal and nominal). We describe an MCMC approach for fitting the model, and our method is compared in several simulation settings to available-case analysis and a rounding method.

Keywords: Multiple Imputation (MI), factor analysis, data augmentation, MCMC , mixed data

1 Introduction

The simplest and most common way to handle incomplete data is to analyze only those cases with all variables observed. This approach, named complete-case analysis by Little and Rubin (2002). is easy to apply and is the default method in most statistical computing packages. However, it is common in applied research to have large numbers of variables measured on a modest number of cases. In this case, even a small number of missing items on each variable can result in a large number of incomplete cases. For example, with 20 variables on 100 cases, if 10 percent of the values on each variable are randomly missing, we would expect only about $100 \times 0.9^{20} \approx 12$ cases with complete records.

Multiple Imputation (MI) (Rubin 1987) is a technique for imputing $m \geq 2$ plausible values to reflect uncertainty about those missing items. When applying multiple imputation to incomplete data sets, it is recommended to include available information to the fullest extent possible because systematic differences between completely and partially observed cases may be reduced by incorporating important covariate information (Rubin 1996). However, when the sample size is modest, even a simple model can be overparameterized when the number of variables is moderately large. For example, for 50 variables, $50 \times 49/2 = 1225$ correlation parameters would need to be estimated in a multivariate normal model with a general covariance matrix. Moreover, sometimes several variables are closely related to one another, which can cause problems with model fitting. Schafer (1997) proposed a method to handle possible overparameterization using a ridge prior distribution under a multivariate normal model. The ridge prior is a limiting case of the normal inverted-Wishart prior. Little and Rubin (2002) hint at an approach to handle missing items in factor analysis, building on the framework of Dempster et al. (1977) and Rubin and Thayer (1982), where factor scores are viewed as missing data even when there are no missing items. Jamshidian (1997) explicitly described an EM algorithm for factor analysis when the data include missing items. Song and Belin (2004) introduced an imputation method using a common factor model. They use the Gibbs sampler to draw factor scores and missing items as well as parameter estimates. However, one can expect to have different types of variables in common applied settings, including continuous, binary, ordinal and nominal variables. The idea of developing methods for a joint model to accommodate multivariate data with mixed data types presents considerable challenges but would be valuable to applied researchers. Both ridge prior method and factor model method are not tailored to an incomplete data setting with mixed data types. The goal of this paper is to develop joint modeling strategy using a common factor model that will accommodate realistic data structures involving large numbers of mixed types variables and modest numbers of cases with general patterns of incomplete data.

In Section 2, we describe a procedure for multiple imputation based on a common factor model. In Section 3, simulation results are displayed to prove the validity of our model. In Section 4, we apply this method to an health care survey research Finally, we discuss future directions of this research in Section 5.

2 Method

2.1 Multivariate probit model

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

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

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

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

2.2 Factor model

For the purpose of fixing ideas, we assume here a scenario with only continuous and binary data, although the intention is to expand the idea to incorporate ordinal and nominal categorical data using multivariate probit and multivariate multinomial probit modeling techniques. Let $T_i^T = (v_i^T, c_i^T), i = 1, ..., n$ consists of a continuous proportion $v_i^T = (v_{i1}, ..., v_{ip_1})$ with length p_1 and a binary portion $c_i^T = (c_{i1}, ..., c_{ip_2})$ with length p_2, $p_1 + p_2 = p$. We treat the binary variables in the multivariate probit model framework in Section 2.1 Let z_i is the corresponding latent vector for c_i, z_i is a $p_2 \times 1$ vector. We divide v_i into two parts. $v_i = (v_{i,obs}, v_{i,mis})$, where $v_{i,obs}$ denotes the observed part of v_i, $v_{i,mis}$ denotes the missing part of v_i. Similarly, we can define $z_i = (z_{i,obs}, z_{i,mis})$ and $c_i = (c_{i,obs}, c_{i,mis})$. The factor model is:

$$y_i = \alpha + \phi_i\Lambda + \epsilon_i \tag{1}$$

where α is a $1 \times p$ intercept vector, Λ is a $k \times p$ factor loading matrix. ϕ_i is a $1 \times k$ factor score and $\phi_i \sim N(0, I)$, $\epsilon_i \sim_{iid} N(0, \tau)$, $\tau = diag(\tau_1^2, ..., \tau_p^2)$. Since here normal latent variables are used to accommodate binary or categorical data. Some constraints have to be added to the model to make sure the model is identifiable.

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

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

$$\alpha_j|\tau_j^2 \sim N(\alpha_{0j}, \frac{1}{n_\alpha}) \quad for\ j = p_1 + 1, ..., p \tag{3}$$

$$\Lambda_j|\tau_j^2 \sim N(\Lambda_{0j}, \frac{1}{n_\Lambda}\tau_j^2 I_k) \quad for\ j = 1, 2, ..., p_1 \tag{4}$$

$$\Lambda_j|\tau_j^2 \sim N(\Lambda_{0j}, \frac{1}{n_\Lambda} I_k) \quad for\ j = p_1 + 1, ..., p \tag{5}$$

where α_{0j} and Λ_{0j} are prior means, n_α and n_Λ can be viewed as additional prior degrees of freedom for inference about α and Γ respectively, and I_k is a $k \times k$ identity matrix.

When $n_\alpha \to 0$ and $n_\beta \to 0$, these distributions correspond to noninformative priors for α_j and β_j. Although we could generalize n_α and n_β to depend on j, we have not detected the necessities to choose different values for n_α and n_β based on different values of j.

2.3 Gibbs sampler for the factor model

With the specifications of prior information listed in Section 2.2, we can derive the following Gibbs sampler algorithm to simulate the intercept α, factor loadings Λ, and uniqueness terms τ^2 as well as factor scores Φ and missing items:

- Simulate the missing values of continuous variables from

$$v_{ij,mis}|v_{i,obs}, z_i, \alpha, \Lambda, \phi_i, \tau^2 \sim N(a_j, \tau_j^2), j \in \digamma_v(i) \tag{6}$$

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

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

$$z_{ij,mis}|v_i, z_{i,obs}, \alpha, \Lambda, \phi_i, \tau^2 \sim N(a_j, \tau_j^2), j \in \digamma_{z_1}(i) \tag{7}$$

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

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

$$z_{ij,obs}|v_i, z_{i,mis}, z_{iL,obs}, L \neq j, \alpha, \Lambda, \phi_i, \tau^2 \sim [I_{(z_{ij,obs}>0)}I_{(c_{ij}=1)} + I_{(z_{ij,obs}<=0)}I_{(c_{ij}=0)}] \times N(a_j, \tau_j^2), j \in \digamma_{z_2}(i) \tag{8}$$

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

- Simulate factor scores from

$$\phi_i|y_i, z_i, \alpha, \Lambda, \tau^2 \sim N((y_i - \alpha)(\Lambda'\Lambda + \tau^2)^{-1}\Lambda', I_k - (\Lambda'\Lambda + \tau^2)^{-1}\Lambda') \tag{9}$$
$$\tag{10}$$

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

- Simulate uniqueness terms from

$$\tau_j^2|y_i, z_i, \phi, \alpha, \Lambda \sim IG(\frac{n + \nu_j}{2}, \frac{b'_j}{2}),\ j = 1, ..., p_1 \tag{11}$$

- Simulate mean estimates from

$$\alpha_j^*|\tau_j^2, Y_{obs}, Y_{mis}, Z \sim N(\frac{n\bar{y_j} + n_\alpha\alpha_{0j}^*}{n + n_\alpha}, \frac{\tau_j^2}{n + n_\alpha}),\ j = 1, ..., p \tag{12}$$

where the the explicit formula for term b'_j can be found from Song and Belin (2004), I won't give the details here due to its complicated form.

- Simulate the factor loading from

$$\Lambda_j|\tau_j^2, Y_i, Z_i, \phi_i, \alpha \sim N((\sum_{i=1}^{n}(\phi_i - \bar{\phi})'(\phi_i - \bar{\phi}) + n_\beta I_k)^{-1}(\sum_{i=1}^{n}(\phi_i - \bar{\phi})'(Y_{ij} - \bar{Y_j})$$
$$+n_\beta\beta_{0j}),(\sum_{i=1}^{n}(\phi_i - \bar{\phi})'(\phi_i - \bar{\phi}) + n_\beta I_k)^{-1}\tau_j^2) \tag{13}$$

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

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

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

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

Multiple imputation results in $m > 2$ complete data sets. Standard complete-case analyses treating imputed values as known 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 Studies

In this chapter, we carry out a set of simulation studies to evaluate the validity of the two proposed approaches. The goal will be to recover parameter values used to generate the data based on inference from the incomplete data sets where a missingness pattern has been introduced.

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

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

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

$$\Lambda = \begin{pmatrix} 0.8 & ... & 0.8 & 0 & ... & 0 & 0 & ... & 0 \\ 0 & ... & 0 & 0.8 & ... & 0.8 & 0 & ... & 0 \\ ... & ... & ... & ... & ... & ... & ... & ... & ... \\ 0 & ... & ... & ... & ... & 0 & 0.8 & ... & 0.8 \end{pmatrix} \tag{14}$$

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

To represent a moderate or large sample size, we assumed that the sample size to be 100 or 300. Following the routine of section 6.1, we assume $p = 50$ variables are measured. The 50 variables are made up of 25 continuous and 25 binary variables. We also assume that true underlying factor structure includes 5 or 10 factors. In a real application, we usually don't know the correct number of factors, so it is possible

to use an incorrect number of factors in the model. Therefore, we can explore the performance of the factor model based on 10 factors applied to data generated by 5 true factors as well as the performance of the factor model based on 5 factors applied to data based on 10 true factors. These represent the case that our imputation model is underparameterized or overparameterized, respectively. Meanwhile, we can explore the performance of the factor model based on correct factor numbers as well. Then we explore two missing data mechanisms. In the first mechanism M1, the first 24 continuous variables $y_1, y_2, ..., y_{24}$ and the first 24 categorical variables $y_{26}, ..., y_{49}$ are missing 25% of the time completely at random, while y_{25} and y_{50} are missing according to a logistic regression model. Specifically, I assume:

$$
\begin{aligned}
&p(y_1 = missing) = 0.25 \\
&p(y_2 = missing) = 0.25 \\
&... \\
&p(y_{24} = missing) = 0.25 \\
&logit[p(y_{25}\ missing)] = l_0 + l_1 y_1 + ... + l_{24} y_{24} \qquad (15) \\
&among\ observed\ y_i's, i = 1, ...24 \\
&p(y_{26} = missing) = 0.25 \\
&... \\
&p(y_{49} = missing) = 0.25 \\
&logit[p(y_{.50}\ missing)] = r_0 + r_1 y_{25} + ... + r_{24} y_{49} \qquad (16) \\
&among\ observed\ y_i's, i = 1, ...49
\end{aligned}
$$

where $l_i, r_i, i = 1, ..., 24$ are drawn from N(0,1) and then fixed throughout the simulation. l_0 and r_0 are constants that can be used to adjust the missing rates of y_{25} and y_{50}. Here we choose l_0 and r_0 to assure that the missing rates of y_{25} and y_{50} are around 25%. Technically, this is an MAR mechanism, but because all of the correlations were positive and the coefficients of the logistic regression were distributed symmetrically around zero, we found that prediction errors in one direction tended to be canceled by prediction errors in the other direction, so that even complete-case analysis may perform well. The second missing data mechanism, M2, is similar to M1 except we use absolute values of normal random numbers to be the logistic

regression coefficients, that is:

$$
\begin{aligned}
& p(y_1 = missing) = 0.25 \\
& p(y_2 = missing) = 0.25 \\
& ... \\
& p(y_{24} = missing) = 0.25 \\
& logit[p(y_{25}\ missing)] = l_0 + |l_1|y_1 + ... + |l_{24}|y_{24} \qquad (17) \\
& among\ observed\ y_i's, i = 1, ...24 \\
& p(y_{26} = missing) = 0.25 \\
& ... \\
& p(y_{49} = missing) = 0.25 \\
& logit[p(y_{50}\ missing)] = r_0 + |r_1|y_{25} + ... + |r_{24}|y_{49} \qquad (18) \\
& among\ observed\ y_i's, i = 1, ...49
\end{aligned}
$$

where $l_i, r_i, i = 1, ..., 24$ are drawn from N(0,0.5) and then fixed throughout the simulation. We take the absolute values of $l_i's$ and $r_i's$ to avoid a canceling effect across variables. As before, l_0 and r_0 are constants that can be used to adjust the missing rates of y_{25} and y_{50} to be around 25%. All $l_i's$ and $r_i's$ are fixed throughout the simulation process. The following table shows the combinations used in the simulation study.

Table 1: Combinations of the simulation

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

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

Due to the space limit, we only show part of the simulation outputs. But all the simulation outputs give similar conclusion. Table 2 shows the result of inferences

on the means of y_{25} and y_{50} under the factor model with n=300, p=50, k=5.

Table 2: The means of y_{25} and y_{50} under the factor model with n=300, p=50, k=5, and missing data mechanism M1

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

Table 3 show results of inferences on the means of y_{25} and y_{50} under the factor model with n=300, p=50, k=10. Compared with Table 6.19 and Table 6.20, we can find an overparameterized model (Table 6.19,6.20) results in little bias for mean estimates of y_{25} and y_{50} but an underparameterized model (Table 6.21,6.22) results in more biased mean estimates with lower than the nominal 95% coverage rate when we apply the incorrect number of factors in our model. But the underparameterized factor model still has better behaviors than the rounding method on the inference of y_{50}.

Table 3: The means of y_{25} and y_{50} under the factor model with n=300, p=50, k=10, and missing data mechanism M2

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

Table 4 shows inferences about the means of y_{25} and y_{50} under the factor model with n=100, p=50, k=5. The standard errors are about two times of those with sample size 300. Under the missing data mechanism M1, all methods even available-case analysis show small biases and good 95% coverage probabilities on the inference

about the mean of y_{25}. That again reveals the "close to MCAR" property of the missing data mechanism M1. But for the inference of y_{50}, both available-case analysis and rounding method give smaller nominal 95% coverage rates. The tables also show that the factor model creates little bias and good 95% coverage rate even under overparameterized scenarios. However, the factor model with correct number of factors performs best among all the models we apply here. The output from other simulation settings have similar results.

4 Application

Diabetes is a lifelong (chronic) disease in which there are high levels of sugar in the blood. Diabetes can be caused by too little insulin, resistance to insulin, or both. The California Health Interview Survey (CHIS) collects information for all age groups on health status, health conditions, health-related behaviors, health insurance coverage, access to health care services, and other health and health related issues. CHIS 2009 is the fifth CHIS data collection cycle and was conducted between September 2009 and April 2010. To investigate the relationship between diabetes and 18 health predictors among Filipinos in Southern California, we use a subset of CHIS 2009 data. Most of the variables we use have missing items due to the non-response of the corresponding survey questions. Among 47614 observations, 430 of them are Filipinos. So the sample size we use is 430. Table 1 gives the brief description of the data set we use. Note that the variables"wrkst" and"aheduc" are not binary but categorical variables. We recode them to be two binary variables "employed" and"bsorabove". Here $employed = 1$ if the observation has a full-time or part-time job currently. Or else $employed = 0$. $bsorabove = 1$ if the observation holds a bachelor or higher degree. Or else $bsorabove = 0$. Meanwhile, some of the variables such as "fruit" and "fry" are not really continuous and normally distributed, we need to do the log transformation to make these variables accommodate the imputation models. Only 283 out of 430 observations are fully observed for all the 19 variables.

This data set example highlights the advantage of our factor analysis modeling strategy. If we assume the unrestricted variance-covariance matrix for the data, the model will include $19 * (19 + 1)/2 = 190$ variance and covariance parameters which is a relatively large number to estimate accurately with 283 complete realizations of y.

To apply factor model to the data, it is very important to find an appropriate number of factors. Checking the eigenvalues of the estimated covariance matrix may not work since some of the variables are not continuous but binary. However, simulations in Section 3 show that overparameterization of factor model still gives small bias and good 95% coverage rate. So here I use a 18-factor model. For simplicity, we use the prior distributions defined in Section 5.4. 31000 iterations are generated and the first 1000 iterations are treated as a burn-in period. The 10 imputations are taken from every 20th iterations since the 30820th iteration due to the auto-correlation plots. After generating multiple complete data sets, Rubin's rule is used to combine the logistic regression estimates. Table 4 shows the combined results.

From Table 4, we find that older Filipinos are more likely to have diabetes than

Table 4: Results of the logistic regression

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

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

younger people. The risk of getting diabetes is higher among heavier Filipinos. Drinking coffee or tea can help Filipinos reduce the risk of getting diabetes. Moreover,smoking has a significant effect on increasing the likelihood of diabetes. All above conclusions are in accord with our common sense. It is interesting that the logistic regression outputs indicate that drinking alcohol will be beneficial to reducing the risk of diabetes. One possible reason is there may exist quadratic effect of alcohol use. Another reason may be we should categorize alcohol use to be moderate use and heavy use. Thus this point is worth further research.

5 Discussion and Future Research

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

From Section 3 we know underparameterization of factor model can result in biased estimates, it seems better to choose enough number of factors to assure inclusion of all important variations. On the other hand, it is generally desirable to have a parsimonious model so that fewer parameters need to be estimated. Since the application of the factor model depend upon the number of factors in use, it would be of interest to develop an adaptive procedure to find an appropriate number of

factors.

Choosing the appropriate number of factors is always a subjective matter, not to mention there are missing items and mixed variable types in the data. Since the number of variables is large and the number of observations is moderate, a large-sample test statistic for choosing the number of factors may not be appropriate. A common way to choose the number of factors is using the scree plot. However, when there are many variables, it is sometimes hard to find a suitable choice from the scree plot. Moreover, it has been known the criterion to choose the number of factors as the number of eigenvalues equals to or greater than one sometimes can lead to the overestimation of the number of factors when there are large number of variables in the model. Song and Belin (2008) developed a new method of choosing the number of factors. First they apply EM algorithm to estimate the parameters in the factor model. Then they use AIC or BIC to choose the appropriate number of factors. But their approach need to be extended to handle the mixed variables scenario. Meanwhile, the computation of AIC and BIC may be burdensome when the data is high-dimensional. A reversible-jump MCMC algorithm was proposed by Lopes and West (2004) to find the correct number of factors. It is possible to modify their algorithm (e.g., adding one step of missing data imputation) to accommodate the mixed incomplete data situation.

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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