International Statistical Review (2003), **71**, 3, 581–592, Printed in The Netherlands

Multiple Imputation: Theory and Method

Paul Zhang

Pfizer Consumer Healthcare, 201 Tabor Road, Morris Plains, NJ 07950, USA.
E-mail: paul.zhang@pfizer.com

Summary

In this review paper, we discuss the theoretical background of multiple imputation, describe how to build an imputation model and how to create proper imputations. We also present the rules for making repeated imputation inferences. Three widely used multiple imputation methods, the propensity score method, the predictive model method and the Markov chain Monte Carlo (MCMC) method, are presented and discussed.

Key words: Missing data; Incomplete data; Missingness mechanism; Multiple imputation.

1 Introduction

Missing data, or incomplete data, are very common in statistical situations. The simple application of complete-data based methods without carefully considering the missingness mechanism may result in an incorrect conclusion. The validity and efficiency of complete-data based methods cannot be guaranteed when data are incomplete (Rubin, 1976).

In the last couple of decades, several methods have been developed to make statistical inferences when data are incomplete, such as the EM algorithm (Dempster, Laird & Rubin, 1977; Little & Rubin, 1987), the random effect model (McCullagh & Nelder, 1989; Diggle *et al.*, 1994), etc. These likelihood-based methods utilize the observed data only, because all inferences are based on the observed-data likelihood function. Under certain assumptions, these methods make valid inferences on the unknown parameters, although the likelihood function itself could be complicated by the missing data and the precisions of these inferences are generally reduced by the missing data. Another approach is to fill the missing data with some 'plausible' values, and then apply complete-data based methods to make valid and efficient inferences. This approach is called imputation of missing data, and it avoids the complexity caused by the missing data. However, the uncertainties associated with the imputations need to be appropriately addressed because imputed values are not real observed values. Three types of uncertainties are involved in the imputation process: One is the uncertainty due to modeling the joint distribution of the response variables and the missingness indicators. The second is the uncertainty due to the sampling from a given imputation model assuming that the observed-data and the values of the model parameters are known. The third is the uncertainty about the values of the model parameters. This is the uncertainty for selecting the imputation model.

In this paper, we present the theoretical background of multiple imputation and describe three widely used multiple imputation methods. In section 2, the definitions of missingness mechanisms are reviewed. In section 3, we discuss the properties of imputation models and explain how to build an imputation model. In section 4, the concepts of multiple imputation, especially proper multiple imputation, are presented, and the standard rules for combining the inference results from multiple imputations to make a repeated imputation inference are explained. Two multiple imputation

methods for monotone missing data, the propensity score method and the predictive model method, are presented in section 5. The use of Markov chain Monte Carlo (MCMC) to create multiple imputations for non-monotone missing data is outlined in section 6. The last section presents some issues concerning the application of multiple imputation, and describes some recent developments.

2 Missingness Mechanism

Let Y be an $n \times p$ data matrix, $Y = (y_1, y_2, \ldots, y_n)^T$, where $y_i = (y_{i1}, \ldots, y_{ip})^T$ is a random sample from a p-dimensional multivariate probability distribution $P(Y|\theta)$ governed by parameters θ. In the following, we refer to the rows of Y as observations, denoted by $y_i (i = 1, 2, \ldots, n)$, and the columns of Y as variables, denoted by $Y_j (j = 1, \ldots, p)$. Moreover, we define an $n \times p$ missingness indicator matrix $R = (r_{ij})$ by letting

$$r_{ij} = \begin{cases} 1 & \text{if } y_{ij} \text{ is missing} \\ 0 & \text{if } y_{ij} \text{ is observed} \end{cases} . \tag{2.1}$$

Defining $\Pr\{r_{ij} = 0|y_{ij}\} = \Pr\{y_{ij} \textit{ observed } |y_{ij}\} = p_{ij}$, then R is subject to a probability distribution $P(R|\xi, Y)$ governed by parameters ξ. Under this assumption, the joint probability distribution of the response variables and the missingness indicator variables can be expressed as

$$P(Y, R|\theta, \xi) = P(Y|\theta)P(R|\xi, Y), \tag{2.2}$$

where $P(Y|\theta)$ is the marginal distribution of the response variables, and $P(R|\xi, Y)$ is the conditional distribution of missingness given the response variables.

If data are incomplete, following Little & Rubin (1987) and Rubin (1987), we use the notation Y_{obs} and Y_{mis} to represent the observed portion and the missing portion of Y, i.e., $Y_{obs} = \{y_{ij}|r_{ij} = 0\}$ and $Y_{mis} = \{y_{ij}|r_{ij} = 1\}$. We also use the notation $Y_{obs,j}$ and $Y_{mis,j}$ to represent the observed portion and the missing portion of variable Y_j, and use $y_{i(obs)}$ and $y_{i(mis)}$ to represent the observed portion and the missing portion of the i-th observation.

Note in the probability model (2.2), there are two sets of parameters, the parameters of interest θ and the nuisance parameters ξ. The correct inferences on θ in general need to be conducted based on the joint probability model (2.2). Hence these inferences depend on how the probability model for the missingness is defined, i.e., how the missingness depends on Y. Rubin (1976), Little & Rubin (1987) categorized the missingness mechanism into the following three categories based on the conditional distribution $P(R|\xi, Y)$.

(1) If $P(R|\xi, (Y_{obs}, Y_{mis})) = P(R|\xi)$, the missingness is independent of the responses (observed and missing), then the missingness mechanism is defined as Missing Completely At Random (MCAR).
(2) If $P(R|\xi, (Y_{obs}, Y_{mis})) = P(R|\xi, Y_{obs})$, the missingness is independent of the missing responses given the observed values. In this case the missingness mechanism is defined as Missing At Random (MAR).
(3) If $P(R|\xi, (Y_{obs}, Y_{mis})) \neq P(R|\xi, Y_{obs})$, the missingness depends on both observed and missing responses.

Of the above definitions, MCAR is the most restrictive because the missing values do not depend on the response variables, neither observed values nor missing values. In this case, the missing values for a variable are like a simple random sample of the data for that variable, so the distribution of missing values is the same as the distribution of observed values. In contrast to MCAR, MAR is a less restrictive assumption because the missing values can depend on the response variables through the observed values. In this case, the missing values for a variable are like a simple random sample of the data for that variable within subgroups defined by the observed values, and the distribution of

missing values is the same as the distribution of observed values within each subgroup.

The two sets of parameters θ and ξ are said to be distinct by Rubin (1976), Little & Rubin (1987) and Rubin (1987) if: (1) From a frequentist perspective, the joint parameter space of (θ, ξ) is the Cartesian cross-product of the parameter spaces for θ and ξ. (2) From a Bayesian perspective, the joint prior distribution of (θ, ξ) can be factored into the independent marginal prior distributions for θ and ξ. This assumption is intuitively reasonable in many real situations since knowing θ provide no information about ξ and vice versa.

If θ and ξ are distinct, and if either MCAR or MAR holds, we will show in the next section that the inferences on θ based on the observed-data likelihood function $L(\theta, \xi|Y_{obs}, R)$ will be the same as based on the observed-data likelihood function $L(\theta|Y_{obs})$. The definitions of these two observed-data likelihood functions will be given below. In that situation, the missingness mechanism is said to be *ignorable* (Rubin, 1976; Little & Rubin, 1987; Rubin, 1987), in the sense that the missingness mechanism can be ignored when making likelihood-based or Bayesian statistical inferences on the parameters of interest θ. A missingness mechanism that does not satisfy this definition is called non-ignorable.

3 Imputation Model

When data are incomplete, the full probability model to describe the data is the joint probability model $P(Y_{obs}, Y_{mis}, R|\theta, \xi)$. Since Y_{mis} is unknown, the likelihood function of this distribution cannot be evaluated. We therefore evaluate the observed-data likelihood function. By the definition, the observed-data likelihood function is proportional to the marginal distribution of the joint distribution integrated over Y_{mis}, i.e.,

$$L(\theta, \xi|Y_{obs}, R) \propto P(Y_{obs}, R|\theta, \xi), \tag{3.1}$$

where

$$P(Y_{obs}, R|\theta, \xi) = \int P(Y_{obs}, Y_{mis}, R|\theta, \xi)dY_{mis} = \int P(R|Y_{obs}, Y_{mis}, \xi)P(Y_{obs}, Y_{mis}|\theta)dY_{mis}. \tag{3.2}$$

Under the above definitions of MCAR and MAR, (3.2) becomes

$$P(Y_{obs}, R|\theta, \xi) = \begin{cases} P(R|\xi)P(Y_{obs}|\theta) & \text{if MCAR} \\ P(R|\xi, Y_{obs})P(Y_{obs}|\theta) & \text{if MAR} \end{cases}. \tag{3.3}$$

If θ and ξ are distinct, the likelihood-based inferences on θ can be conducted based on $P(Y_{obs}|\theta)$ alone, without concerning $P(R|\xi)$ or $P(R|\xi, Y_{obs})$. That is, the joint observed-data distribution $P(Y_{obs}, R|\theta, \xi)$ can be replaced by the marginal observed-data distribution $P(Y_{obs}|\theta)$ for the purpose of inferences on θ, and the observed-data likelihood function ignoring the missingness mechanism is proportional to this distribution,

$$L(\theta|Y_{obs}) \propto P(Y_{obs}|\theta). \tag{3.4}$$

Similarly, the observed-data posterior distribution of θ is given by

$$P(\theta|Y_{obs}) \propto L(\theta|Y_{obs}) \times \pi_\theta(\theta), \tag{3.5}$$

where $\pi_\theta(\theta)$ is a prior distribution of θ.

Frequentists likelihood inferences on θ are based on the observed-data likelihood function (3.4), whereas Bayesians inferences on θ are based on the observed-data posterior distribution (3.5). These inferences utilize the observed data Y_{obs} only and are valid as long as the missingness mechanism is ignorable. However, the precision of the inference is reduced if a large amount of information is missing.

Another approach to handle missing data is to impute the missing data Y_{mis}, and then apply complete-data based methods to the imputed complete data to make inferences on the parameters θ. The conditional probability distribution of Y_{mis} given Y_{obs} can be derived by integrating over the parameter space of θ, i.e.,

$$P(Y_{mis}|Y_{obs}) = \int P(Y_{mis}|Y_{obs}, \theta)P(\theta|Y_{obs})d\theta. \tag{3.6}$$

where $P(Y_{mis}|Y_{obs}, \theta)$ is the *conditional predictive* distribution of Y_{mis} given Y_{obs} and θ, and $P(\theta|Y_{obs})$ is the observed-data posterior distribution of θ. The conditional distribution $P(Y_{mis}|Y_{obs})$ is called the *posterior predictive distribution* of Y_{mis} given Y_{obs} (Rubin, 1987), and it is the conditional predictive distribution $P(Y_{mis}|Y_{obs}, \theta)$ averaged over the observed-data posterior distribution of θ. The random draws from the posterior predictive distribution $P(Y_{mis}|Y_{obs})$ involve three types of uncertainties. One is the uncertainty due to the choice of model for $P(Y, R|\theta, \xi)$. The second is the uncertainty due to the random sampling from $P(Y_{mis}|Y_{obs}, \theta)$ when the observed data Y_{obs} and the values of the parameters θ are known. The third is the uncertainty due to the random sampling of θ from the posterior distribution $P(\theta|Y_{obs})$.

Because of the integration in (3.6), the posterior predictive distribution can rarely be expressed in a closed form. Also, it will be difficult to draw samples from this distribution directly. However, the conditional predictive distribution is often easy to obtain once the observed data and the values of the parameters θ are given. For example, when a multivariate normal distribution is assumed for the response variables, the conditional predictive distribution of Y_{mis} given Y_{obs} and θ is multivariate or univariate normal. Therefore, if the values of the parameters θ can be drawn from their posterior distribution $P(\theta|Y_{obs})$, then the corresponding draws from the conditional predictive distribution $P(Y_{mis}|Y_{obs}, \theta)$ given Y_{obs} and θ are the draws from the posterior predictive distribution $P(Y_{mis}|Y_{obs})$.

One set of imputed values of Y_{mis} can be obtained as follows: Each element of Y_{mis} is filled by a random draw from the conditional predictive distribution $P(Y_{mis}|Y_{obs}, \widetilde{\theta})$, where $\widetilde{\theta}$ is a random draw from the observed-data posterior distribution $P(\theta|Y_{obs})$. In order to generate m sets of conditionally independent imputations given Y_{obs}, m simulated values of θ are independently drawn from the observed-data posterior distribution, say $\widetilde{\theta}^{(t)}$ $(t = 1, \ldots, m)$. For each simulated value $\widetilde{\theta}^{(t)}$, one set of imputed values of Y_{mis} is obtained by taking random draws from each corresponding conditional predictive distribution $P(Y_{mis}|Y_{obs}, \widetilde{\theta}^{(t)})$.

4 Multiple Imputation

Multiple imputation (MI) was first proposed by Rubin in the 1970's as a possible solution to the problem of survey non-response (Rubin, 1977, 1978). He emphasized that missing data should be handled based on some principled methods, rather than ad hoc. Multiple imputation is a principled method which consists of three steps. The first step is to create $m (m > 1)$ complete data sets by filling each missing value m times using m independent draws from an appropriate imputation model given the observed values. The imputation model should be constructed to reflect the true distributional relationship between the missing values and the observed values. In the second step the m imputed complete data sets are analyzed by treating each imputed complete data set as a 'real' complete data set. Standard complete-data procedures and software can be utilized directly. In the third step, the analysis results from m imputed complete data sets are combined in a simple, appropriate way to obtain the so-called repeated imputation inference (Rubin, 1987). The variances of combined estimates consist of the within imputation and the between imputation variances, and therefore the uncertainties in the imputed data are generally correctly incorporated into the final inference. This method overcomes the drawback of single imputation, which underestimates the standard errors of the estimates because it has zero between imputation variance.

The first step, drawing random samples from an imputation model, is the most fundamental part of the MI. This task involves building an imputation model and drawing random samples from it. A multiple imputation is called *proper* if it satisfies certain frequentist properties (Rubin, 1987). If the imputation model does not preserve the distributional relationships between the missing values and the observed values, then the inferences on these relationships from imputed complete data will generally be biased. For example, if the imputation model does not include variables to be used in the inferences from the imputed complete data, then the correlations between these omitted variables and the imputed variables will be biased towards zero. If the multiple imputations are not based on conditionally independent samples from the imputation model given Y_{obs}, the between imputation variance typically will be underestimated.

The theory for making a repeated imputation inference is derived from a Bayesian model (Rubin, 1987), and also justified from a frequentist perspective when multiple imputations are proper. Let Q be a generic scalar quantity to be estimated, such as treatment effect, odds ratio or regression coefficient. Then the observed-data posterior distribution of Q is given by

$$P(Q|Y_{obs}) = \int P(Q|Y_{obs}, Y_{mis})P(Y_{mis}|Y_{obs})dY_{mis}. \tag{4.1}$$

That is, the observed-data posterior distribution of Q is the completed-data posterior distribution of Q averaged over the posterior predictive distribution of Y_{mis}. Let $\hat{Q} = \hat{Q}(Y)$ denote the statistic that would be used to estimate Q if complete data are available and let $\hat{U} = \hat{U}(Y)$ be its squared standard error. Then the moment summaries can be obtained from the observed-data posterior distribution as

$$E(Q|Y_{obs}) = E\,[E(Q|Y_{obs}, Y_{mis})|Y_{obs}] \approx Ave(\hat{Q}),$$

$$V(Q|Y_{obs}) = E\,[V(Q|Y_{obs}, Y_{mis})|Y_{obs}] + V\,[E(Q|Y_{obs}, Y_{mis})|Y_{obs}] \approx Ave(\hat{U}) + (1 + m^{-1})V(\hat{Q}),$$

where $\hat{Q}$ and $\hat{U}$ are calculated from the imputed complete data, and $Ave(\hat{Q})$, $Ave(\hat{U})$ and $V(\hat{Q})$ are the averages and variance over the repeated imputations. The inflation factor $(1 + m^{-1})$ accounts for the additional variance due to the fact that a finite number of imputations is used to calculate $Ave(\hat{Q})$ (Rubin & Schenker, 1986; Rubin, 1987).

After the missing data Y_{mis} have been imputed by m sets of conditionally independent draws $Y_{mis}^{(t)}$ from the posterior predictive distribution $P(Y_{mis}|Y_{obs})$, the repeated imputation inference is obtained as follows. Calculate the repeated estimates $\hat{Q}^{(t)} = \hat{Q}(Y_{obs}, Y_{mis}^{(t)})$ along with their estimated squared standard errors $\hat{U}^{(t)} = \hat{U}(Y_{obs}, Y_{mis}^{(t)})$ from the imputed complete data sets $\{Y_{obs}, Y_{mis}^{(t)}\}(t = 1, 2, \ldots, m)$. Then the overall estimate of Q is simply the average of these repeated estimates,

$$\overline{Q} = \frac{1}{m}\sum_{t=1}^{m} \hat{Q}^{(t)}. \tag{4.2}$$

The standard error of $\overline{Q}$ is

$$T = \left\{(1 + m^{-1})B + \overline{U}\right\}^{1/2}, \tag{4.3}$$

where $B = \frac{1}{m-1}\sum_{t=1}^{m}(\hat{Q}^{(t)} - \overline{Q})^2$ is the between imputation variance and $\overline{U} = \frac{1}{m}\sum_{t=1}^{m}\hat{U}^{(t)}$ is the within imputation variance.

When data are complete, it is assumed that the hypothesis test and the confidence interval are based on the standard normal reference distribution

$$(\hat{Q} - Q)/\sqrt{\hat{U}} \sim N(0, 1). \tag{4.4}$$

One may need to make a transformation to have the estimand better meet this assumption. When data are incomplete, if the size of complete data is large and the number of imputations, m, is small,

the hypothesis test and the confidence interval are based on a Student-t reference distribution

$$(\overline{Q} - Q)/T \sim t_v, \tag{4.5}$$

with the degrees of freedom being given by

$$v = (m-1)\left[1 + \frac{\overline{U}}{(1+m^{-1})B}\right]^2. \tag{4.6}$$

Note that formula (4.6) is based on the assumption that the complete data inference is based on the normal reference distribution, i.e., the complete data degrees of freedom, v_{com}, is infinite and the number of imputations is finite. When v_{com} is small and there is only a modest proportion of missing data, the degrees of freedom defined by formula (4.6) can be much larger than v_{com}. In this case, formula (4.6) is inappropriate. Barnard & Rubin (1999) proposed a modified formula for calculating the degrees of freedom. The modified formula has the following properties. For fixed m and an estimated fraction of missing information: (1) the modified degrees of freedom monotonically increases in v_{com}, (2) the modified degrees of freedom is always less than or equal to v_{com}, and (3) the modified degrees of freedom equals to the original degrees of freedom when v_{com} is infinite. To describe the modified degrees of freedom, we denote the original degrees of freedom by $v_m = (m-1)(1 + r_m^{-1})^2$ where $r_m = (1 + m^{-1})B/\overline{U}$. The modified degrees of freedom $\tilde{v}_m$ can then be expressed by

$$\tilde{v}_m = \left(\frac{1}{v_m} + \frac{1}{v_{obs}}\right)^{-1}, \tag{4.7}$$

where

$$v_{obs} = \left(\frac{v_{com}+1}{v_{com}+3}\right) v_{com}(1 - r_m) \tag{4.8}$$

is the observed data degrees of freedom. This formula for the modified degrees of freedom will generate a larger percentile t-value, which will result in a wider confidence interval and a larger p-value than the original formula.

5 Monotone Pattern of Missing Data

A data matrix is said to have a monotone missing pattern if, whenever an element y_{ij} is missing, the elements y_{ik} are also missing for all $k > j$ (Anderson, 1957; Rubin, 1974; Little & Rubin, 1987). Monotone missing patterns are often seen in clinical trials with repeated measurements. A subject may prematurely discontinue the trial at any time during the trial, so that all the subsequent measurements after this time from this subject will be missing. Let n_j denote the number of observed values of variable Y_j, then when missing data follow a monotone pattern, we have $n = n_1 \geq n_2 \geq \cdots \geq n_{p-1} \geq n_p$.

5.1 *Propensity Score Method*

Lavori *et al.* (1995) proposed this propensity score method of multiple imputations. Basically, this method follows a nonparametric approach in which the missing values are imputed by resampling of the observed values.

By defining the missingness indicator variables r_{ij} as in (2.1), the missingness can be predicted by a linear logistic regression model. The conditional probability of observing y_{ij}, given the previous history $y_{i1}, \ldots, y_{i,j-1}$, can be called a propensity score s_{ij} (in analogy with Rosenbaum & Rubin's

definition of the propensity score, 1983), that is,

$$s_{ij} = \Pr\{r_{ij} = 0 | y_{i1}, \ldots, y_{i,j-1}\}. \tag{5.1.1}$$

Since the missing data have a monotone pattern, the propensity score can be modeled by

$$\log(\frac{s_{ij}}{1 - s_{ij}}) = \beta_0 + \beta_1 y_{i1} + \cdots + \beta_{j-1} y_{i,j-1}, \tag{5.1.2}$$

where $\beta_0, \beta_1, \ldots, \beta_{j-1}$ are the regression coefficients. After the regression coefficients in (5.1.2) are estimated from the observed r_{ij} for the response variable Y_j and the complete data for the covariates $Y_1, \ldots, Y_{j-1}$, each observation can be assigned an estimated propensity score,

$$\hat{s}_{ij} = \frac{\exp\{\hat{\beta}_0 + \hat{\beta}_1 y_{i1} + \cdots + \hat{\beta}_{j-1} y_{i,j-1}\}}{1 + \exp\{\hat{\beta}_0 + \hat{\beta}_1 y_{i1} + \cdots + \hat{\beta}_{j-1} y_{i,j-1}\}} \tag{5.1.3}$$

and then all observations are stratified into q strata based on the quantiles of estimated propensity scores. Within each stratum, a 'donor pool' is created by applying the approximate Bayesian bootstrap (ABB), (Rubin & Schenker, 1986). That is, a random sample is created by random draws with replacement from the observed values of Y_j within each stratum. The ABB method is applied in order to reflect additional uncertainty about the posterior distribution of the underlying parameters, given the observed values of Y_j within each stratum. This method is roughly equivalent to choosing the values of the parameters θ for the conditional posterior predictive distribution $P(Y_{mis}|Y_{obs}, \theta)$ from the observed-data posterior distribution $P(\theta|Y_{obs})$.

After the donor pools are created, each missing value of Y_j is then imputed by a single random draw from its donor pool. m sets of multiple imputations are obtained by creating m conditionally independent donor pools for each individual missing value and then taking a random draw from each donor pool. Note that imputing a missing value by a random draw from its stratum rather than from its donor pool would result in an improper multiple imputation in the sense that the between imputation variance would be underestimated because the uncertainty due to selecting the imputation model is not incorporated into the imputation.

It is important to note that it is the *missingness* being modeled rather than the *missing values* being modeled. The logistic model (5.1.2) explains the relationships between the missingness indicator r_{ij} and the previous history $(Y_1, \ldots, Y_{j-1})$, but it does not model the relationships between $Y_{mis,j}$ and $(Y_1, \ldots, Y_{j-1})$. This is pointed out by Schafer (`www.stat.psu.edu/~jls`) on his web page regarding the use of Solas (Statistical Solutions Ltd.). He commented that this method is effective for the analysis pertaining to the marginal distribution of the missing values of Y_j, but it is not appropriate in general for the analysis involving the relationships between Y_j and $(Y_1, \ldots, Y_{j-1})$. One hypothetical example in his comment is that if some variables are highly correlated with $Y_{mis,j}$ but unrelated to the missingness indicator r_{ij}, then those variables have no influence in the logistic regression model. Therefore, the imputed values of Y_j bear no relationship to those variables, and the estimate of the correlation between Y_j and those variables from the multiple imputed data sets will be biased towards zero. In this case, the propensity score method is unable to preserve important features of the joint distribution of Y_j and $(Y_1, \ldots, Y_{j-1})$.

5.2 Predictive Model Method

When missing data have a monotone pattern, the joint observed-data likelihood function can be factored into the independent observed-data likelihood functions, that is,

$$L(\phi_1, \ldots, \phi_p|Y_{obs}) = \prod_{j=1}^{p} L(\phi_j|Y_{obs}),$$

where

$$L(\phi_j|Y_{obs}) \propto \prod_{i=1}^{n_j} P(y_{ij}|y_{i1}, \ldots, y_{i,j-1}, \phi_j),$$

and $P(Y_j|Y_1, \ldots, Y_{j-1}, \phi_j)$ is the conditional distribution of Y_j given $Y_1, Y_2, \ldots, Y_{j-1}$ and ϕ_j is the conditional distribution parameters. Thus missing data with a monotone pattern can be imputed from independent univariate distributions given the previous observed data (Little & Rubin, 1987).

If a multivariate normal is assumed for the response variables $Y_1, Y_2, \ldots, Y_p$, the observed-data likelihood function $L(\phi_j|Y_{obs})$ becomes a linear regression of Y_j on $Y_1, Y_2, \ldots, Y_{j-1}$ based on first n_j observations, and the conditional distribution parameters ϕ_j become the regression coefficients and the residual variance. The missing values of Y_j can be imputed by the predicted values from this linear regression model, given observed values of $Y_1, \ldots, Y_{j-1}$ and simulated regression parameters which are randomly drawn from their observed-data posterior distribution. In this way, the additional uncertainty due to the fact that the regression coefficients can be estimated, but not determined, from the observed values of Y_j and $Y_1, \ldots, Y_{j-1}$ is reflected. Using fixed estimates of regression parameters (rather than simulated values from their observed-data posterior distribution) will result in improper multiple imputations in the sense that the between imputation variance would be underestimated (Rubin, 1987).

The computational details of the predictive model method of multiple imputations are as follows: Let X be the data matrix for $Y_1, \ldots, Y_{j-1}$ with augmented ones in the first column to incorporate the intercept, and let X_{obs} and X_{mis} be the rows of X corresponding to $Y_{obs,j}$ and $Y_{mis,j}$ respectively. Note that since the missing data have a monotone pattern, the first n_j observations of Y_j are observed and the remaining of $(n_j + 1)$ to n observations are missing. Since the probability model of Y_j given $Y_1, \ldots, Y_{j-1}$ is an univariate normal distribution $Y_j \sim N_1(\mu_j, \sigma_j^2)$, where $\mu_j = \beta_0 + \beta_1 Y_1 + \cdots + \beta_{j-1} Y_{j-1}$, the observed-data likelihood function of the regression parameters $\phi_j = (\beta_0, \beta_1, \ldots, \beta_{j-1}, \sigma_j^2)$ is

$$L(\mu_j, \sigma_j^2|Y_{obs}) \propto \sigma_j^{-n_j} \exp\left\{-\frac{1}{2\sigma_j^2}\sum_{i=1}^{n_j}(y_{ij} - X_{obs(i)}\beta)^2\right\},$$

where $\beta = (\beta_0, \beta_1, \ldots, \beta_{j-1})$ is a vector of regression coefficients. Suppose a non-informative prior distribution $\pi(\phi_j) \propto \sigma_j^{-1}$ is assumed for (β, σ_j^2). After some manipulations, the observed-data posterior distribution of (β, σ_j^2) can be expressed as

$$\sigma_j^{-\frac{p}{2}} \exp\left\{-\frac{1}{2\sigma_j^2}(\beta - \hat{\beta})^T (X_{obs}^T X_{obs})(\beta - \hat{\beta})\right\} \times \sigma_j^{-(\frac{n_j - p}{2})-1} \exp\left\{-\frac{1}{2\sigma_j^2}\sum_{i=1}^{n_j}(y_{ij} - X_{obs(i)}\hat{\beta})^2\right\}$$

which is the product of a multivariate normal and a scaled inverted-chisquare distribution,

$$\beta|Y_{obs}, \sigma_j^2 \sim N_j(\hat{\beta}, \sigma_j^2 (X_{obs}^T X_{obs})^T), \tag{5.2.1}$$

$$\sigma_j^2|Y_{obs} \sim \hat{\varepsilon}^T \hat{\varepsilon} \chi_{n_j - p}^{-2}, \tag{5.2.2}$$

where $\hat{\beta} = (X_{obs}^T X_{obs})^{-1} X_{obs}^T Y_{obs,j}$ is the MLE of β from the observed-data likelihood function, and $\hat{\varepsilon} = Y_{obs,j} - X_{obs}\hat{\beta}$ is the residual vector.

The values of (β, σ_j^2) can be simulated from their observed-data posterior distribution (5.2.1)–(5.2.2) by letting $\tilde{\sigma}_j^2 = \hat{\varepsilon}^T \hat{\varepsilon}/\sigma^*$, where σ^* is a random draw from $\chi_{n_j - p}^2$, and then taking a random

draw $\widetilde{\beta}$ from (5.2.1) given $\widetilde{\sigma}_j^2$. After a random draw $(\widetilde{\beta}, \widetilde{\sigma}_j^2)$ has been taken from their observed-data posterior distribution, the missing values of Y_j are imputed by independent random draws from the conditional predictive distribution $P(Y_{mis.j}|Y_{obs}, \widetilde{\beta}, \widetilde{\sigma}_j^2)$, which is a univariate normal distribution $N_1(X_{mis}\widetilde{\beta}, \widetilde{\sigma}_j^2)$. In order to obtain m sets of multiple imputations, m conditionally independent random draws are taken from the observed-data posterior distribution (5.2.1)–(5.2.2), say $(\widetilde{\beta}^{(t)}, \widetilde{\sigma}_j^{2(t)})$, $t = 1, \ldots, m$. For each simulated values of regression parameters $(\widetilde{\beta}^{(t)}, \widetilde{\sigma}_j^{2(t)})$, the missing values of Y_j are imputed by independent random draws from $N_1(X_{mis}\widetilde{\beta}^{(t)}, \widetilde{\sigma}_j^{2(t)})$.

6 Markov Chain Monte Carlo (MCMC) Method

When missing data have a non-monotone missing pattern, the joint observed-data likelihood function cannot be factored into the independent observed-data likelihood functions because of the missing values of $Y_1, Y_2, \ldots, Y_{j-1}$. Thus missing data with a non-monotone pattern cannot be imputed from independent univariate distributions as for missing data with a monotone pattern. To impute the missing data with a non-monotone pattern, more complicated computational procedures need to be applied. In this case, Markov chain Monte Carlo (MCMC) is very useful. The MCMC is a Monte Carlo integration method using Markov chains. This method has been successfully applied in a broad range of statistical situations (Gilks *et al.*, 1996). Schafer (1997) applied the MCMC method for the purpose of multiple imputations by utilizing the data augmentation algorithm developed by Tanner & Wong (1987).

The MCMC method is to draw the pseudo random samples from a target probability distribution. When missing data have a non-monotone pattern, the target distribution is the joint conditional distribution of Y_{mis} and θ given Y_{obs},

$$P(Y_{mis}, \theta|Y_{obs}). \tag{6.1}$$

The MCMC method for imputing missing data is conducted as follows: Replace the missing data Y_{mis} by some assumed values, then θ can be simulated from the resulting complete data posterior distribution $P(\theta|Y_{obs}, Y_{mis})$. Let $\theta^{(t)}$ be the current simulated value of θ from the complete data posterior distribution, then the next iterative sample of Y_{mis}, $Y_{mis}^{(t+1)}$, can be drawn from the conditional predictive distribution of Y_{mis} given Y_{obs} and $\theta^{(t)}$, i.e.,

$$Y_{mis}^{(t+1)} \sim P(Y_{mis}|Y_{obs}, \theta^{(t)}). \tag{6.2}$$

Conditioning on $Y_{mis}^{(t+1)}$, the next simulated value of θ can be drawn from its complete data posterior distribution,

$$\theta^{(t+1)} \sim P(\theta|Y_{obs}, Y_{mis}^{(t+1)}). \tag{6.3}$$

Repeating the random draws of (6.2)–(6.3) from a starting value of $\theta^{(0)}$ yields a Markov chain $\{(\theta^{(t)}, Y_{mis}^{(t)}) : t = 1, 2, \ldots\}$. The stationary distribution of this chain is the joint distribution of θ and Y_{mis} given Y_{obs}, $P(Y_{mis}, \theta|Y_{obs})$. Consequently, the marginal stationary distributions of the subsequence $\{\theta^{(t)} : t = 1, 2, \ldots\}$ and $\{Y_{mis}^{(t)} : t = 1, 2, \ldots\}$ are the observed-data posterior distribution $P(\theta|Y_{obs})$ and the posterior predictive distribution $P(Y_{mis}|Y_{obs})$ respectively. When t is sufficiently large, $\theta^{(t)}$ can be viewed as a single simulation from the observed-data posterior distribution $P(\theta|Y_{obs})$, and $Y_{mis}^{(t)}$ can be viewed as a single imputation from the posterior predictive distribution $P(Y_{mis}|Y_{obs})$. The random draw of (6.2) is to impute the missing data Y_{mis}, and the random draw of (6.3) is to simulate the unknown parameter θ. Therefore (6.2) and (6.3) are referred to as the Imputation or I-step and the Posterior or P-step respectively (Tanner & Wong, 1987), in analogy with the E-step and M-step of the EM algorithm. The first use of this algorithm seems to have been made by Li (1988) who presented an argument for convergence and used it to impute the missing data Y_{mis}.

Multiple imputations of Y_{mis} ideally should be independent given Y_{obs} . However, even after a long period of burn-in, such multiple imputations cannot be acquired by successive iterations from a single chain because the successive iterations tend to be correlated. One way to get proper multiple imputations is to take a widely separated sub-sample from a single chain. For example, take every k-th iteration after the burn-in period, where k is large enough so that the dependence between the imputed values is negligible. Alternatively, one can generate m independent chains, after the burn-in period, take the final value of each chain as the imputed values of Y_{mis}.

The MCMC method avoids complicated analytic calculation of the observed-data posterior distribution of unknown parameters θ and the posterior predictive distribution of missing data Y_{mis} given Y_{obs}, which appears necessary when the missing pattern is non-monotone. But as for any iterative methodology, convergence is an issue one needs to face. In the MCMC, the convergence of the Markov chain $\{(\theta^{(t)}, Y_{mis}^{(t)})\}$ is a *distribution* convergence. When the Markov chain converges, the distribution of (θ, Y_{mis}) no longer changes from one iteration to the next, but the random draws from this distribution do change from iteration to iteration. Therefore, monitoring convergence of Markov chain is more complicated than that of pointwise convergence, and consequently the methods designed for monitoring this type of convergence should be applied (Gelman & Rubin, 1992).

7 Remarks

Multiple imputation is a useful tool for handling missing data because of its flexibility to enable one to conduct the imputation and subsequent analysis separately. The validity of the analysis can depend on the imputation model's capability to correctly capture the missingness mechanism. Also, the separation of imputation model and analysis model raises the issue of compatibility between these two models. The impact of incompatibility on the repeated imputation inferences has been investigated by Fay (1992), Meng (1995) and Rubin (1996), and it is discussed in greater depth by other articles in this journal. When the imputation model is more general (has fewer assumptions) than the analysis model, the MI leads to valid inferences with perhaps some loss of efficiency because the extra generality may add more variation among the m sets of imputed missing data $Y_{mis}^{(1)}, \ldots, Y_{mis}^{(m)}$. When the imputation model is more restrictive (has more assumptions) than the analysis model, the situation is twofold. If those extra assumptions are plausible, then the MI leads to valid inferences with perhaps more efficiency than the estimates from the observed data alone. However, if the extra assumptions are unwarranted, the MI may lead to biased estimates. For example, if the missing values of a variable are imputed from a regression model with no interaction, and the analysis model is to investigate the potential interaction, then the MI estimate of interaction will be biased towards zero (Schafer, 1999). Therefore, the imputation model should correctly incorporate the missingness mechanism and reasonably preserve the distributional relationship between Y_{mis} and Y_{obs}.

When missing data are present, three issues are of main concern: (1) loss of efficiency, (2) complications in data handling and statistical inferences and (3) potential serious bias due to the systematic differences between the observed data and the missing data (Barnard & Meng, 1999). If an appropriate imputation model is employed, the MI can improve the efficiency. By creating imputed complete data, (2) is no longer an issue. If ignorable missingness mechanism is tenable, the bias caused by missing data may not be a serious problem. However, it is impossible to test the MAR assumption against a non-ignorable alternative without additional information beyond the data (Little & Rubin, 1987). Thus, Barnard & Meng (1999) suggested that various model checking, in particular, the posterior predictive checks (Rubin, 1984; Gelman *et al.*, 1996), should be performed in all situations where missing data are present. In applications, there may be numerous causes for missing data and the missingness mechanism could be a mixture of ignorable and non-ignorable. As long as a large portion of missing data is ignorable and the non-ignorable portion of missing data are also included into the imputation process, the bias caused by treating all missing data as ignorable

would be negligible.

Rubin & Schenker (1986), Rubin (1987), and Schafer (1997) have shown that a small number of multiple imputations can provide estimates of standard errors that are almost fully efficient. Schafer (1999) suggested that no more than 10 imputations are usually required. However, in the situations where the proportion of missing data is large, the inferences based on the observed-data likelihood may be unstable and the inferences based on the multiple imputations may have quite large variation. In this case, Horton & Lipsitz (2001) suggested that a closer investigation should be conducted. An increase in the number of imputations will stabilize the estimate of variation.

Multiple imputation, as a useful and convenient tool for handling missing data, is becoming more and more popular because of the availability of easy-to-use statistical software. Such software facilitates users to apply the MI in a broader range of missing data settings. However, correct application of multiple imputation requires the user to be careful in checking the assumptions underlying each application. As a summary, we present a brief list of software which implement the three multiple imputation methods discussed in this paper:

Solas v3.0:	The propensity score method and the predictive model method.
BMDP Professional 2.0:	The predictive model method.
SAS v8.2:	The propensity score method and the predictive model method for monotone missing pattern. The MCMC method for non-monotone missing pattern.
NORM:	The MCMC method. (free at http://www.stat.psu.edu/~jls)

For a detailed and most recent review of the multiple imputation software packages, please see Horton & Lipsitz (2001).

Acknowledgement

The author thanks the editor, Dr. Elja Arjas, for his diligent work and thoughtful coordination during the revision of the manuscript. The author also thanks Professor Rubin for his thorough review, valuable comments and insightful suggestions to improve this manuscript.

References

Anderson, T.W. (1957). Maximum likelihood estimates for the multivariate normal distribution when some observations are missing. *Journal of the American Statistical Association*, **52**, 200–203.

Barnard, J. & Meng, X.L. (1999). Applications of multiple imputation in medical studies: from AIDS to NHANES. *Statistical Methods in Medical Research*, **8**, 17–36.

Barnard, J. & Rubin, D.B. (1999). Small-sample degrees of freedom with multiple imputation. *Biometrika*, **86**, 948–955.

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

Diggle, P.J., Liang, K.Y. & Zeger, S.L. (1994). *Analysis of Longitudinal Data*. Oxford: Clarendon Press.

Fay, R.E. (1992). When are inferences from multiple imputation valid? *Proceedings of the Survey Research Methods Section of the American Statistical Association*, pp. 227–232.

Gelman, A. & Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science*, **7**, 457–472.

Gelman, A., Meng, X.L. & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statistica Sinica*, **6**, 733–807.

Gilks, W.R., Richardson, S. & Spiegelhalter, D.J. (1996). *Markov chain Monte Carlo in practice*. London: Chapman & Hall.

Horton, N.J. & Lipsitz, S.R. (2001). Multiple imputation in practice: Comparison of software package for regression models with missing variables. *The American Statistician*, **55**, 244–254.

Lavori, P.W., Dawson, R. & Shera, D. (1995). A multiple imputation strategy for clinical trials with truncation of patient data. *Statistics in Medicine*, **14**, 1913–1925.

Li, K.H. (1988). Imputation using Markov chains. *Journal of Statistical Computation and Simulation*, **30**, 57–79.

Little, R.J.A. & Rubin, D.B. (1987). *Statistical analysis with missing data*. New York: John Wiley.

McCullagh, P. & Nelder, J.A. (1989). *Generalized linear models*. New York: Chapman and Hall.

Meng, X.L. (1995). Multiple-imputation inferences with uncongenial sources of input (with discussion). *Statistical Science*, **10**, 538–573.

Rosenbaum, P.R. & Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, **70**, 41–55.
Rubin, D.B. (1974). Characterizing the estimation of parameters in incomplete data problems. *Journal of the American Statistical Association*, **69**, 467–474.
Rubin, D.B. (1976). Inference and missing data. *Biometrika*, **63**, 581–592.
Rubin, D.B. (1977). Formalizing subjective notions about the effect of nonrespondents in sample surveys. *Journal of the American Statistical Association*, **72**, 538–543.
Rubin, D.B. (1978). Multiple imputations in sample surveys. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, pp. 20–34.
Rubin, D.B. (1984). Bayesianly justifiable and relevant frequency calculations for applied statisticians. *The Annals of Statistics*, **12**, 1151–1172.
Rubin, D.B. (1987). *Multiple imputation for non-response in surveys*. New York: John Wiley.
Rubin, D.B. (1996). Multiple imputation after 18+ years. *Journal of American Statistical Association*, **91**, 473–489.
Rubin, D.B. & Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, **81**, 366–374.
SAS/STAT Software: Changes and Enhancements, Release 8.2 (2001). Carry, North Carolina: SAS Institute.
Schafer, J.L. (1997). *Analysis of incomplete multivariate data*. London: Chapman & Hall.
Schafer, J.L. (1999). Multiple imputation: A primer. *Statistical Methods in Medical Research*, **8**, 3–15.
Solas 3.0. User Reference (1999). Statistical Solutions Ltd., Ireland.
Tanner, M.A. & Wong, W.H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *Journal of American Statistical Association*, **82**, 528–550.

Résumé

En cet exposé synoptique, nous discutons le fond théorique de l'imputation multiple, décrivons comment établir un modèle d'imputation et comment créer imputations appropriée. Nous présentons règles pour faire des inférences répétées d'imputation. Trois méthodes multiples largement répandues d'imputation, imputation multiple de points de propension, imputation multiple modèle prédictive et imputation multiple de Monte Carlo de chaîne de Markov (MCMC), sont présentées et discutées.

[*Received November 2002, accepted April 2003*]