TNO report
PG/VGZ/99.054

Flexible multivariate imputation by MICE

TNO Prevention and Health

Public Health
Wassenaarseweg 56
P.O.Box 2215
2301 CE Leiden
The Netherlands

Tel + 31 71 518 18 18
Fax + 31 71 518 19 20

Date

Oktober 1999

Authors

Stef van Buuren
Karin Oudshoorn

TNO Prevention and Health contributes to the improvement of quality of life and to an increased healthy human life expectancy. Research and consultancy activities aim at improving health and health care in all stages of life.

Netherlands Organization for Applied Scientific Research (TNO)

Author

S. van Buuren

Project number

011.40684/01.01

ISBN-number

90-6743-623-2

This report can be ordered from TNO-PG by transferring ƒ 21,- (incl. VAT) to account number 99.889 of TNO-PG Leiden. Please state TNO-PG publication number PG/VGZ/99.054

Executive summary

Several approaches for (multiple) imputation of multivariate data have been proposed recently. Schafer (1997) presents a methodology to describe the data by an encompassing multivariate model, derives theoretically sound posterior distributions under this model, and draws imputations from these by Gibbs sampling and other methods. The `transcan` function (Alzola & Harrell, 1999) imputes each incomplete variable by cubic spline regression given all other variables.

Multivariate Imputation by Chained Equations (MICE) is an attempt to combine the most attractive aspects of both approaches. The MICE user specifies a conditional distribution for the missing data in each incomplete variable, for example in the form of a linear or (polytomous) logistic regression of the incomplete column given a set of predictors. Predictors themselves can be incomplete. It is assumed that a multivariate distribution exists from which these conditional distributions can be derived, and that iterative Gibbs sampling from the conditionals can generate draws from it.

The algorithm is implemented as an S-PLUS function. For each incomplete variable the user can choose a set of predictors that will be used for imputation. This is useful for imputing large data sets containing hundreds of variables. Passive imputation is a built-in feature that takes care that transformed data are always in sync with their original values. This can be used, for example, to impute categorical variables along with their dummies (needed for imputing other variables). In addition, the user can alter the visiting scheme of the Gibbs sampler, or plug-in his or her own imputation method. Features like these make it easy to include complex imputation constraints in a practical but principled way.

Key words:
item nonresponse, Gibbs sampler, large data sets, multiple imputation, imputation strategy

Contents

1 Introduction

Nonresponse in surveys can cause substantial loss of information in analysis of multivariate data. For example, suppose that one wishes to fit a regression model on survey data with 10 explanatory variables, and that each of these has randomly 10% of missing entries. Then, on the average 65 percent of the units will have at least one missing value. Besides potential problems regarding selective nonresponse, it will be clear that simply deleting incomplete records amounts to substantial losses of costly collected data.

Multiple imputation (Rubin 1987, 1996) is one of the best, currently available and general techniques to deal with nonresponse. Rubin's book, however, does not contain methods for imputing *multivariate data*, as is typical in surveys. Specific practical problems in multivariate data imputation are:

- For large sets of data, it is necessary to select a sensible set of potential predictor variables used for imputation;

- Predictors themselves may be incomplete, leading to a cascade of imputation problems;

- Circularities may occur, where Y_1 is imputed given Y_2, and Y_2 given Y_1;

- The order in which data are imputed can be meaningful, e.g. in experiments with repeated measurements;

- Transformed versions of imputed data might be needed, e.g. Y_1 and $\log(Y_1)$, or continuous and discretized versions of the same data;

- Variables can have different measurement levels: nominal, ordinal or interval;

- The optimal imputation model may be nonlinear, and could contain interaction terms;

- The units could be weighted to account for the sampling design;

- The complete data models applied to the imputed multivariate data could be quite different.

Several approaches to imputing multivariate data have been developed over the last decade. Li (1988) and Rubin and Schafer (1990) presented techniques for generating multivariate multiple imputations. Both are Bayesian simulation algorithms that draw imputations from the posterior predictive distribution of the missing data given the observed data. The Rubin-Schafer method assumes that the data have a multivariate normal distribution and are missing at random. Schafer (1997) applied the underlying principle to other multivariate distributions, and derived imputation algorithms for multivariate numerical, categorical and mixed data. Though theoretically sound, these methods rest on distributional assumptions that may sometimes be unrealistic in practice (e.g. assuming normality of a binary variable).

The S-PLUS `transcan` function (Alzola & Harrell 1999) represents a somewhat different approach to multivariate imputation. The function imputes each incomplete variable by cubic spline regression given all other variables, thus without assuming that the data can be modeled by a multivariate probability distribution. Though conceptually easy and flexible, the transcan algorithm lacks a sound theoretical rationale, so it is unknown whether the generated imputations are proper in the sense of Rubin (1987).

This paper describes a method that combines the most attractive aspects of both approaches. The method is called *Multivariate Imputation by Chained Equations* (MICE). It assumes that, for each incomplete variable, the user specifies a conditional distribution for the missing data given the other data. For example, logistic regression could be used for incomplete binary variables, polytomous regression for categorical data, and linear regression for numerical data. Under the assumption that a multivariate distribution exists from which these conditional distributions can be derived, MICE constructs a Gibbs sampler from the specified conditionals. This sampler is used to generate multiple imputations. A number of papers document the method (Van Buuren et al. 1999; Brand 1999).

The present paper describes the major functions in the S-PLUS library MDM, which stands for Missing Data Machine. The library is available on-line at our website http://www.multiple-imputation.com. We note that, for those that use SAS, a program called IVEWARE (Raghunathan et al. 1999) implements an approach that is related to ours. IVEWARE is geared toward imputing data with mixed measurement levels. Like our method, the IVEWARE algorithm iterates over the variables and requires only a specification of each conditional distribution of missing data. Many details in the implementation differ however.

2 Method

2.1 Multiple imputation

Multiple imputation is a statistical technique for analyzing incomplete data sets. The main idea is that each missing value is replaced by several (*m*) values, thus producing *m* imputed data sets. The differences between these data sets reflect the uncertainty of the missing values. Each imputed data set is analyzed by standard complete-data procedures, which ignore the distinction between real and imputed values. The *m* resulting analyses are then combined into one final analysis. Figure 1 illustrates the flow of operations in multiple imputation.

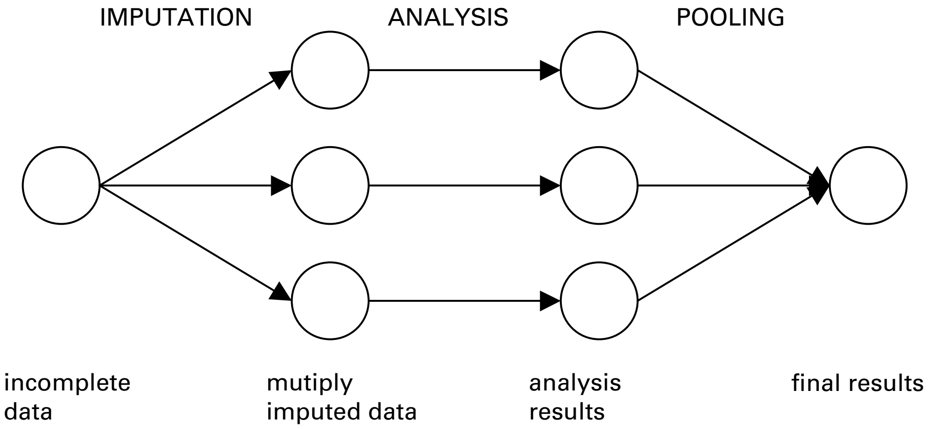

Figure : Schematic representation of the steps in multiple imputation. The process starts with an incomplete data set (on the left side), which is imputed m times (m=3 here) thus creating m completed data sets. Each complete data set is analyzed by using standard complete-data software, thus resulting in m analysis results. Finally, these m results are pooled into one final result that adequately reflects the amount of uncertainty in the estimates.

The primary advantage of multiple imputation is that it leads to valid statistical inferences in the presence of nonresponse. A second advantage is that only familiar complete-data software is needed to analyze the data.

Despite these virtues, the application of multiple imputation is not without problems. Though simple and sound procedures exist for pooling the *m* analyses, generating proper multiple imputations is not a trivial task. In practical applications, a major difficulty is the derivation of an appropriate predictive distribution from which imputations are to be drawn. Closed form analytic solutions are often unavailable, and some form of iterative algorithm is needed. The algorithm given in the next section requires only the specification of conditional distribution for the missing data in each incomplete variable.

2.2 MICE imputation algorithm

Let $X = (X_1, X_2,\ldots, X_k)$ be a set of k random variables, where each variable $X_j = (X_j^{\text{obs}}, X_j^{\text{mis}})$ may be partially observed, with $j = 1,\ldots,k$. The imputation problem is to draw from $P(X)$, the unconditional multivariate distribution of X. Let t denote an iteration counter. Assuming that data are missing at random (MAR), one may repeat the following sequence of Gibbs sampler iterations:

For X_1: draw imputations X_1^{t+1} from $P(X_1 \mid X_2^t, X_3^t,\ldots, X_k^t)$

For X_2: draw imputations X_2^{t+1} from $P(X_2 \mid X_1^{t+1}, X_3^t,\ldots, X_k^t)$

:

For X_k: draw imputations X_k^{t+1} from $P(X_k \mid X_1^{t+1}, X_2^{t+1},\ldots, X_{k\text{-}1}^{t})$,

i.e., condition each time on the most recently drawn values of all other variables. Properties of this general iteration scheme have been described by Gelfand and Smith (1990). Rubin and Schafer (1990) show that if $P(X)$ is multivariate normal, then iterating linear regression models like $X_1 = X_2^t\beta_{12} + X_3^t\beta_{13} + \ldots + X_k^t\beta_{1k} + \varepsilon_1$ with $\varepsilon_1 \sim N(0, \sigma_1^2)$ will produce a random draw from the desired distribution. Schafer (1997) generalizes this result to other multivariate distributions.

The implemented algorithm differs slightly from Schafer's approach in that the conditional models can be specified directly, thus without the need to choose an encompassing multivariate model for the entire data set. It is *assumed* that a multivariate distribution exists, and that draws from it can be generated by iteratively sampling from the conditional distributions. In this way, the multivariate problem is split into a series of univariate problems. Similar ideas have been applied by Kennickell (1991), Brand (1999) and Van Buuren *et al* (1993, 1999). The approach is also known as *regression switching* or *variable-by-variable* imputation.

It is not always certain whether the multivariate distribution actually exists. It is possible that the specification of two conditional distributions $P(X_1|X_2)$ and $P(X_2|X_1)$ are incompatible, so that no joint distribution $P(X_1, X_2)$ exists. Since there is no distribution to converge to, the algorithm will then alternate between isolated conditional distributions. In the linear case, this is probably more an exception than a rule. The subject of incompatible conditionals is, however, still an open research problem. Brand (1999) studied the performance of a variety of regression switching algorithms based on possibly incompatible conditionals, with quite encouraging results.

2.3 Specification of the imputation model

The most complex step in multiple imputation is the specification of the imputation model. As a general rule, using every available bit of available information yields multiple imputations that have minimal bias and maximal certainty. It is desirable that the algorithm produces imputations that preserve the structure in the data, as well as the uncertainty about this structure. Ideally, the complete-data model plays no role in imputation: a data set is multiply imputed only once, and will subsequently be used for any purpose. Such imputations are called *mindless* and a method that produces them is called a *mindless method* (Van Buuren et al 1993, 1994). In practice, achieving global mindless imputations (i.e. imputation suited for any purpose) is problematic since the imputation model may have excluded the relationship of interest. Following this line of reasoning, the number of predictors used for imputation should be chosen as large as possible. In addition, a large set of predictors tends to make the MAR assumption more plausible, thus reducing the need to make special adjustments for mechanisms that are not MAR.

In its extreme form, every variable will imputed from all other variables in the data using the most general model. In practice however, data sets often contain several hundreds of variables, all of which are potential predictors. It is not feasible (because of multicollinearity, computational and empty cell problems) to include all these variables. It is also not necessary. The increase in explained variance in linear regression is typically negligible after the best, say, 15 variables have been included. For imputation purposes, it is expedient to select a suitable subset of data that contains no more than 15 to 25 variables. Van Buuren et al (1999) provide the following strategy for selecting predictor variables from a large data base:

1. Include all variables that appear in the complete-data model. Failure to do so may bias the complete-data analysis, especially if the complete-data model contains strong predictive relations.

2. In addition, include the variables that appear in the response model. Factors that are known to have influenced the occurrence of missing data (stratification, reasons for nonresponse) are to be included on substantive grounds. Others variables of interest are those for which the distributions differ between the response and nonresponse groups.

3. In addition, include variables that explain a considerable amount of variance of the target variable. Such predictors help to reduce the uncertainty of the imputations.

4. Remove from the variables selected in steps 2 and 3 those variables that have too many missing values within the subgroup of incomplete cases.

Note that predictors may be incomplete themselves. In principle, one could apply the above modeling steps for each incomplete predictor in turn, but this may lead to a cascade of auxiliary imputation problems. In doing so, one runs the risk that every variable needs to be included after all. In practice, there is often a small set of key variables for which imputations are needed, which suggests that steps 1 through 4 are to be performed for key variables only. This was the approach taken in Van Buuren et al (1999), but it may miss important second level predictors in the data. A safer and more efficient, though more laborious, strategy is to perform the modelling steps for all incomplete level-1 predictors. This is done in Oudshoorn et al. (1999). We expect that it is rarely necessary to go beyond this level.

The `mice()` function of the MDM library features predictor selection. For each incomplete variable, the user can specify which predictors are used to generate the imputations. In addition, the statistical imputation model can be specified for each predictor.

2.4 Monitoring convergence

The Gibbs sampler is not a conventional algorithm in the sense that a particular criterion value is optimised. The Gibbs sampler aims for convergence in distribution, which is more difficult to assess than convergence in value. According to Gelman and Rubin (1992), the best method is to examine parallel sequences of the Gibbs sampler for a set of model parameters. At convergence, the sequences should overlap and be free of trend. Convergence is diagnosed when the variance between different sequences is no larger than the variance with each individual sequence.

The MICE algorithm creates *m* multiply imputed matrices in parallel, so it is possible to monitor the development of *m* separate strains of a given set of parameters. For example, one could monitor, for each incomplete variable, the mean and variance of the imputations. The choice of which parameters to monitor often depends on the scientific problem. See Gelman (1996) and Raftery and Lewis (1996) for discussions of this topic.

In our experience, MICE needs less iteration than is common in modern Markov Chain methodology, that often require thousands of iterations. For a given variable, the method creates statistically independent imputations. No iterations need to be wasted for achieving independence between successive draws, as is typical for Markov Chain methods. Brand's simulation study was done with just 5 iterations, with satisfactory performance (Brand 1999). For large amounts of missing data, more iterations will often be needed.

3 Implementation

3.1 General structure

The S-PLUS library MDM contains functions for imputing and analysing incomplete data by multiple imputation. The library defines three data classes, each of which corresponds to a particular step in Figure 1. Suppose that the incomplete data are in the form of a matrix or a data frame. Specific functions convert the input data into objects of the following three data classes:

- `mids`: multiply imputed data set (the result of imputation)
- `mira`: multiply imputed repeated analyses (the results of repeated complete data analyses)
- `mipo`: multiple imputed pooled results (the result of pooling the repeated analyses)

Table 1 contains a short description of the most important functions in the MDM library.

Table1: S-PLUS functions in the MDM library for generating, storing and analyzing multiply imputed data.

Function	Input	Output	Description
md.pattern	incomplete data	matrix	summarizes the pattern of the missing data
mice	incomplete data	mids	creates a multiply imputed data set
complete	mids	data.frame	converts mids into various forms of completed data
lm.mids	mids	mira	applies the linear regression model to the imputed data
glm.mids	mids	mira	applies the generalized linear model to the imputed data
gam.mids	mids	mira	applies the generalized additive model to the imputed data
nbrm.mids	mids	mira	applies the negative binomial model to the imputed data
analysis	mira	fit	extracts the j'th (1..m) complete data analysis
pool	mira	mipo	pools the analysis results

The analysis functions (`lm.mids`, `glm.mids`, `gam.mids`, `nbrm.mids`) are called as `lm()`, `glm()`, `gam()` and `nbrm()` through the standard S-PLUS dispatch mechanism. The other functions are called by the full name. The `mice()` function implements the algorithm of Section 2.2.

3.2 Elementary imputation methods

For each incomplete variable, one can specify an *elementary imputation method*. This is the method that the Gibbs sampling algorithm uses for imputing the variable, for example linear or logistic regression. Several elementary imputation methods are available. For numeric data these are: Bayesian linear regression imputation with normal errors, improper linear regression with normal errors, predictive mean matching, and unconditional mean imputation. Logistic regression imputation is used for binary data, and polytomous logistic regression for categorical data with more than two categories. Also, a simple random sample can be taken as imputations. This is useful if the data are supposed to be missing completely at random (MCAR).

In addition, users can write their own customized elementary imputation algorithms, and call these from within the Gibss sampler. This allows for specialized imputation methods for specific variables, e.g. imputation under particular editing constraints.

3.3 Passive imputation

There is often a need for transformed versions of the (imputed) data. In the case of incomplete data, one could 1) impute the original, and transform the completed original afterwards, or 2) transform the incomplete original and impute the transformed version. If, however, both the original and the transformed versions are needed within the imputation algorithm, neither of these approaches work because one cannot be sure that the transformation is synchronized between the original and transformed versions.

A special built-in elementary imputation method, called *passive imputation*, maintains the consistency among different functions of the same imputed data. Passive imputation synchronizes the transform with the most recently imputed original. The user can specify the transformation function. For example, the formula "~log(income)" searches for a column called "income", computes the logarithm of the values whenever income is imputed, and stores the result. This mechanism provides a convenient way to maintain synchronized dummy variables (e.g. specify "~color=="green", "~color=="red", and so on). In the current implementation, passive imputation is linked to only one original. It is not yet possible to define a passive variable that depends on two or more columns, for example, as the product of two variables.

3.4 Visiting scheme

The standard algorithm imputes each incomplete column in the data from left to right. It is known that the visiting scheme of the Gibbs sampler is essentially irrelevant to the results, but some schemes might be more efficient than others. It is possible to alter the default visiting scheme. If, for example, variables are ordered in time, it could be sensible to reflect the time order in the visiting sequence. The visiting scheme is also needed to keep passive variables synchronized with their imputed originals. In addition, some key variables could be visited (and imputed) more often than others.

4 Illustration

Table 2 is a small example data set created by Schafer to mimick the response pattern in NHANES III.

Table 2: Sample NHANES data with simulated patterns of missingness. Source: Schafer (1997, p. 237)

	age	bmi	hyp	chol
1	1			
2	2	22.7	1	187
3	1		1	187
4	3			
5	1	20.4	1	113
6	3			184
7	1	22.5	1	118
8	1	30.1	1	187
9	2	22.0	1	238
10	2			
11	1			
12	2			
13	3	21.7	1	206
14	2	28.7	2	204
15	1	29.6	1	
16	1			
17	3	27.2	2	284
18	2	26.3	2	199
19	1	35.3	1	218
20	3	25.5	2	
21	1			
22	1	33.2	1	229
23	1	27.5	1	131
24	3	24.9	1	
25	2	27.4	1	186

The function `md.pattern()` summarizes the missing data pattern in these data (1=observed, 0=missing). Rows and columns are sorted in according to the amount of missingness.

```
> md.pattern(nhanes)
   age hyp bmi chl
13   1   1   1   1  0
 1   1   1   0   1  1
 3   1   1   1   0  1
 1   1   0   0   1  2
 7   1   0   0   0  3
     0   8   9  10 27
```

The simplest way to create a multiply imputed data matrix is by calling the `mice()` function with its defaults set as

```
> imp_mice(nhanes)
```

The function returns an object of class `mids`. Imputations are generated according to the default method, which is predictive mean matching for numerical data (`bmi` and `chl`) and logistic regression for binary data (`hyp`). Individual imputations are found by listing specific parts of the `mids` object.

```
> imp$imputations$bmi

      1    2    3    4    5
 1 30.1 30.1 20.4 30.1 33.2
 3 30.1 29.6 30.1 29.6 30.1
```

...

```
> imp$imputations$hyp:

   1 2 3 4 5
 1 1 1 1 1 1
 4 2 2 1 2 2
```

...

The list element `pred.mat` is a square matrix containing 0/1 data, specifying the set of predictors to be used for each incomplete column. The predictor matrix can be found by the command

```
> imp$pred.mat

     age  bmi  hyp  chl
age    0    0    0    0
bmi    1    0    1    1
hyp    1    1    0    1
chl    1    1    1    0
```

Rows correspond to target variables, in the sequence as they appear in data. A value of '1' indicates that the column variable is used as a predictor for the target (row) variable. Thus, in the above example, `age`, `hyp` and `chl` are predictors for imputing `bmi`. The diagonal of `pred.mat` is zero. In its default setting, every column predicts all other columns. In the above example, `age` is complete, so it has no predictors by default.

The `complete()` function return various forms of completed data set. For example, the second imputed data set can be obtained as

```
> complete(imp,2)

   age  bmi hyp chl
 1   1 30.1   1 187
 2   2 22.7   1 187
 3   1 29.6   1 187
```

...

It is possible to fit linear regression on the imputed data as usual:

```
> fit_lm(chl~bmi+hyp+age,data=imp)

> summary(fit)

                est     se       t      df  Pr(>|t|) missing  fmi

(Intercept) -58.809 52.69 -1.116 632.00 0.26       NA    0.082

        bmi   7.273  1.73  4.204 464.80 0.00        9    0.097

        hyp -16.843 26.45 -0.636  12.68 0.53        8    0.617

        age  42.719 13.84  3.087  18.71 0.00        0    0.512
```

The `lm()` function recognizes the multiply imputed data, repeats the linear analysis *m* times, and returns a fit object of class `mira`. The next statement pools the repeated analyses into an object of class `mipo`, and extracts the table of coefficients of the linear model. The column termed ‘fmi’ contains the fraction of missing information about the estimate (Rubin 1987, p. 77). Similar analyses are possible for `glm()` and `gam()` functions.

5 Conclusion

The MICE algorithm is a conceptually simple, flexible and practical way to generate multivariate multiple imputations. For each incomplete variable the user can choose a set of predictors that will be used for imputation. This is useful for imputing large data sets. Passive imputation is a built-in feature that takes care that transformed data are always in sync with their original values. This can be used, for example, to impute categorical variables along with their dummies (needed for imputing other variables). In additional, the user can alter the visiting scheme of the Gibbs sampler, or plug-in customized imputation methods. Features like these make it easy to include complex imputation constraints in a practical but principled way.

A weakness of the approach is that convergence of the Gibbs sampler is guaranteed only in a number of special cases, e.g. under the multivariate normal model. Simulation studies (Brand, 1999) indicated that the method performs well in some other cases, but of course this is no guarantee that this will be true in general. In our experience, the method should be used carefully if the amount of missing information is large. More work is needed to develop reliable methods for checking convergence.

Another practical problem is that the specification of imputation models from large data bases containing hundreds of variables may involve a lot of work. It is not uncommon that multicollinearity and other instability problems show up if too many predictors are thoughtlessly added to the imputation model. It would be certainly be useful and efficient if predictor selection could be automated, where stability issues are taken into account. It could also be worthwhile to investigate more robust regression forms, like ridge regression (Schafer, 1997), or intermediate dimension reduction strategies (Belin, 1999).

Other practical enhancements include the use of constrained imputations, the allowance for survey weights, the possibility to specify interaction terms, and the development of additional elementary imputation methods, e.g. for Poisson regression. We expect that most features can be built into the existing software without too much trouble.

6 References

ALZOLA C, HARRELL F. An introduction of S-Plus and the Hmisc and Design libraries, 1999. (download from http://www.med.virginia.edu/medicine/clinical/hes)

BELIN TR. Strategies for handling nonresponse in highly multivariate surveys. Contributed Paper. International Conference on Survey Nonresponse, Oct 1999, Portland.

BRAND JPL. Development, implementation and evaluation of multiple imputation strategies for the statistical analysis of incomplete data sets. Thesis University of Rotterdam/TNO Prevention and Health, 1999.

BUUREN S van, RIJCKEVORSEL JLA van, RUBIN DB. Multiple imputation by splines. Bull Int Stat Inst, Contributed Papers II *1993;* 503–4.

BUUREN S van, MULLIGEN EM, BRAND JPL. Routine multiple imputation in statistical databases. In: French JC, Hinterberger H, eds. Proceedings of the Seventh International Working Conference on Scientific and Statistical Database Management. IEEE Computer Society Press, Los Alamitos, CA, 1994:74-78.

BUUREN S van, BOSHUIZEN HC, KNOOK DL. Multiple imputation of missing blood pressure covariates in survival analysis. Stat Med 1999;18, 681-94.

GELFAND AE, SMITH AFM. Sampling-based approaches to calculating marginal densities. J Am Stat Assoc 1990;85:398-409.

GELMAN A. Inference and monitoring convergence. In: Gilks WR, Richardson S, Spiegelhalter DJ, eds, Markov Chain Monte Carlo in practice. London: Chapman & Hall, 1996:131-43.

GELMAN A, RUBIN DB. Inference from iterative simulation using multiple sequences (with discussion). Stat Sci 1992;7:457-511.

KENNICKELL AB. Imputation of the 1989 survey of consumer finances: Stochastic relaxation and multiple imputation. Proceedings of the Section on Survey Research Methods, 1-10. Alexandria: ASA, 1991.

LI, K-H. Imputation using Markov chains. J Statist Comput Simul. 1988;30:57-79.

OUDSHOORN K, BUUREN S van, RIJCKEVORSEL JLA van. Flexible multiple imputation by chained equations of the AVO-95 survey. Contributed Paper. International Conference on Survey Nonresponse, Oct 1999, Portland.

RAGHUNATHAN TE, SOLENBERGER PW, HOEWYK J van. IVEWARE: Imputation and variance estimation software. Installation Instructions and User Guide. Survey Research Center, Institute of Social Research, 1999.

RAFERTY AE, LEWIS SM. Implementing MCMC. In:. Gilks WR, Richardson S, and Spiegelhalter DJ, eds, Markov Chain Monte Carlo in practice. London: Chapman & Hall, 1996:115-130.

RUBIN DB. Multiple imputation for nonresponse in surveys. New York: Wiley, 1987.

RUBIN DB. Multiple imputation after 18+ years (with discussion). J Am Stat Assoc 1996;*91*, 473-518.

RUBIN DB, SCHAFER JL. Efficiently creating multiple imputations for incomplete multivariate normal data, Proceedings of the Statistical Computing Section*,* 83-88. Alexandria: ASA, 1990:83-8.

SCHAFER JL. Analysis of incomplete multivariate data. London: Chapman & Hall, 1997.