Primer: Statistical Inference and Computational Formulae for Missing Data, Synthetic Data, Partially Synthetic Data, and Partially Synthetic Data with Missing Data Using Multple Imputation

John M. Abowd

April 12, 2005 (incomplete)

1 Introduction

This primer contains a summary of the statistical formulae appropriate for scientific inference when multiple imputation methods have been used for either missing data or confidentiality protection. The classical use of multiple imputation is the

imputation of missing values in general purpose data files. The newer use is the creation of synthetic or partially synthetic data as a general confidentiality protection mechanism.

2 Notation and Definitions

The notation and definitions follow Rubin (1987) with enhancements for synthetic data from Raghunathan *et al.* (2003), partially synthetic data from Reiter (2003), and partially synthetic data with missing data from Reiter (2004). I have attempted to make the notation consistent in this primer. Hence, it does not match the original authors' notation. Some formulae have been stated below as matrix generalizations of univariate formulae published by the referenced authors. In particular, inference theory based on the stated asymptotoc normal distribution is correct for the multivariate formulae; however, asymptotics require both the sample size and the number of independent imputations to be large. Inference theory for moderate numbers of implicates has only been established in the univariate case.

Although all the techniques discussed in this primer are based on Bayesian methods, the inference system supported by the computational formulae is frequentist. See Rubin (1996) for an explanation. Statistical validity is defined using finite population statistics. Quoting Rubin:

> First and foremost, for statistical validity for scientific estimands, point estimation must be approximately unbiased for the scientific estimands averaging over the sampling and posited nonresponse mechanisms. ... Second, interval estimation and hypothesis testing must be valid in the sense that nominal levels describe operating characteristics over sampling and posited nonresponse mechanisms. (1996, p. 474)

This definition should be modified to include the phrase "confidentiality protection mechanisms" wherever "nonresponse mechanisms" appears. It is in this sense that the methods outlined below deliver inference-valid general purpose data files.

A finite population contains N entities whose population frame characteristics are known and constitute the f columns of X, $(N \times f)$. A sample of size $n < N$ is drawn from the population. Let the vector I $(N \times 1)$ be defined as $I_i = 1$ if entity i is sampled and $I_i = 0$, otherwise. Data are collected for p variables denoted by the matrix Y $(N \times p)$. Note that the matrix Y is defined for the entire population, not just for the sampled units. Of course, some elements of Y are missing because the entity that they characterize was not sampled. Other elements of Y are missing because of item nonresponse in the sample. (In administrative record data item nonresponse is equivalent to missing data items on an in-scope administrative record.) Let the matrix R $(N \times p)$ be defined as $R_{ij} = 1$ if the data represented by item Y_{ij}

are available in the sample and $R_{ij} = 0$, otherwise. Certain submatrices of Y and R are of interest. Let Y_{inc} $(n \times p)$ be the submatrix of Y that corresponds to the rows for which $I_i = 1$. So Y_{inc} contains the data for all the sampled entities. The complement of Y_{inc} is Y_{exc}, the rows of Y that correspond to the rows for which $I_i = 0$. So Y_{inc} contains the data for all the unsampled entities..Similarly, let R_{obs} $(n \times p)$ be the submatrix of R corresponding to the item missingness for the sampled entities; i.e., those rows for which $I_i = 1$. Finally, define the submatrices Y_{obs} and Y_{mis} as follows

$$Y_{obs,ij} = \left\{ \begin{array}{c} Y_{ij}, \text{ if } I_i = 1 \text{ and } R_{ij} = 1 \\ \\ undefined, \text{ otherwise} \end{array} \right\}$$

and

$$Y_{mis,ij} = \left\{ \begin{array}{c} Y_{ij}, \text{ if } I_i = 1 \text{ and } R_{ij} = 0 \\ \\ undefined, \text{ otherwise} \end{array} \right\}$$

So, the matrix Y_{obs} contains all the sampled values of Y_{ij} that contain data and the matrix Y_{mis} contains all the sampled values of Y_{ij} that are item missing. The observed data are summarized by the set D=$\{X, Y_{obs}, I, R\}$.

3 Complete Data Estimation

Interest focuses on a complete data estimand $Q(X,Y)$ $(c \times 1)$. This estimand can be any computable, vector-valued function of the data. For example, it could be the average value of Y, many moments of Y, conditional moments of Y, given X, parameters of a model relating columns of (X,Y), and so on. The essential feature of Q is that it is computable from complete data on the population and, therefore, is not random.

Estimates of Q are random because they are based on D, which involves sampling from the finite population and incomplete observation of Y in the sample. If $R_{obs} = 1_{np}$, a matrix of ones, then there are no missing data. An estimator of Q is random because of the sample design embodied in I. The complete data estimator is $q(D)$ and its variance estimator is $u(D)$. Notice that because of the definition of complete data q and u depend only on (X, Y_{obs}, I). The analyst is assumed to have an inference system for $q(D)$ and $u(D)$. In particular, complete data inference is based on $(q(D) - Q) \sim N(0, u(D))$, which may be exact or an approximation but is taken as given.

4 Inference Frameworks Using Multiple Imputation

4.1 Missing Data Only

In the classic Rubin (1987) missing data application Y_{mis} is imputed m times by sampling from $p\left(Y_{mis}\left|D\right.\right)$, the posterior predictive distribution of Y_{mis} given D. The completed data consist of m sets $D^{(\ell)} = \left\{D, Y_{mis}^{(\ell)}\right\}$, where $Y_{mis}^{(\ell)}$ is the ℓ^{th} draw from $p\left(Y_{mis}\left|D\right.\right)$ and is called the ℓ^{th} implicate. Inference is based on the following formulae:

$$q^{(\ell)} = q\left(D^{(\ell)}\right)$$

$$\bar{q}_m = \sum_{\ell=1}^{m} \frac{q^{(\ell)}}{m}$$

$$b_m = \sum_{\ell=1}^{m} \frac{\left(q^{(\ell)} - \bar{q}_m\right)\left(q^{(\ell)} - \bar{q}_m\right)'}{m-1}$$

$$u^{(\ell)} = u\left(D^{(\ell)}\right)$$

$$\bar{u}_m = \sum_{\ell=1}^{m} \frac{u^{(\ell)}}{m}$$

$$T_m = \bar{u}_m + \left(1 + \frac{1}{m}\right) b_m$$

$$\text{if } c = 1, \text{ then } \nu_m = (m-1)\left(1 + \frac{\bar{u}_m}{\left(1+\frac{1}{m}\right) b_m}\right)^2$$

When n and m are large, inference is based on $(\bar{q}_m - Q) \sim N(0, T_m)$. When m is moderate and the estimator $\bar{q}_m$ is univariate, inference is based on $(\bar{q}_m - Q) \sim t_{\nu_m}(0, T_m)$. Proofs and further details can be found in Rubin (1987, 1996).

4.2 Synthetic Data Only

If there are no missing data, then fully synthetic data (Rubin, 1993; Raghunathan et al. 2003) are constructed by forming r synthetic populations as follows. For $\ell = 1, \ldots, r$, draw $Y_{exc}^{(\ell)}$ from $p(Y_{exc}|D)$. Since there are no missing data, $Y_{obs} = Y_{inc}$ and $R_{obs} = 1_{np}$. Hence $D = \{X, Y_{inc}, I\}$. Call $Y^{(\ell)} = \left(Y_{inc}, Y_{exc}^{(\ell)}\right)$. Synthetic population ℓ is defined as $\left(X, Y^{(\ell)}\right)$. Synthetic sample ℓ is a simple random sample of the rows of $\left(X, Y^{(\ell)}\right)$ denoted by $I^{(\ell)}$, where $I_i^{(\ell)} = 1$ for entities in the ℓ^{th} synthetic sample. Let $D^{(\ell)} = \left\{X, Y^{(\ell)} | I_i^{(\ell)} = 1\right\}$, where the notation means that only the rows of $\left(X, Y^{(\ell)}\right)$ corresponding to $I_i^{(\ell)} = 1$ are included in the set $D^{(\ell)}$. Inference is based on the following formulae:

$$q^{(\ell)} = q\left(D^{(\ell)}\right)$$

$$\bar{q}_r = \sum_{\ell=1}^{r} \frac{q^{(\ell)}}{r}$$

$$b_r = \sum_{\ell=1}^{r} \frac{\left(q^{(\ell)} - \bar{q}_r\right)\left(q^{(\ell)} - \bar{q}_r\right)'}{r-1}$$

$$u^{(\ell)} = u\left(D^{(\ell)}\right)$$

$$\bar{u}_r = \sum_{\ell=1}^{r} \frac{u^{(\ell)}}{r}$$

$$T_r = \left(1 + \frac{1}{r}\right) b_r - \bar{u}_r$$

When n and r are large, inference is based on $(\bar{q}_r - Q) \sim N(0, T_r)$. The theoretical discussion and proofs are in Raghunathan et al (2003).

4.3 Partially Synthetic Data Only

If there are no missing data but the synthetic data are constructed using the same sampling plan as the original data, $I_{inc}^{(\ell)} = I_{inc}$ for all $\ell = 1, \ldots, r$, then the data are called partially synthetic. Assuming that there are no missing data $Y_{obs} = Y_{inc}$ and $R_{obs} = 1_{np}$. Let the vector Z $(n \times 1)$ denote entities i for which values any values of Y_{obs} have been synthesized. So, $Z_i = 1$ if any of the values of $Y_{obs,i}$ have been synthesized. Partition Y_{obs} into Y_{nrep} containing the rows where $Z_i = 0$ and Y_{rep} containing the rows where $Z_i = 1$. Construct the ℓ^{th} synthetic implicate by drawing Y_{rep} from $p(Y_{rep}|D, Z)$. Denote the ℓ^{th} synthetic data set as $D^{(\ell)} = \left\{X, Y_{nrep}, Y_{rep}^{(\ell)}, I, Z\right\}$.

Inference is based on the following formulae:

$$q^{(\ell)} = q\left(D^{(\ell)}\right)$$

$$\bar{q}_r = \sum_{\ell=1}^{r} \frac{q^{(\ell)}}{r}$$

$$b_r = \sum_{\ell=1}^{r} \frac{\left(q^{(\ell)} - \bar{q}_r\right)\left(q^{(\ell)} - \bar{q}_r\right)'}{r-1}$$

$$u^{(\ell)} = u\left(D^{(\ell)}\right)$$

$$\bar{u}_r = \sum_{\ell=1}^{r} \frac{u^{(\ell)}}{r}$$

$$T_r = \bar{u}_r + \frac{b_r}{r}$$

$$\text{if } c = 1, \text{ then } \nu_r = (r-1)\left(1 + \frac{\bar{u}_r}{b_r/r}\right)^2$$

When n and r are large, inference is based on $(\bar{q}_r - Q) \sim N(0, T_r)$. When m is moderate and the estimator $\bar{q}_r$ is univariate, inference is based on $(\bar{q}_r - Q) \sim t_{\nu_r}(0, T_r)$. Proofs and details are contained in Reiter (2003).

4.4 Missing and Partially Synthetic Data

To combine the analyses in sections (4.1) and (4.3) the missing data imputation and the synthetic data sampling must be done sequentially. First, complete m versions of D by sampling from $p\left(Y_{mis}|D\right)$. Denote the m completed data sets as $D^{(\ell)} = \left\{X, Y_{obs}, Y_{mis}^{(\ell)}, I, R\right\}$. Then, for each completed data set, partially synthesize r implicates by sampling from $p\left(Y_{rep}|D^{(\ell)}, Z\right)$. Denote the r completed partially synthetic data sets as $D^{(\ell,k)} = \left\{X, Y_{nrep}^{(\ell)}, Y_{rep}^{(\ell,k)}, I, R, Z\right\}$, where $Y_{nrep}^{(\ell)}$ corresponds to the rows of $\left(Y_{obs}, Y_{mis}^{(\ell)}\right)$ for which $Z_i = 0$ and $Y_{rep}^{(\ell,k)}$ corresponds to the rows of $\left(Y_{obs}, Y_{mis}^{(\ell)}\right)$ for which $Z_i = 1$. Note that $Y_{nrep}^{(\ell)}$ contains no synthetic data but may contain missing data imputations whereas $Y_{rep}^{(\ell,k)}$ may contain both missing data implicates (an element of $Y_{rep}^{(\ell,k)}$, say ij, for which item j is missing for entity i but not synthesized; entity i is in this set because $Z_i = 1$ whenever any element of Y_{inc} is synthesized) and synthetic data (an element of $Y_{rep}^{(\ell,k)}$, say ij, for which item j is missing for entity i and is synthesized; entity i is in this set because $Z_i = 1$ and element j element of $Y_{inc,i}$ is synthesized). Inference is based on the following formulae:

$$q^{(\ell,k)} = q\left(D^{(\ell,k)}\right)$$

$$\bar{q}^{(\ell)} = \sum_{k=1}^{r} \frac{q^{(\ell,k)}}{r}$$

$$\bar{q}_M = \sum_{\ell=1}^{m} \sum_{k=1}^{r} \frac{q^{(\ell,k)}}{mr} = \sum_{\ell=1}^{m} \frac{\bar{q}^{(\ell)}}{m}$$

$$b^{(\ell)} = \sum_{k=1}^{r} \frac{\left(q^{(\ell,k)} - \bar{q}^{(\ell)}\right)\left(q^{(\ell,k)} - \bar{q}^{(\ell)}\right)'}{r-1}$$

$$\bar{b}_M = \sum_{\ell=1}^{m} \sum_{k=1}^{r} \frac{\left(q^{(\ell,k)} - \bar{q}^{(\ell)}\right)\left(q^{(\ell,k)} - \bar{q}^{(\ell)}\right)'}{m\left(r-1\right)} = \sum_{\ell=1}^{m} \frac{b^{(\ell)}}{m}$$

$$B_M = \sum_{\ell=1}^{m} \frac{\left(\bar{q}^{(\ell)} - \bar{q}_M\right)\left(\bar{q}^{(\ell)} - \bar{q}_M\right)'}{m-1}$$

$$u^{(\ell,k)} = u\left(D^{(\ell,k)}\right)$$

$$\bar{u}^{(\ell)} = \sum_{k=1}^{r} \frac{u^{(\ell,k)}}{r}$$

$$\bar{u}_M = \sum_{\ell=1}^{m} \sum_{k=1}^{r} \frac{u^{(\ell,k)}}{mr} = \sum_{\ell=1}^{m} \frac{\bar{u}^{(\ell)}}{m}$$

$$T_M = \left(1 + \frac{1}{m}\right) B_M - \frac{\bar{b}_M}{r} + \bar{u}_M$$

$$\text{if } c = 1, \text{ then } \nu_M = \frac{1}{\left(\frac{\left(\left(1+\frac{1}{m}\right)B_M\right)^2}{(m-1)T_M^2} + \frac{\left(\bar{b}_M/r\right)^2}{m(r-1)T_M^2}\right)}$$

When n, m and r are large, inference is based on $(\bar{q}_M - Q) \sim N\left(0, T_M\right)$. When m and r are moderate and the estimator $\bar{q}_M$ is univariate, inference is based on $(\bar{q}_M - Q) \sim t_{\nu_M}\left(0, T_M\right)$. Proofs and details can be found in Reiter (2004).

5 Methods for Computing Posterior Predictive Distributions

In all of the sections above, reference is made to sampling from posterior predictive distributions such as $p\left(Y_{mis}|D\right)$. This section discusses computational formulae for performing this sampling. More general methods exist, such as Markov Chain Monte Carlo, but to-date the methods summarized herein are the ones used by various parts of synthetic data projects in LEHD.

5.1 General Setup

An explicit representation of D is required. As defined above $D = \{X, Y_{obs}, I, R\}$. While in principle the analyst at Census has access to X, the population frame characteristics, in the applications described in this section, only the rows of X corresponding to $I_i = 1$ are used. Hence, there is no practical difference between X and Y_{obs}. Complete data are guaranteed for X but many variables in X require confidentiality protection before they can be placed in a public use data file. In this section, I adopt the notational convention that a variable appears in X if it is always available when $I_i = 1$ and it never requires confidentiality protection. Otherwise, the variable is included in Y_{obs}. This set of X variables can be empty without affecting

the discussion below.

In all of the methods described below, the columns of D are partitioned into four mutually exclusive sets: grouping variables, conditioning variables, dependent variables, and ignored variables. Grouping variables are used to stratify D such that a separate posterior distribution is estimated in each stratum. Conditioning variables are a list of potential right-hand-side variables to be entered linearly in model-based estimation of the posterior predictive distribution. Dependent variables are those for which the posterior distribution is being estimated, Finally, ignored variables are all other columns of (X, Y_{obs}). For purposes of doing the computations below, the data matrix (X, Y_{obs}) should be interpreted as including any variables that have been calculated as exact functions of the available data. Hence, the dimensionality of the matrices used below potentially exceeds $f + p$.

5.2 Bayesian Bootstrap

The Bayesian bootstrap (BB) was originally defined by Rubin (1981). As explained therein, the BB is used to simulate the posterior distribution of parameter whereas the regular bootstrap simulates the sampling distribution of the parameter. Whereas a conventional bootstrap assumes that the sample CDF is equal to the population CDF, the BB properly accounts for the uncertainty of the sample CDF.

5.2.1 Generic BB Algorithm

Notation to describe the BB algorithm is generic and does not refer to the matrices defined elsewhere. Let X $(n \times k)$ be the source data matrix and Y $(s \times k)$.be the target data matrix. This means that we want to construct an $s \times k$ Bayesian bootstrap sample from an $n \times k$ matrix of source data. Each BB replicate ℓ is a unique $Y^{(\ell)}$.

1. Draw $n-1$ random variables from $U(0,1)$.
2. Sort u_i ascending and let $u_{(i)}$ denote the order statistics from lowest to highest. Define $u_{(0)} = 0$ and $u_{(n)} = 1$.
3. For $i = 1, \ldots, n$, let $\hat{p}_i = u_{(i)} - u_{(i-1)}$.
4. For $j = 1, \ldots, s$ sample with replacement from the rows X using $\hat{p}_i$ as the probability of selecting row i. Place the sampled row into Y_j.
5. Repeat from step 1 for as many BB replicates as desired.

5.2.2 Application to Missing Data

Choose grouping variables such that the rows of (X, Y_{obs}) can be assumed to come from the same joint distribution within each group defined by the unique combinations of values of the grouping variables. Some collapsing of categories may be needed but I won't clutter the description with a formalization of that collapsing.

What is required is the creation of G groups based on the values of the variables in the grouping variable list. In the BB application, none of the grouping variables can contain missing data. There are no conditioning variables because no linear model is used. The dependent variables consist of all columns j of Y for which $R_{ij} = 0$ for some i. The ignored variable list consists of all variables that are neither grouping variables nor dependent variables.

The application of BB to the missing data problem is complicated if the missing data pattern is non-monotone (Rubin 1987). Assume that the missing data pattern is monotone. Then, proceed through the dependent variables in groups constructed as follows:

1. All dependent variables with missing data exactly comparable to the variable with the least missing data; *i.e.,* all j for which $R_{ij} = 0$ if and only if $R_{ij^*} = 0$, where j^* is the column index of the variable with the least missing data. This is dependent variable group 1.

2. Remove all variables from the dependent variable list that are already in a group. Let j^* represent the column index of the variable with the least missing data from among those dependent variables that remain. Group all dependent variables with missing data exactly comparable to the variable indexed by j^*; *i.e.,* all j for which $R_{ij} = 0$ if and only if $R_{ij^*} = 0$. This is dependent variable

group h.

3. Repeat 2 until no dependent variables remain.

This defines H dependent variable groups. Initialize the BB missing data algorithm by placing all dependent variables into the ignored variable list and setting $h = 1$.

1. Remove the variables in group h from the ignored variable list and place them in the dependent variable list.

2. For $g = 1, \ldots G$, BB the rows of Y_{mis} (target data matrix) using the rows of Y_{obs} as the source data matrix. Repeat the BB m times to get m imputations $Y_{mis}^{(\ell)}$.

3. Put the dependent variables in group h back into the list of ignored variables.

4. If $h < H$ then increment h and return to step 1; otherwise, stop.

The result is m completed data sets for which the formulae in section (4.1) apply. When the missing data are not monotone, the BB algorithm can be used to get starting values for other algorithms described below, in particular, Sequential Regression Multivariate Imputation.

5.2.3 Application to Synthetic Data

This is idea is originally due to Rubin (1993) and was extended by Feinberg (1994), a paper that is very difficult to obtain because it was never published. The idea is repeated in Feinberg *et al.* (1998). The simplest version is easy to state and is included here for completeness but it can only be applied to very special case without modification.

Recall that there are no missing data. Choose grouping variables from X such that rows of (X, Y_{inc}) can be assumed to come from the same joint distribution within each group $g = 1, \ldots G$, as described above in section (5.2.2) . Dependent variables consist of all columns of Y. There are no conditioning or ignored variables. For $g = 1, \ldots G$, BB the rows of Y_{exc} using the rows of Y_{inc} as the data matrix. Repeat the BB r times to form $Y_{exc}^{(\ell)}$. For each BB synthetic population $\left(X, Y^{\ell}\right)$, randomly sample n rows to form the synthetic data matrix $D^{(\ell)}$. Apply the inferential formulae from section (4.2) . See Rubin (1993) and Feinberg et al. (1998) for caveats in using BB excusively as the confidentiality protection method.

5.2.4 Application to Partially Synthetic Data

The BB method cannot be applied to partially synthetic data because no synthetic population is formed.

5.2.5 Application to Partially Synthetic Data with Missing Data

The BB method can be applied to the missing data imputation to generate the m completed data sets, which are then synthesized using a different method for estimating the posterior predictive distribution. BB is also useful for generating starting values for missing data imputation in the case of non-monotone missing data imputation.

5.3 Sequential Regression Multivariate Imputation

Sequential Regression Multivariate Imputation (SRMI) was first proposed by as a general technique for multiple imputation of missing data. See Schenker et al. (2002) for an extensive discussion of the SRMI computing code developed at the Survey Research Center at the University of Michigan. Raghunathan et al. (2003) extend the method to confidentiality protection. Abowd and Woodcock (2001) use the SRMI method for confidentiality protection combined with missing data imputation. Abowd and Woodcock (2004) propose the kernel density estimation extension discussed below. Although the formulae for SRMI can be stated generically using joint probability distributions like $p\left(Y_{mis}|D\right)$, almost all applications assume that the entities that constitute the rows of (X, Y_{inc}) have been sampled independently. Nothing in the generic statement of the problem prohibits dependent sampling; however, as

a practical matter, formalizing this dependence while implementing SRMI is complicated. Abowd and Woodcock (2001) illustrate these complications for the case of longitudinally linked employer-employee data. The algorithms are summarized below ignoring the complications associated with dependent sampling until section (5.3.11) .

5.3.1 Definitions and General Algorithm

In SRMI, the analyst cycles through the dependent variable list. Let Y_j denote the current dependent variable and let $Y_{\sim j}$ denote all other columns of Y. The general algorithm is most cleanly stated for the missing data case.

For each dependent variable, the analyst selects grouping variables, conditioning variables and ignored variables. The grouping variables stratify the estimation into G mutually exclusive and exhaustive groups as illustrated in section (5.2.2) . The conditioning variables may include all columns of $(X, Y_{\sim j})$. The ignored variables are all columns of $(X, Y_{\sim j})$ not included among the conditioning variables. We wish to generate m implicates $Y_{mis}^{(\ell)}$. SRMI is an iterative procedure. Denote the interim values of implicate ℓ as $Y_{mis}^{(\ell,s)}$. Initialize $\ell = 1$ and $s = 1$. Initialize $Y_{mis}^{(1,0)}$ using BB.

1. For $j = 1, \ldots, p$:

(a) If $\ell = 1$ then estimate

$$p\left(Y_j|X, Y_{obs,\tilde{}j}, Y_{mis,1}^{(\ell,s-1)}, \ldots, Y_{mis,j-1}^{(\ell,s-1)}, Y_{mis,j+1}^{(\ell,s)}, \ldots, Y_{mis,p}^{(\ell,s)}\right)$$

(b) Fill $Y_{mis,j}^{(\ell,s)}$ with data sampled from

$$p\left(Y_j|X, Y_{obs,\tilde{}j}, Y_{mis,1}^{(\ell,s-1)}, \ldots, Y_{mis,j-1}^{(\ell,s-1)}, Y_{mis,j+1}^{(\ell,s)}, \ldots, Y_{mis,p}^{(\ell,s)}\right)$$

2. If converged then

(a) Set $Y_{mis}^{(\ell)} = Y_{mis}^{(\ell,s)}$.

(b) Increment ℓ.

(c) Reinitialize $Y_{mis}^{(\ell,0)} = Y_{mis}^{(\ell-1,s)}$

(d) Reinitialize $s = 1$.

3. If $\ell \leq m$, go to 1.

The test for convergence is not formal. In practice s is often limited to 10 or less. The algorithm estimates the joint distribution $p\left(Y_{mis}|D\right)$ by iterating over each conditional distribution $p\left(Y_{mis,j}|D\right)$ and filling the "data matrix" with imputed values based on the previous iteration's estimate of $p\left(Y_{mis}|D\right)$. Once the estimation has

converged, the implicates are all drawn from the same estimate of $p(Y_{mis}|D)$ however, the completion of D for each implicate results in different conditioning data for the draws. In the implementation of the algorithm, one cycles over the grouping variables $g = 1, \ldots, G$ performing the entire algorithm for each homogeneous group. In steps $1.a$ and $1.b$ only the conditioning variables appropriate for $Y_{mis,j}$ in conditioning group g are actually included in the conditioning set. The selection of these variables is dependent on the analyst but see section (5.3.3).

Raghunathan, T. E., Reiter, J. P., and Rubin, D. B. (2003). "Multiple imputation for statistical disclosure limitation." *Journal of Official Statistics* 19, 1–16.

Reiter, J. P. (2003). "Inference for partially synthetic, public use microdata sets," *Survey Methodology* 181–189.

Reiter, J. P. (2004). "Simultaneous Use of Multiple Imputation for Missing Data and Disclosure Limitation," working paper.

Rubin, D. B. (1987) *Multiple Imputation for Nonresponse in Surveys* (Hoboken, NJ, Wiley Classics Library).

Rubin, D.B. (1993), "Satisfying Confidentiality Constraints Through Use of Synthetic Multiply-imputed Microdata," (published title: "Discussion: Statistical Disclosure Limitation") *Journal of Official Statistics* 9, 461-468.

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