A Semi-Parametric Multiple Imputation Data Augmentation Procedure

Gabriele B. Durrant
Southampton Statistical Sciences Research Institute, University of Southampton, UK

ABSTRACT

Multiple imputation procedures (MI) are a useful tool to adjust for item non-response but are often based on fully parametric assumptions, such as multivariate normality. For many applications such assumptions may not hold in practice, for example if the data are skewed and affected by rounding and truncation effects. Hot deck imputation methods, however, make less or no assumptions about underlying distributions and may be more appropriate to use in such circumstances. The basic idea is to combine multiple imputation and hot deck approaches with the aim of preserving advantageous properties of MI and at the same time relaxing distributional assumptions. This paper develops a semi-parametric data augmentation method to generate MI under the missing at random assumption and a nonignorable missing data mechanism. The use of predictive mean matching as a form of hot deck imputation is considered as part of the data augmentation procedure to improve the robustness of the multiple imputation method.

Keywords: missing data, hot deck imputation; predictive mean matching imputation; nonignorable non-response.

1. Introduction

Item non-response is often a problem in survey data and various imputation and weighting techniques have been developed to compensate for non-response bias. One imputation method is multiple imputation (MI) which has found considerable attention in the literature in recent years. The availability of multiple imputation procedures in a number of software packages has facilitated an easier use of the method. Although other imputation methods may be more efficient than MI and might be more suitable for some applications or certain types of estimators (Kim and Fuller, 2004), MI can be regarded as a flexible approach to imputation with the advantage of providing a relatively simple variance estimation technique. The basic principles of MI have been described in detail in Rubin (1987), Schafer (1997; 1999) and Zhang (2003). Multiple imputation procedures are often based on fully parametric assumptions, such as multivariate normality. An imputation model should, however, preserve distributional features of variables subject to missing data. For example, imputing variables, that are skewed or truncated, under an assumption of joint normality, will lead to biased results if distributional features are of interest, such as the estimation of the tail behavior (see also Schafer, 1999). This may be of concern for example when considering income and earnings variables, in particular when estimating the distribution of such variables.

An attractive approach to imputation is hot deck imputation. An advantage is that actually observed values are imputed and, depending on the type of hot deck method, less or even no assumptions about underlying distributions of the variables of interest are made. Hot deck imputation is common in practice, in particular in official statistics. The combination of both approaches, multiple and hot deck imputation, could lead to imputation methods with desirable properties. Such a combination of methods may have certain advantages, for example overcoming distributional assumptions by using a hot deck method and at the same time providing a simple variance estimation formula by using MI. Some approaches of how to do this have been discussed in the literature such as the use of the approximate Bayesian bootstrap (Rubin and Schenker, 1986) or the semi-parametric methods described in Schenker and Taylor (1996).

In this paper, the use of hot deck methods is proposed as a means of relaxing distributional assumptions made by parametric model-based multiple imputation methods. The paper develops a method that combines hot deck and multiple imputation as a semi-parametric approach under the missing at random (MAR) assumption and a nonignorable non-response assumption.

The structure of the paper is as follows. Section 2 introduces notation, the missing data mechanisms and the estimation problem considered. Section 3 reviews some existing approaches that combine MI and hot deck methods. In section 4 the basic ideas of data augmentation for imputation are introduced and section 5 develops the use of predictive mean matching imputation in the data augmentation procedure under the MAR assumption. The method is extended to a nonignorable missing data setting in section 6. Inference under semi-parametric data augmentation is described in section 7 and section 8 makes some concluding remarks.

2. Missing Data Assumptions and Estimation

To facilitate the discussion the following notation is introduced. The focus is, at least initially, on the univariate non-response case, where only one variable is subject to missing data. Let y_i be the variable of interest, which is missing for some units i in a random sample s. Let r_i be the binary variable indicating whether y_i is observed, i.e. $r_i = 1$ if y_i is observed and $r_i = 0$ if not. The fully observed auxiliary variables in sample s are denoted in a row-vector w_i.

When considering a non-response problem an assumption needs to be made about the non-response mechanism. The standard assumption is that the data are missing at random (MAR), i.e.

$$y_i \perp r_i \mid w_i, \qquad (1)$$

which means that conditioning on the observed variables w_i, y_i is independent of r_i (Rubin, 1987).

Alternatively, if data from a validation sample or a follow up study are available, a nonignorable non-response assumption may be considered. Let us assume that surrogate data are available, that is let x_i be a variable, which measures y_i with error, but is observed for all units in the sample. Then, the following missing data assumption may be made,

$$x_i \perp r_i \mid y_i, w_i, \qquad (2)$$

which means that conditioning on the true variable y_i and w_i, the measurement error variable x_i is independent of r_i. This may be regarded as an attractive assumption in the presence of surrogate data, since it allows conditioning on the true variable y_i. The assumption will be referred to as the common measurement assumption (CME), since it can be interpreted as a measurement error model assuming the same measurement error for respondents and non-respondents. Other nonignorable assumptions may be conceivable depending on the type of data available. (for a discussion on the plausibility of assumptions (1) and (2) for a specific application see Durrant and Skinner, 2006).

In the following, the case is considered where the variable y_i does not follow a normal distribution, but may be skewed and truncated. The aim of the analysis is estimation of the cumulative distribution function of y_i. The parameter of interest, θ, may be expressed as

$$\theta(y) = \frac{1}{N}\sum_{i\in U} I(y_i \leq y), \qquad (3)$$

where U is the population of interest, N is the size of U, $I(.)$ is the indicator function indicating if a condition is true or false and y is a specified threshold. Under imputation this parameter may be estimated by:

$$\hat{\theta}.(y) = \frac{1}{n}\sum_{i\in s} I(y._i \leq y), \qquad (4)$$

where $y._i = y_i$ if $r_i = 1$ and $y._i = \hat{y}_i$ if $r_i = 0$ and $\hat{y}_i$ is the imputed value. For simplicity, it is assumed that $(y_i, x_i, w_i, r_i) \sim iid$. The aim is to define an appropriate imputation method to obtain $\hat{y}_i$ for $r_i = 0$, either under MAR or under nonignorable non-response.

3. Multiple Imputation and Hot Deck

One way of generating MI is to use a Markov chain Monte Carlo method defined in a Bayesian framework (Rubin, 1996; Schafer, 1997; Lipsitz, Zhao and Molenberghs, 1998). Particularly suitable for imputation of missing data is the data augmentation algorithm by Tanner and Wong (1987). Such an approach, however, is fully parametric and requires making assumptions about underlying distributions. As emphasized in Schafer (1997) for some applications a parametric approach might perform reasonably well even if the assumptions do not hold in practice. However, for applications where components of the data are skewed or show certain features, such as truncation and rounding effects, or where the estimation of distributional quantities is of interest, fully parametric approaches do not seem suitable. In such circumstances, it is important to focus on semi-parametric or non-parametric imputation methods that make less or even no distributional assumptions about the variables to be imputed.

Hot deck imputation methods are non-parametric (or semi-parametric) and aim to avoid distributional assumptions. A particularly attractive form of hot deck imputation is predictive mean matching imputation (Little, 1988). Under the MAR assumption, an imputation model is defined relating y_i to auxiliary variables w_i, which may be written as

$$y_i = \eta_i\beta + \varepsilon_i, \ \varepsilon_i \sim N(0, \sigma^2), \qquad (5)$$

where η_i is a vector of covariates, functions of w_i, ε_i are the residual terms and β and σ^2 are the unknown parameters.

The predicted values of y_i, $\hat{y}_i^{pred}$, are obtained from the model estimating the parameters based on respondents data only, and are calculated for all $i \in s$. The value y_{i^*} is replaced for the missing item, where the respondent i^* is the donor for non-respondent j if

$$D_{ji^*} = \min_i \mid \hat{y}_j^{pred} - \hat{y}_i^{pred} \mid,$$

i.e. $\hat{y}_j = y_{i^*}$, where $r_j = 0$ and $r_{i^*} = 1$.
The advantage of the predictive mean matching imputation method is that it uses the linear regression model based on normal theory only to define the distance. The method is therefore expected to be less sensitive to model misspecifications than fully parametric approaches.

One way of implementing a combination of hot deck imputation and multiple imputation under MAR is to use the approximate Bayesian bootstrap (ABB) (Rubin and Schenker, 1986), which may be regarded as a non-parametric approach to MI. Having defined imputation classes, for example based on categories of w_i, the donors within each imputation class are sampled (bootstrapped) with replacement of the same size as respondents are available in each class. For each non-respondent in a class one donor is selected with replacement from the set of bootstrapped respondents for that class at random. This is repeated M times.
The ABB for predictive mean matching imputation is described in Heitjan and Little (1991), and requires bootstrapping the sample s with replacement creating M bootstrap samples $s^{(m)}$, $m = 1,...,M$. The parameters of the imputation model (5) are estimated based on respondents only for each bootstrap sample separately, to reflect parameter uncertainty, and the predicted values, $\hat{y}_i^{pred(m)}$, for all $i \in s$, are defined for each bootstrap sample. Based on these values predictive mean matching imputation is performed, by drawing at random one donor value from a set of nearest neighbors, e.g. defined as the nearest 5 above and 5 below the predicted value of $\hat{y}_j^{pred(m)}$, $r_j = 0$.

Alternatively, at least for the univariate missing data case under MAR, the partially parametric techniques proposed in Schenker and Taylor (1996) may be used. The method requires drawing parameters of the imputation model from their posterior distribution given the observed data, calculated analytically. Then, the M multiple draws of the estimated parameters are used in the imputation model to derive the predicted values of y_i, $\hat{y}_i^{pred(m)}$, $m = 1,...,M$, for all $i \in s$. Using predictive mean matching, a nearest neighbor is defined based on these predicted values and a donor value is chosen for imputation for $m = 1,...,M$.

An alternative, semi-parametric MI approach would be to incorporate a hot deck method in the MI data augmentation procedure. The novelty here is the use of predictive mean matching in the imputation step instead of regression imputation with the aim of relaxing residual assumptions, commonly made in standard data augmentation procedures. The aim is to improve the robustness of the MI procedure to model misspecification. This approach will be discussed in the following sections, first based on the MAR assumption in (1), then under the nonignorable assumption in (2).

4. Data Augmentation

Data augmentation is a Markov chain Monte Carlo method, which enables imputation for complex missing data problems by iteratively solving more tractable complete data problems (Schafer, 1997 and Gelman et al., 1998). In the context of missing data, the data augmentation algorithm consists of a series of imputation steps (I-steps), which impute the missing values given all the observed data and a current set of parameters, and posterior steps (P-steps), in which the parameters of the model are drawn from their posterior distribution given the complete data formed from the I-step. On convergence, the algorithm should provide imputed values from the conditional distribution of the missing values given the observed data, where the distribution is integrated over any unknown parameters in the model with respect to the posterior distribution of these parameters given the data.

The following notation is used. The vectors of length n, containing the sample values are denoted Y, X and R, for example $Y = (y_1,...,y_n)'$. Similarly, W denotes a matrix with values of the covariates. Suppose without loss of generality that for the direct variable only the first n_r elements are observed in sample s and the following $n - n_r$ elements are missing. It is $Y = (Y'_{obs}, Y'_{mis})'$, where $Y_{obs} = (y_1,...,y_{n_r})'$ is the observed part of Y and $Y_{mis} = (y_{n_r+1},...,y_n)'$ is the missing part.

5. A Semi-Parametric Data Augmentation Approach under MAR

In data augmentation under the assumption of MAR as specified in (1) the imputation and the posterior steps are as follows. Let us write $f(Y \mid W, \zeta)$, i.e. ζ is defined as the parameter for the complete data model $f(Y \mid W)$. The predictive distribution of the direct variable required for the *I*-step is $f(Y \mid W, \zeta)$ and the complete-data posterior for the *P*-step is $f(\zeta \mid Y, W)$.

5.1 The Imputation Step

Given a current estimate of the parameters $\zeta^{(d)}$, where d denotes the iteration of the data augmentation procedure, $d = 0,...,D$, the imputation step draws the imputed value from

$$\hat{y}_i^{(d+1)} \sim f(y_i \mid w_i, \zeta^{(d)}), \qquad (6)$$

where $\hat{y}_i^{(d+1)}$ is the imputed value for non-respondent i in iteration $d+1$ and $\zeta^{(d)}$ is the vector of parameters of the complete data model. The model $f(y_i \mid w_i, \zeta)$ is referred to as the imputation model.

Drawing values $\hat{y}_i^{(d+1)}$ from $f(y_i \mid w_i, \zeta^{(d)})$ is usually based on a parametric regression model. A standard approach would be to assume that

$$y_i \mid w_i, \zeta \sim N(\eta_i \beta; \sigma^2_{Y|W}), \qquad (7)$$

where η_i is a vector of covariates, functions of w_i, β is a vector of coefficients and $\sigma^2_{Y|W}$ denotes the conditional variance of y_i given w_i. The vector of parameters is $\zeta = (\beta', \sigma^2_{Y|W})'$. Regression imputation can then be performed adding a normal error to the predicted values from the model (David et al. 1986). Similar assumptions as in (7) have been made by Raghunathan et al. (2001) and by Heitjan and Rubin (1990), even in the presence of skewed data. However, they are unlikely to hold in many applications. It may for example be the case that the assumption of homoscedasticity may be violated in reality.

To relax the distributional assumptions made in (7), and to improve the robustness of the standard parametric approach, the imputation step may be modified by using predictive mean matching imputation. Predictive mean matching still makes an assumption about the form of the linear relationship but does not make any assumptions about the distribution of the residuals, such as constant variance and normality. Two forms of predictive mean matching are suggested to be implemented in the imputation step. These are:

a.) Hot deck imputation within classes: In each iteration, imputation classes, C_t, $t = 1,...,T$, are defined based on the range of the predicted values of the imputation model as specified in (7). For example, for each non-respondent in class C_t, 10 donor values are selected from the same class without replacement. Then, one donor value is selected at random from this set for imputation. After convergence of the data augmentation algorithm for example $M = 10$ imputed sets are selected. How to draw imputed values after convergence and inference under the data augmentation algorithm is discussed in section 7.

b.) Nearest neighbor imputation: for example, the 10 nearest neighbors for each non-respondent are defined and one donor value is selected at random for imputation. Then, for example $M = 10$ imputations are selected after convergence.

5.2 The Posterior Step

In the posterior step for iteration $d+1$, the required vector of parameters $\zeta^{(d+1)} = (\beta^{(d+1)\prime}, \sigma^2_{Y|W}{}^{(d+1)})'$ is drawn from the posterior distribution $f(\zeta \mid Y.^{(d+1)}, W)$, where $Y.^{(d+1)}$ denotes $(Y'_{obs}, \hat{Y}^{(d+1)\prime}_{mis})'$ and $\hat{Y}^{(d+1)}_{mis}$ are the imputed values from the I-step in iteration $(d+1)$. Under an uninformative prior for ζ and (7) the posterior distribution of interest, following derivations in Box and Tiao (1992), is

$$f(\zeta \mid Y, W) \propto (\sigma^2_{Y|W})^{-1/2} \prod_{i=1}^{n} (\sigma^2_{Y|W})^{-1/2} \exp\left\{ \frac{-1}{2\sigma^2_{Y|W}} (y_i - \eta_i \beta)^2 \right\}$$
$$= (\sigma^2_{Y|W})^{-(\frac{n+1}{2})} \exp\left\{ \frac{-1}{2\sigma^2_{Y|W}} \sum_{i=1}^{n} (y_i - \eta_i \beta)^2 \right\}. \qquad (8)$$

The required parameters for the P-step can therefore be drawn from the posterior distribution as

$$\sigma^2_{Y|W} \mid W \sim \hat{\varepsilon}'_{Y.} \hat{\varepsilon}_{Y.} \chi^{-2}_{n-2} \text{ and}$$

$$\beta \mid \sigma^2_{Y|W}, W \sim N(\hat{\beta}, \sigma^2_{Y|W} (\eta'\eta)^{-1}), \qquad (9)$$

where χ^{-2}_{n-2} is the inverted chi-square distribution, η defines the corresponding matrix to η_i, $\hat{\beta}$ is the maximum likelihood estimate, $\hat{\beta} = (\eta'\eta)^{-1} \eta Y.$, and $\hat{\varepsilon}_{Y.} = Y. - \eta \hat{\beta}$, both based on augmented data $Y.^{(d+1)}$ for iteration $(d+1)$.

Note that under the assumption of MAR and in the case of a monotone missing data structure, it is not necessary to use an iterative procedure such as data augmentation to obtain draws from the observed-data posterior. Under these conditions, it is possible to express the observed-data posterior in a tractable form. In the case of a monotone missing data structure and if the prior density factors into independent densities, then the observed-data posterior distribution also factors into independent posteriors and Bayesian inference is possible without iteration. In this case the observed-data posterior is the product of a multivariate normal and a scaled inverted-chisquare density based on observed data only, rather than on augmented data, i.e. observed and imputed data. Under such conditions, it would be enough to draw the parameters from their posterior distribution given the data as in Schenker and Taylor (1996). However, for illustrative purposes, since the interest is on computational properties of such an iterative method and to allow extensions to more complex problems a data augmentation procedure is used.

6. Semi-Parametric Data Augmentation Approach under Nonignorable Non-Response

Let us now turn to using multiple imputation under the CME assumption, representing nonignorable non-response as specified in (2). Here, a surrogate variable of y_i, namely x_i, is assumed to exist, which measures y_i with error but is fully observed. Under noningorable non-response the data augmentation procedure is more complex since the non-response model cannot be ignored and needs to be incorporated in the imputation procedure.

Given the data under CME, a model for Y, X, R conditional on W is considered which is expressed as

$$f(Y, X \mid W, \zeta) f(R \mid Y, X, W, \psi),$$

where ζ and ψ are the parameters of the complete data and the missing data mechanism respectively. A prior density for ζ and ψ is written as $f(\zeta, \psi)$. The predictive distribution of the variable y_i required for the *I*-step is $f(Y \mid X, R, W, \zeta, \psi)$ and the complete-data posterior required for the *P*-step is $f(\zeta, \psi \mid Y, X, R, W)$. It is convenient to express the parameter ζ as $(\zeta_1', \zeta_2')'$, where ζ_1 is the vector of parameters of $f(Y \mid X, W, \zeta_1)$ and ζ_2 is the vector of parameters of $f(X \mid W, \zeta_2)$. Using the CME assumption, the factorisation

$$\begin{aligned} & f(Y, X, R \mid W, \zeta, \psi) \\ & = f(Y \mid X, W, \zeta_1) f(X \mid W, \zeta_2) f(R \mid Y, W, \psi) \qquad (10) \end{aligned}$$

appears convenient for the implementation of the *I*- and the *P*-steps. This factorization into three models has a simple interpretation. The first model represents the predictive distribution of the true variable, the second the predictive distribution of the variable measured with error and the third factor represents a model for the non-response under the CME assumption. Let us now specify the imputation step and posterior step for data augmentation under the nonignorable non-response assumption.

6.1 The Imputation Step

The imputation step requires drawing imputed values for missing values of y_i from the predictive distribution $f(y_i \mid x_i, w_i, r_i = 0, \zeta, \psi)$. Using the CME assumption and (10), it is

$$\begin{aligned} f(y_i \mid x_i, w_i, r_i = 0, \zeta, \psi) &= \frac{f(y_i, x_i, r_i = 0 \mid w_i, \zeta, \psi)}{f(x_i, r_i = 0 \mid w_i, \zeta, \psi)} \\ = f(y_i \mid x_i, w_i, \zeta_1) & \frac{f(r_i = 0 \mid y_i, w_i, \psi)}{f(r_i = 0 \mid x_i, w_i, \zeta, \psi)} \end{aligned}$$

and therefore

$$\begin{aligned} & f(y_i \mid x_i, w_i, r_i = 0, \zeta, \psi) \\ & \propto f(y_i \mid x_i, w_i, \zeta_1) f(r_i = 0 \mid y_i, w_i, \psi). \qquad (11) \end{aligned}$$

The *I*-step may thus be implemented as follows. Given current values of the parameters $\zeta_1^{(d)}$ and $\psi^{(d)}$, where d denotes the iteration of the data augmentation procedure, a possible imputed value for non-respondent i, denoted $\hat{y}_i^{(d+1)*}$, is drawn

$$\hat{y}_i^{(d+1)*} \sim f(y_i \mid x_i, w_i, \zeta_1^{(d)}).$$

Rejection sampling (Tanner, 1996 and Gelman et al., 1998) is then performed based on the non-response model, accepting $\hat{y}_i^{(d+1)*}$ for imputation with probability $f(r_i = 0 \mid \hat{y}_i^{(d+1)*}, w_i, \psi^{(d)}) = \rho_i^{(d+1)*}$, where $\rho_i^{(d+1)*}$ denotes the probability of non-response. If accepted, it is $\hat{y}_i^{(d+1)*} = \hat{y}_i^{(d+1)}$, where $\hat{y}_i^{(d+1)}$ is the imputed value for non-respondent i at iteration $d+1$. If rejected, another value $\hat{y}_i^{(d+1)*}$ is drawn and so on. The *I*-step in (11) has therefore a simple interpretation and is easy to implement. The model $f(y_i \mid x_i, w_i, \zeta_1)$ is henceforth referred to as the imputation model and $f(r_i = 0 \mid y_i, w_i, \psi)$ as the non-response model. An advantage of the factorisation in (10) is that a model for X does not need to be fitted, and therefore no assumptions need to be made about this distribution.

To illustrate the procedure how to draw values $\hat{y}_i^{(d+1)*}$ from $f(y_i \mid x_i, w_i, \zeta_1^{(d)})$ in practice, initially a standard parametric regression model is described, which is then modified. The standard approach, similarly to (7), would be to assume that

$$y_i \mid x_i, w_i, \zeta_1 \sim N(\eta_i \beta; \sigma^2_{Y|X,W}), \qquad (12)$$

where η_i is a vector of covariates, functions of x_i and w_i, β is a vector of coefficients and $\sigma^2_{Y|X,W}$ denotes the conditional variance of y_i given x_i and w_i. The vector of parameters is $\zeta_1 = (\beta', \sigma^2_{Y|X,W})'$. As before, to relax the distributional assumptions, two forms of predictive mean matching imputation are proposed: a.) hot deck imputation within classes and b.) nearest neighbor imputation, where the classes and the nearest neighbors are defined based on the predictions of the regression model. Due to the nonignorable nonresponse model in the I-step these procedures are slightly more complex than under MAR.

Under hot deck imputation within classes Q donor values, denoted $\hat{y}_{i1}^{*},...,\hat{y}_{iQ}^{*}$, are selected with simple random sampling without replacement for non-respondent i in class C_t from that class. Under nearest neighbor imputation the $Q/2$ responding nearest neighbors above and below the predicted value for non-respondent i are used to obtain the Q possible values for imputation, where the value for Q is an even number, e.g. $Q = 10$. However, under hot deck imputation within classes and nearest neighbor imputation the number of values that can be chosen for imputation is restricted due to the definition of the classes and the nearest neighbors. The acceptance-rejection procedure based on the probability $\rho_i = 1 - f(r_i = 1 \mid y_i, w_i, \psi)$ is therefore modified using a weighted bootstrap method as described in Carroll et al. (1995) and Tanner (1996), since classical rejection sampling requires being able to generate a large number of potential imputed values, which is only possible under parametric random regression imputation. Under the weighted bootstrap method the value for imputation, $\hat{y}_i^{(d+1)}$, for iteration $d+1$, is sampled out of the Q possible values $\hat{y}_{i1}^{(d+1)*},...,\hat{y}_{iQ}^{(d+1)*}$ with probabilities

$$\breve{\rho}_{iq}^{(d+1)^{*}} =$$

$$f(r_i = 0 \mid y_{iq}^{(d+1)*}, w_i, \psi^{(d)}) / \sum_{q=1}^{Q} f(r_i = 0 \mid y_{iq}^{(d+1)*}, w_i, \psi^{(d)}) \qquad (13)$$

for all $q = 1,...,Q$. Note that under both, rejection sampling and the weighted bootstrap method, in each *I*-step only one value $\hat{y}_j^{(d+1)}$ is imputed for each non-respondent. The difference to the data augmentation procedure under MAR in section 5 is that the values are drawn with the addition of rejection sampling or weighted bootstrap from the predictive distribution of y_i.

6.2 Posterior Distributions for the Posterior Step

The *P*-step requires drawing values of the parameters from the complete data posterior distributions. The required posteriors are $f(\zeta_1 \mid Y, X, W)$ and $f(\psi \mid Y, R, W)$, i.e. the posteriors of ζ_1 and ψ. Following the imputation step a posterior for ζ_2 does not need to be fitted. The derivation of the posterior step under CME has been discussed in Durrant and Skinner (2006). Here, only the main results are presented. Under a computationally uninformative prior and assumption (12), the resulting posterior distribution for $\zeta_1 = (\beta', \sigma_{Y|X,W}^2)'$, discarding proportionality constants, can be expressed as

$$\begin{aligned} & f(\zeta_1 \mid Y, X, W) \\ & \propto (\sigma_{Y|X,W}^2)^{-3/2} \prod_{i=1}^{n} (\sigma_{Y|X,W}^2)^{-1/2} \exp\left\{ \frac{-1}{2\sigma_{Y|X,W}^2} (y_i - \eta_i\beta)^2 \right\} \\ & = (\sigma_{Y|X,W}^2)^{-(\frac{n+3}{2})} \exp\left\{ \frac{-1}{2\sigma_{Y|X,W}^2} \sum_{i=1}^{n} (y_i - \eta_i\beta)^2 \right\}. \qquad (14) \end{aligned}$$

In the special case that the data has a monotone missing-data pattern (Little and Rubin, 2002) and since the parameters are independent, the required parameters can be drawn from the posterior distribution similarly to the MAR case as follows

$$\sigma_{Y|X,W}^2 \mid Y, X, W \sim \hat{\varepsilon}_{Y.}' \hat{\varepsilon}_{Y.} \chi_{n-1}^{-2} \qquad \text{and}$$

$$\beta \mid \sigma_{Y|X,W}^2, Y, X, W \sim N(\hat{\beta}, \sigma_{Y|X,W}^2 (\eta'\eta)^{-1}), \qquad (15)$$

where η is the corresponding matrix to η_i, $\hat{\beta}$ is the maximum likelihood estimate, $\hat{\beta} = (\eta'\eta)^{-1}\eta Y.$, and $\hat{\varepsilon}_{Y.} = Y. - \eta\hat{\beta}$, both based on augmented data $Y.$.

To compute the posterior for ψ, $f(\psi \mid Y, R, W)$, also a noninformative prior is assumed. For the response model, let $f(r_i = 1 \mid y_i, w_i, \psi) = G(\tau_i\psi) = p_i$, where τ_i is a row-vector including functions of y_i and w_i, p_i denotes the probability of response and G the logistic regression model,

$$G(\tau_i\psi) = \frac{\exp(\tau_i\psi)}{1 + \exp(\tau_i\psi)}.$$

Following the approach adopted in Zellner and Rossi (1984) the posterior can be specified as

$$\psi \sim N(\hat{\psi}, T^{-1}). \qquad (16)$$

The matrix T is defined as

$$T = -\left[\frac{\partial^2 L(\psi)}{\partial\psi\partial\psi'} \right]_{\psi=\hat{\psi}} = \tau' V \tau,$$

where τ is a matrix including functions of Y and W and V is a diagonal matrix with element

$$v_i = \left[\frac{r_i}{G_i^2} + \frac{1 - r_i}{(1 - G_i)^2} \right] g_i^2 - \frac{(r_i - G_i) g_i'}{G_i(1 - G_i)}, \qquad (17)$$

where

$$G_i = G(\tau_i\hat{\psi}),$$

$$g_i = \left[\frac{dG(\mathrm{z}_i)}{d\mathrm{z}_i}\right]_{\mathrm{z}_i=\tau_i\hat{\psi}} = g(\tau_i\hat{\psi})$$, and

$$g'_i = \left[\frac{dg(\mathrm{z}_i)}{d\mathrm{z}_i}\right]_{\mathrm{z}_i=\tau_i\hat{\psi}}.$$

7. Inference Under Both Semi-Parametric Data Augmentation Procedures

Inference under both cases of data augmentation in section 5 and 6 follows the standard procedure for multiple imputations. Suppose that the data augmentation algorithm has run long enough to achieve approximate stationarity and to be independent of the initial starting values $\zeta^{(0)}$ for the MAR case and $\zeta_1^{(0)}$ and $\psi^{(0)}$ for nonignorable non-response, i.e. d is large enough such that the vectors of parameters $\zeta^{(d)}$ ($\zeta_1^{(d)}$ and $\psi^{(d)}$) are essentially draws from the observed-data posterior. Imputed values $\hat{y}_i^{(m)}$, $m = 1,\ldots,M$, $M > 1$ can be determined for each non-respondent i from repeated *I*-steps. The resulting point estimators from each of the M completed datasets, denoted $\hat{\theta}.^{(m)}(y)$ for $m = 1,\ldots,M$, can then be combined (Rubin, 1987) to give the point estimator:

$$\hat{\theta}.(y) = \frac{1}{M}\sum_{m=1}^{M}\hat{\theta}.^{(m)}(y). \qquad (18)$$

Under the model assumptions this estimator will be approximately unbiased. The method of multiple imputation, moreover, suggests a method of variance estimation in the context of data augmentation (Little and Rubin, 2002). For the purpose of variance estimation, the M sets of multiple imputations should not be obtained from successive sets of imputed values Y_{mis} since they are correlated. Instead, the Markov chain may be subsampled after an initial burn-in period using every k -th iterate to achieve approximate independence of repeated imputations. An estimator of the variance of $\hat{\theta}.(y)$ is then given by (Rubin, 1987):

$$\hat{\mathrm{var}}_{MI}(\hat{\theta}.(y)) = \overline{A}. + (1 + 1/M)\hat{B}. \qquad (19)$$

where

$$\overline{A}. = \frac{1}{M}\sum_{m=1}^{M}\hat{A}.^{(m)}$$

is the within imputation variance, and $\hat{A}.^{(m)}$ is the standard variance estimator valid for complete data, applied to Y_{obs} and the imputed values $Y_{mis}^{(m)}$ for the m -th imputation, and

$$\hat{B}. = \frac{1}{M-1}\sum_{m=1}^{M}(\hat{\theta}.^{(m)}(y) - \hat{\theta}.(y))^2$$

is the between imputation variance.

An application of the semi-parametric data augmentation method and a simulation study evaluating the properties of this method for a specific example on hourly pay distributions are described in Durrant and Skinner (2006). They find in their application that standard parametric imputation approaches, including fully parametric multiple imputation, do not perform well, for skewed and truncated hourly pay data. However, the semi-parametric method described above incorporating predictive mean matching imputation in the imputation step of the data augmentation procedure instead of fully parametric regression imputation performs much better and shows robustness against model misspecifications. This coincides with findings described in Schenker and Taylor (1996), investigating fully parametric imputation methods under model misspecifications and comparing the performance to a partially parametric predictive mean matching method.

8. Conclusions

The combination of multiple imputation and hot deck imputation can have several advantages. Such an approach can make use of the flexibility of the MI procedure, with the advantage of providing a relatively simple variance estimation formula, as well as preserving attractive features of hot deck imputation, such as making less parametric assumptions, and being able to impute actually observed values. Further work is needed to evaluate the properties of such a method in more detail and to develop the combination of multiple imputation and hot deck procedures for the multivariate missing data case.

9. REFERENCES

Box, G.E.P. and Tiao, G.G. (1992): *Bayesian Inference in Statistical Analysis*, Reading.

Carroll, R.J., Ruppert, D. and Stefanski, L.A. (1995): *Measurement Error in Nonlinear Models*, London.

David, M., Little, R.J.A., Samuhel, M.E. and Triest, R.K. (1986): Alternative Methods for CPS Income Imputation, *Journal of the American Statistical Association*, 81, 29-41.

Durrant, G.B. and Skinner, C. (2006): Using Data Augmentation to Correct for Nonignorable Non-Response when Surrogate Data are Available: An Application to the Distribution of Hourly Pay, *Journal of the Royal Statistical Society, Series A*, 169, 3, to appear.

Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (2004): *Bayesian Data Analysis*, London.

Heitjan, D.F. and Little, R. (1991): Multiple Imputation for the Fatal Accident Reporting System, *Applied Statistics*, 40, 13-29.

Heitjan, D.F. and Rubin, D.B. (1990) Inference from Coarse Data via Multiple Imputation with Application to Age Heaping, *Journal of the American Statistical Association*, 85, 304-314.

Kim, J.K. and Fuller, W. (2004): Fractional Hot Deck Imputation, *Biometrika*, 91, 3, 559-578.

Lipsitz, S.R., Zhao, L.P. and Molenberghs, G. (1998): A Semiparametric Method of Multiple Imputation, *Journal of the Royal Statistical Society, Series B*, 60, 1, 127-144.

Little, R.J.A. (1988): Missing-Data Adjustments in Large Surveys, *Journal of Business and Economic Statistics*, 6, 287-301.

Little, R.J.A. and Rubin, D.B. (2002): *Statistical Analysis with Missing Data*, New York.

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

Rubin, D.B. (1987): *Multiple Imputation for Non-Response in Surveys*, New York.

Rubin, D.B. (1996): Multiple Imputation after 18+ Years, *Journal of the American Statistical Association*, 91, 434, 473-489.

Rubin, D.B. and Schenker. N. (1986): Multiple Imputation for Interval Estimation from Simple Random Samples With Ignorable Non-Response, *Journal of the American Statistical Association*, 81, 394, 366-374.

Schafer, J.L. (1997): *Analysis of Incomplete Multivariate Data*, London.

Schafer, J.L. (1999): Multiple Imputation: A Primer, *Statistical Methods in Medical Research*, 8, 3-15.

Schenker, N. and Taylor, J.M.G. (1996): Partially Parametric Techniques for Multiple Imputation, *Computational Statistics and Data Analysis*, 22, 425-446.

Tanner, M.A. (1996): *Tools for Statistical Inference, Methods for the Exploration of Posterior Distributions and Likelihood Functions*, Springer, New York.

Tanner, M.A. and Wong, W.H. (1987): The Calculation of Posterior Distributions by Data Augmentation, *Journal of the American Statistical Association*, 82, 398, 528-540.

Zellner, A. and Rossi, P.E. (1984): Bayesian Analysis of Dichotomous Quantal Response Models, *Journal of Econometrics*, 365-393.

Zhang, P. (2003): Multiple Imputation: Theory and Method, *International Statistical Review*, 71, 3, 581-592 (with discussions).