Fully Conditional Specification in Multivariate Imputation

S. van Buuren[1,2]

J.P.L. Brand[3]

C.G.M. Groothuis-Oudshoorn[4]

D.B. Rubin[5]

Version: January 11, 2005

1 TNO Quality of Life, Leiden, The Netherlands (S.vanBuuren@pg.tno.nl)

2 University of Utrecht, The Netherlands

3 Pennington Biomedical Research Center, Baton Rouge, LA 70808 (BrandJP@pbrc.edu)

4 Roessingh Research and Development, Enschede, The Netherlands (k.groothuis@rrd.nl)

5 Harvard University, Cambridge MA 02138 (rubin@stat.harvard.edu)

Address for correspondence:

Prof Stef van Buuren

Dept. of Statistics, TNO Quality of Life

P.O. Box 2215

2301 CE LEIDEN

The Netherlands

Email: S.vanBuuren@pg.tno.nl

Abstract

The use of the Gibbs sampler with fully conditionally specified models, where the distribution of each variable given the other variables is the starting point, has become a popular method to create imputations in incomplete multivariate data. The theoretical weakness of this approach is that the specified conditional densities can be incompatible, and therefore the stationary distribution to which the Gibbs sampler attempts to converge may not exist. This study investigates practical consequences of this problem by means of simulation. Missing data are created under four different missing data mechanisms. Attention is given to the statistical behavior under compatible and incompatible models. The results indicate that multiple imputation produces essentially unbiased estimates with appropriate coverage in the simple cases investigated, even for the incompatible models. Of particular interest is that these results were produced using only five Gibbs iterations starting from a simple draw from observed marginal distributions. It thus appears that, despite the theoretical weaknesses, the actual performance of conditional model specification for multivariate imputation can be quite good, and therefore deserves further study.

Key words: multivariate missing data, multiple imputation, distributional compatibility, Gibbs sampling, simulation, proper imputation

1 Introduction

Missing data often plague the statistical analysis of multivariate data. When confronted with incomplete data, the analyst can choose a variety of strategies: ad-hoc methods (e.g., analysis of the complete cases only, available case methods, use of some indicator variables with means filled in), likelihood-based approaches that allow for missing data (e.g., EM algorithm, structural equations or mixed models), weighting, or imputation-based methods. The relative merits of these approaches have been discussed elsewhere (Little & Rubin, 2002; Schafer, 1997). Multiple imputation (Rubin 1987, 2004; 1996) is a general and statistically valid method for dealing with missing data. This paper studies a particular method for creating imputations (single or multiple) in multivariate data.

Let y denote an $n \times k$ matrix with data from n individuals on k variables. Let Y_j be the jth variable, and y_j the jth column of y (j=1,…,k). We define y_j^{obs} as the observed part of y_j , and y_j^{mis} as the missing part of in y_j. Let $y^{\text{obs}} = \left(y_1^{\text{obs}},...,y_k^{\text{obs}}\right)$ and $y^{\text{mis}} = \left(y_1^{\text{mis}},...,y_k^{\text{mis}}\right)$ stand for the collection of all observed and missing data, respectively. Imputation of y_j^{mis} will typically be based on the relation between the incomplete variable Y_j and the k–1 predictors $Y_{-j} = (Y_{1,...,} Y_{j-1,} Y_{j+1,...,} Y_k)$, where the nature of the relation is primarily estimated from the units contributing to y_j^{obs} . For notational convenience, we suppress notation for all variables that are fully observed, and so all distributions are implicitly conditional on the fully observed variables. Thus, each of the k columns in y has some missing values.

A number of practical problems can occur in general when $k > 1$:

- The predictors Y_{-j} themselves contain missing values;

- "Circular" dependence occurs, where y_j^{mis} depends on y_h^{mis}, and y_h^{mis} depends on y_j^{mis} ($h \neq j$), because in general Y_j and Y_h are correlated, even given other variables;
- Especially with large k and small n, collinearity or empty cells occur;
- Rows or columns can be ordered, e.g., as with longitudinal data;
- Variables are of different types (e.g., binary, unordered, ordered, continuous), thereby making the application of theoretically convenient models, such as the multivariate normal, theoretically inappropriate;
- The relation between Y_j and predictors Y_{-j} is complex, e.g., nonlinear, or subject to censoring processes;
- Imputation can create impossible combinations such as pregnant fathers.

This list is by no means exhaustive, and other complexities may appear for particular data. Two general strategies for imputing multivariate data have surfaced during the last decade: joint modeling and fully conditional specification (FCS).

The first common strategy, joint modeling, begins by specifying a parametric multivariate density $P(Y|\theta)$ for the data Y given the model parameters θ. Given this specification and appropriate prior distributions for θ, one can use the Bayesian framework to generate imputations as draws from the posterior predictive distribution $P\left(y^{mis} \mid y^{obs}\right)$, usually under the assumption of an ignorable missing data mechanism. Using this approach, Schafer (1997) described sophisticated methods for creating multivariate imputations under the multivariate normal, the log-linear, and the general location model. These methods are available as tools in S-Plus 6.2 and SAS V8.2.

The other common approach, FCS does not start from an explicit multivariate density, but instead implicitly defines $P(Y|\theta)$ by specifying a separate conditional density $P(Y_j|Y_{-j},\theta_j)$ for each

Y_j. This density is used to impute y_j^{mis} given y_{-j}, for example by linear or logistic regression applied to the cases in y_j^{obs}, where y_{-j} refers to the columns of y excluding y_j. Imputation under FCS is done by iterating over all conditionally specified imputation models, each iteration consisting of one cycle through all Y_j.

FCS has several important practical advantages over joint modeling. First, FCS allows for the creation of flexible multivariate models because it splits a k-dimensional problem into k one-dimensional problems. One can easily specify models that are outside any standard multivariate density $P(Y|q)$. Second, FCS may help to preserve investments in specialized imputation methods that are difficult to formulate as a part of a multivariate density $P(Y|q)$. For example, it is easy to incorporate imputation methods that preserve unique features in the data, e.g., bounds, skip patterns, interactions, bracketed responses, and so on. Also, it is relatively straightforward to maintain constraints between different variables. Such constraints might be needed to avoid logical inconsistencies in the imputed data. Gelman and Raghunathan (2001) observed that in such situations "separate regressions often make more sense than joint models". Third, generalization to models under nonignorable missing data mechanisms might be easier. Finally, the idea of specifying a separate imputation model for each variable is easy to communicate to users.

On the other hand, FCS is not without drawbacks. First, each conditional density has to be specified separately, so substantial modeling effort can be needed for data sets with many variables. Second, FCS is often computationally more intensive than joint modeling. Typical computational shortcuts (e.g. using the sweep operator, Little & Rubin, 2002) may not apply. Last, and very importantly, relatively little is known about the quality of the resulting imputations

because the implied joint distributions may not exist theoretically, and that convergence criteria are ambiguous.

Variations of the FCS idea have appeared before. Buck (1960) computed estimates for all missing entries by multiple regression, where the regression coefficients are computed using the complete cases, i.e. all individuals with fully complete data. The observed data for the individual constitute the independent variables in the equation predicting the missing variables for that individual. Gleason and Staelin (1975) extended Buck's method to include multivariate regression, and noted that their ad-hoc method could also be derived more formally from the multivariate normal distribution. These authors also brought up the possibility of an iterated version of Buck's method, suggesting that missing entries from one iteration could be used to form an improved estimate of the correlation matrix for use in a subsequent iteration. Variations of iterated regression imputation have been studied later by Finkbeiner (1979), Raymond and Roberts (1987), Jinn and Sedransk (1989), and Gold and Bentler (2000). Systems for creating multiple imputations by iterated regressions have been developed include FRITZ (Kennickell, 1991), IVEWARE (Raghunathan, Solenberger, van Hoewyk, 2000), HERMES missing data engine (Brand, 1999) and MICE (Van Buuren, Van Rijckevorsel & Rubin, 1993; Van Buuren & Oudshoorn, 2000). Royston (2004) created a version of MICE in Stata. Rubin (2003) pioneers a technique that attempts to take the best of both worlds by combining a FCS model for some missing values with joint modeling on other missing data. Other applications of FCS can be found in Kennickell (1999), Van Buuren, Boshuizen and Knook (1999), Raghunathan and Siscovick (1996), Oudshoorn, van Buuren and Van Rijckevorsel (1999), Heeringa, Little and Raghunathan (2002), Gelman and Raghunathan (2001) and Faris *et al*. (2002).

There appears to be yet no really satisfactory theory, but in many examples FCS seems to work well, is of great importance in practice, and is easily applied. Some simulation work is

available (Horton & Lipsitz, 2001; Raghunathan *et al*., 2001; Brand *et al.*, 2003), but this is relatively limited in scope and complexity. This paper presents a more extensive simulation-based evaluation of FCS.

2 Imputation by Fully Conditional Specification

2.1 Definitions and Introduction

Suppose $Y=(Y_1,Y_2,\ldots,Y_k)$ is a vector of k random variables with k-variate distribution $P(Y|q)$. We assume that the joint multivariate distribution of Y is completely specified by q, a vector of unknown parameters. For example, if Y is multivariate normally distributed, $q = (m,S)$, with m a k-dimensional mean vector and S a $k \times k$ covariance matrix. Let the matrix $y=(y_1,\ldots, y_n)$ with $y_i =(y_{i1} , y_{i2},\ldots,y_{ik})$, $i=1,\ldots,n$ be an i.i.d. sample of the vector Y. The matrix y is partially observed, in the sense that each column in y has missing data.

The standard procedure (cf. Rubin 1987) for creating multiple imputations y^* of y^{mis} is as follows:

1. Calculate the posterior distribution $p(q \mid y^{obs})$ of q based on the observed data y^{obs};
2. Draw a value q^* from $p(q \mid y^{obs})$;
3. Draw a value y^* from $p(y^{mis} \mid y^{obs}, q=q^*)$, the conditional posterior distribution of y^{mis} given $q=q^*$.

Repeat steps 2 and 3 for more imputations, e.g. 5-10. Appendix A gives algorithms for the cases where Y is univariate normally distributed, dichotomous, or polytomous.

For multivariate Y, the central problem is how to get the multivariate distribution of q, either explicitly or implicitly. FCS proposes to obtain a posterior distribution of q by sampling iteratively from conditional distributions of the form

$$\begin{aligned} &P(Y_1 | Y_{-1}, q_1), \\ &\dots \\ &P(Y_k | Y_{-k}, q_k). \end{aligned} \tag{1}$$

The parameters $q_1, \dots, q_k$ are treated as specific to the respective conditional densities and are not necessarily the product of some factorization of the "true" joint distribution $P(Y|q)$. More precisely, the tth iteration of the method consists of the following successive draws of the Gibbs sampler:

$$\begin{aligned} q_1^{*(t)} &\sim P(q_1 \mid y_1^{obs}, y_2^{(t-1)}, \dots, y_k^{(t-1)}) \\ y_1^{*(t)} &\sim P(y_1^{mis} \mid y_1^{obs}, y_2^{(t-1)}, \dots, y_k^{(t-1)}, q_1^{*(t)}) \\ &\dots \\ q_k^{*(t)} &\sim P(q_k \mid y_k^{obs}, y_1^{(t)}, y_2^{(t)}, \dots, y_{k-1}^{(t)}) \\ y_k^{(t)} &\sim P(y_k^{mis} \mid y_k^{obs}, y_1^{(t)}, \dots, y_{k-1}^{(t)}, q_k^{*(t)}) \end{aligned} \tag{2}$$

Observe that no information about y_j^{mis} is used to draw $q_j^{*(t)}$, which differs from Markov Chain Monte Carlo approaches to joint modeling. Our method is just a concatenation of univariate procedures applied to the cases with complete y_j, and deviates from MCMC theory at this point.

The iterations of (2) are executed m times in parallel to generate m multiple imputations. The number of iterations is fixed to a small number, say 5 or 10. This procedure implicitly assumes that the joint distribution is specified by (1), and that the Gibbs sampler in Equation (2) provides draws from it. With k incomplete variables, the vector parameters $q_1,\ldots,q_k$ will generally depend on each other, and so the sampler can be overparametrized. For example, the space spanned by $q_1,\ldots,q_k$ generally has more dimensions than appropriate. If this occurs, the implicit joint distribution does not exist. This issue is known as the problem of compatibility of conditionals (Arnold & Press, 1989).

2.2 Compatibility

Bhattacharryya (1943) observed that the combination of two conditional normal densities with linear regressions and constant variance defines a joint bivariate normal density. Two conditional densities are compatible if a joint distribution exists that has the given densities as its conditional density. In general, two conditional densities $f(x|y)$ and $g(y|x)$ are compatible if and only if (apart from a technical condition on the support of the densities) their density ratio $f(x|y)/g(y|x)$ factorizes into $u(x)v(y)$ for some integrable functions u and v (Besag, 1974). So, either the joint distribution exists and is unique, or does not exist. If $f(x|y)$ and $g(y|x)$ are known functions, we can calculate $f(x|y)/g(y|x)$ on a grid of x and y values, and infer compatibility if the matrix of $f(x|y)/g(y|x)$ values is of rank 1.

In actual data analysis, compatibility is not an all-or-nothing phenomenon. For example, even simple rounding errors can destroy exact compatibility. Measures have been proposed to measure the amount of compatibility (Arnold *et al*, 1999). The effects of near compatibility and

clear incompatibility on the quality of statistical inference are yet unknown, except in special cases. Section 6 therefore addresses the robustness of FCS under clearly incompatible models.

3 Evaluation of Univariate Imputation

The section provides simulation results for both compatible (univariate and multivariate) and incompatible FCS imputation methods. Gibbs sampling is not actually needed in univariate models, but forms an important ingredient for the multivariate case because of the use of univariate distributions in (1).

3.1 Setup: Data and simulations

We study the quality of univariate linear and logistic imputation methods using a data set of Irish wind speeds (Haslett & Raftery, 1989), which contains the average daily wind speed measured at 12 meteorological stations in Ireland during the years 1961-1978 (6574 time points). The correlations among these stations are high, ranging from 0.59 to 0.84, thus enabling the use of MAR (Rubin, 1976) missing data mechanisms that generate large differences between the complete and incomplete records. A random sample of n=400 was taken. No attempts were made here to model the temporal variation between the measures. To investigate the linear imputation method, the following five locations from the Irish wind speed data were selected: y_1 = Rosslare, x_1 = Roche's Pt, x_2 = Shannon, x_3 = Dublin, and x_4 = Clones. The original data of y_1 were replaced by new data that were generated according to the conditional probability of y_1 given $x_1,\ldots,x_4$ under a linear model, which was done to avoid any issues of inaccuracy of model fit. Missing

data in y_1 were subsequently created, where the response probability possibly depended on $x_1,\ldots,x_4$ using methods that are described below.

For the logistic method, we selected y_1 = Valentia, x_1 = Roche's Pt, x_2 = Rosslare, x_3 = Shannon, and x_4 = Dublin, dichotomized y_1 in equally sized groups, replaced the original data in y_1 by data generated according to a logistic regression model with linear predictors $x_1,\ldots,x_4$. Missing entries were then created in these substitutes.

We used data from Hosmer and Lemeshow (2000, p. 265) to study the polytomous model. This data set contains six responses from a survey of 412 women on knowledge, attitude and behavior towards mammography. The target variable y_1 was Mammographic Experience (ME) with three response categories (0=never, 1=during past year, 2=over year ago). As before, original values of y_1 were first replaced by data conforming to the polytomous logistic model conditional on the other data, and subsequently made missing.

Simulations were done using 1000 replications. In every replication, approximately 50% missingness in y_1 was generated under four different MAR missing data mechanisms: MCAR, MARRIGHT, MARTAIL and MARMID. Mechanism MCAR (missing completely at random) deletes observations in a completely random fashion, MARRIGHT creates more missingness for larger values, MARTAIL deletes more cases from both tails, whereas MARMID introduces more non-response in the center of the distribution. The latter three mechanisms introduce biases in the statistical analysis based on complete cases. Appendix B describes the methodology for generating the missing entries in detail. Figure 1 graphs the resulting distributions of the complete and incomplete target variable for one replication.

--- INSERT FIGURE 1 ABOUT HERE ---

3.2 Results

Let Q be the quantity of interest, and let $\hat{Q}$ be the associated complete-data estimate with variance U. For each replication i=1,…,1000, multiple imputation with m=5 is applied to the incomplete data. This results in the pooled estimates $\overline{Q}_m^{(i)}$ of $\hat{Q}$, $\overline{U}_m^{(i)}$ of U, and $B_m^{(i)}$ as the variance between the m-complete data estimates (Rubin, 1987, p. 76). Validity conditions for proper imputation similar to those presented by Rubin (1996) are: $\hat{\mathrm{E}}[\overline{Q}_m] \approx \hat{Q}$, $\hat{\mathrm{E}}[\overline{U}_m] \approx U$, and $\mathrm{var}(\overline{Q}_m) \approx (1+m^{-1})\,\hat{\mathrm{E}}[B_m]$, where $\hat{\mathrm{E}}[]$ is the mean and var[] the variance over the replications. Computations were done in SAS-IML.

--- INSERT TABLE 1 ABOUT HERE ---

Table 1 reports the following statistics based on the simulation:

- $\hat{Q}$, the complete data statistic based on the underlying values of all cases,
- ?, the fraction of information about Q missing due to nonresponse,
- $\hat{\mathrm{E}}[\hat{Q}_{ac}]$, the average $\hat{Q}$ computed from the available cases
- the coverage of the 95% c.i. of $\hat{Q}$ for Q for the available cases,
- $\hat{\mathrm{E}}[\overline{Q}_m]$, the average $\hat{Q}$ after multiple imputation,
- the coverage of the 95% c.i. under multiple imputation.

Choices for $\hat{Q}$ reflect aspects of the distribution of the incomplete variable (e.g., mean, probability of a category, quantiles) or quantify the relation with the predictors (e.g., correlations, conditional means).

The results for $\hat{E}[\hat{Q}_{ac}]$ indicate that available case analysis is often severely biased under MARRIGHT, MARTAIL and MARMID. Note that, unlike MARRIGHT and MARTAIL, mechanism MARMID generally increases the correlations with the predictors. Almost everywhere, the difference $\hat{E}[\overline{Q}_m] - \hat{Q}$ is much smaller than $\hat{E}[\hat{Q}_{ac}] - \hat{Q}$, so multiple imputation corrects for the biases introduced by MARRIGHT, MARTAIL and MARMID. For example, the bias of the median (P50) estimated by available case analysis under MARRIGHT is quite large (11.88-9.93=1.95), although it is negligible (11.88-11.84=0.04) after imputation. Note that under MARMID, the median bias of available cases (11.88-11.52=0.36) is larger than that of P25 (8.28-8.09=0.19) or P75 (15.32-15.29=0.03). This may seem surprising because estimates of the median are generally more stable estimates of the quartiles. Observe, however, that MARMID deletes more data from the center of the distribution, thus affecting the stability of the median.

The coverages under available case analysis are low. The worst case occurs when the mean of Y_1 is estimated from the available cases under the MARRIGHT mechanism. Here, the 95% confidence interval covered the true value only once (!) in 1000 simulations. In general, the confidence intervals of the available cases have acceptable coverage only under MCAR. In all other cases, confidence intervals are much too short and lead to incorrect statistical inferences.

In contrast, the actual coverage of the 95% confidence interval for multiple imputation is generally close to the nominal levels, and is nowhere below 93%. Coverage under MARMID is usually even larger than the nominal level. All in all, the results show that the linear, logistic and polytomous multiple imputation methods for univariate data are on target, well calibrated and

adhere to the validity conditions for proper imputation under a variety of missing data mechanisms.

4 Evaluation of FCS for Multivariate Imputation of Continuous Data

This section studies the performance of FCS for multivariate missing values with continuous data.

4.1 Setup: Data and simulations

Six locations from the Irish wind speed data are selected: y_1 = Roche's Pt, y_2 = Rosslare, y_3 = Shannon, y_4 = Dublin, x_1 = Clones, and x_2 = Malin Head. Two complete data sets, one simulated and one real, are created. The simulated data consist of approximately 400 cases drawn from the multivariate normal distribution with mean vector and covariance matrix equal to that estimated from the raw wind speed data. This simulated data set presents an idealized case where there are no issues of inaccuracy of model fit. The real data set is a random sample of about 400 observations from the raw wind speed data. Imputing this data yields insight into the robustness of the imputation model in more practical situations. Conditional on the observed data, non-monotone multivariate missing data were created in $Y_1,\ldots, Y_4$ using the method described in appendix B. The percentage of cases that were made incomplete was 62.5%.

The incomplete data were multiply imputed using m=10, a relatively high value chosen to account for larger fractions of missing information. The fully conditional specification consists of the set of linear regressions of each Y_j on all other variables, Y_{-j}. The number of Gibbs sampling

iterations is set to 5. This is a low value in a Gibbs sampling context, but we found it to work well in this type of application using starting values of y^{mis} drawn from each variable's observed marginal distribution. The execution time for generating, imputing and analyzing the 1000*10=10000 incomplete data sets was approximately 30 minutes on a Intel 1.7Ghz processor running SAS 6.2.

4.2 Results

--- INSERT TABLE 2 ABOUT HERE ---

Table 2 provides estimates of the bias and coverage of multiple imputation for many descriptive statistics estimating various quantities of the 6-variate distribution (Y_1, Y_2, Y_3, Y_4, X_1 , X_2). Under MAR, the available case analysis is often biased, but multiple imputation consistently moves in the right direction, and nearly always repairs the damage done by the missing data mechanism. For example, the correlation between Y_2 (Rosslare) and Y_3 (Shannon) drops from 0.59 to 0.22 for the available cases, but is 0.61 after imputation.

--- INSERT FIGURE 2 ABOUT HERE ---

Figure 2 illustrates some properties of the imputation process for Rosslare and Shannon in more detail. The figure on the left is a scatter plot of a sample of about 400 cases from the original Irish Wind Speed data. The middle figure portrays a subset of this data set. The slope of the regression line is biased downwards, and both variable ranges are limited to about half of the scale. The panel on the right is the first imputed data set after imputation. Note that imputation

‘restores’ the slope of the regression, the ranges of the variables, and the correlation between them.

Coverages are often close to the nominal value. The average coverage percentage over all statistics is 94.1 for the simulated data, and 93.5 for the real data, and thus both are remarkably close to the nominal value of 95. Coverage is occasionally lower than 90, especially for statistics with large fractions of missing information. We observed this also at other runs of the simulations, but at different places. The results provide firm evidence that for both simulated and real data, the Gibbs sampling imputation algorithm is on target and well calibrated under the studied conditions.

6 Evaluation of FCS for Multiple Imputation of Mixed Data

5.1 Setup: Data and simulations

Multivariate missing data were created in the Mammography Experience data set (Hosmer & Lemeshow, 2000, p. 265). As before, imputation of real and simulated data is studied. Non-monotone missing data under a MAR mechanism were created in four missing data patterns. Each of these patterns was characterized by missing data on one of the following pair of variables: (SYMPT,BSE), (ME,SYMPT), (BSE,DETC) and (SYMPT,DETC). Each pattern occurs in approximately 15.6% of all cases, so the total percentage of incomplete cases is 4 * 15.6 = 62.5%. Incomplete data in SYMPT and BSE are imputed by logistic regression, whereas ME and DETC are imputed by polytomous logistic regression. Imputation was always done conditional on the five other variables, with m = 10 and using 5 Gibbs sampling iterations from the marginal starting values.

5.2 Results

--- INSERT TABLE 3 ABOUT HERE ---

The results in Table 3 indicate that the performance of multiple imputation for multivariate data is quite satisfactory for both data sets. The point estimates are nearly always closer to the true values than under available case analysis, and the empirical coverage of the 95% intervals is close to the nominal value. The line labeled '*E*[PB|SYMPT=1]' illustrates that multiple imputation is clearly superior to available case even if the amount of missing information is small.

6 Evaluation of FCS for Imputation based on Incompatible Models

This section addresses potential consequences of incompatibility by means of simulation.

6.1 Setup: Data and Simulations

For each replication, 1000 draws were made from the bivariate normal distribution $P(Y_1, Y_2)$ with $m_1 = m_2 = 5$, $s_1^2 = s_2^2 = 1$, and $r_{12} = 0.6$. All values generated were positive. Missing data in Y_1 and Y_2 were in three ways:

MARRIGHT: logit(Pr(Y_1=missing)) = -1 + Y_2/5 and logit(Pr(Y_2=missing)) = -1 + Y_1/5;

MARTAIL: logit(Pr(Y_1=missing)) = -1 + 0.4|Y_2| and logit(Pr(Y_2=missing)) = -1 + 0.4|Y_1|;

MARMID: 1-Pr(MARTAIL).

--- INSERT FIGURE 3 ABOUT HERE ---

Figure 3 plots the probability to be missing under each mechanism as a function of the data. When taken together, these specifications led to zero, one or two missing observations in the pair (Y_1, Y_2). Under MARRIGHT, there were approximately 50% missing entries, 75% incomplete cases, and about 25% of the cases for which both Y_1 and Y_2 were missing. There were proportionally more missing data for the higher values Y_1 and Y_2, like in the univariate MARRIGHT mechanism in Table 1. The multivariate missing data were not entirely MAR because the cases where Y_1 or Y_2 (or both) is (are) missing were more frequent for the higher values. The regression lines are however not affected because the nonresponse is generated symmetrically around the regression lines.

Multiple imputations Y_1^* and Y_2^* using m=5 were created using five iterations of the Gibbs sampler. One iteration of the compatible Gibbs sampler consisted of a chain of two univariate imputation models $Y_2^* | Y_1 \sim N(m_1^* + b_1^* Y_1, s_1^{2*})$ and $Y_1^* | Y_2 \sim N(m_2^* + b_2^* Y_2, s_2^{2*})$, where m_1^*, b_1^*, s_1^{2*}, m_2*, b_2^*, and s_2^{2*} were draws from the appropriate posterior distributions. The first incompatible conditionally specified model was formulated by replacing the imputation step for Y_2 by $Y_2^* | Y_1 \sim N(m_1^* + b_1^* Y_1^2, s_1^{2*})$, so imputation was conditional on Y_1^2 instead of Y_1. A second incompatible model used log(Y_1) instead of Y_1. The linear model $Y_1 = a + b\, Y_2 + e$ was taken as the complete-data model, where scientific interest focused on b. The number of replications was

set to 500. The average fraction of missing information about b was approximately 0.63, so the imputation problem was quite difficult.

6.2 Results

--- INSERT TABLE 4 ABOUT HERE ---

Table 4 contains the results. The equality $E(b_1^*/s_1^{2*}) = E(b_2^*/s_2^{2*})$ must hold for the compatible normal model. Note that this equality is empirically obtained for the conditionally specified linear model. For both incompatible models, we find $E(b_1^*/s_1^{2*}) \neq E(b_2^*/s_2^{2*})$, thus indicating serious incompatibility. Column 5 lists the marginal mean of Y_1. Under MARRIGHT, CCA is biased since there are proportionally more missing data for higher Y_1, whereas the imputation methods are closer to the theoretical value, though not completely unbiased. Note that the imputation methods assume MAR, and therefore cannot completely account for all selectivity induced by the missing data mechanism. Column 6 is the mean of the regression weight b over all replications. All methods produce essentially the same regression weight, except for CCA under MARMID and MARTAIL. The standard errors indicate that all multiple imputation models, even the incompatible ones, are more efficient than CCA, which is due to the fact that they use the incomplete data in a more efficient way. The last column is the coverage coefficient, which is equal to the percentage of cases in which the 95% confidence interval includes the true value. Coverage is excellent in all cases, again except for CCA under MARMID and MARTAIL.

It appears that the forms of incompatibility as used here do not influence the statistical properties of multiple imputation in any major way. Somewhat surprisingly, we found that, in the

cases studied, applying a deliberately specified incompatible method gives less bias and more efficiency than CCA. Thus, imputation using the Gibbs sampler seems to be robust against incompatible specified conditionals in terms of bias and precision, thus suggesting that incompatibility may be a relatively minor problem in multivariate imputation.

7 Discussion

Fully conditional specification (FCS) is a convenient and powerful approach for creating imputations in multivariate missing data. Its theoretical weakness is that convergence can only be guaranteed under compatibility of conditionals, a condition that is often difficult to verify in practice. Our simulations show that this weakness does not seem to affect the quality of imputation in the cases considered. In fact, even for clearly incompatible models, the method produces reasonable multiple imputations with appropriate coverage. Thus, FCS appears to be robust against incompatibility. We suspect this result will hold more generally, but more work is needed to explore the boundaries of this conclusion.

The amount of work per iteration can be substantial, but only relatively few iterations appear to be needed when using well-chosen starting values. All simulations were done with just five iterations. Increasing the number of iterations did not result into a lower bias and a better coverage. This is unlike many Markov Chain Monte Carlo methods that often require thousands of iterations. Of course, we only studied simple models with no more than four incomplete variables, and therefore do not suggest that a small number of iterations would also be sufficient for larger or more complex models. We have also seen applications with large fractions of missing data and high correlations that required several hundreds of iterations before reaching

some stability. In our limited experience, using 20 iterations for modest missing data problems (<10-15% missing data) is ample. In more demanding problems, convergence of critical parameters should be carefully monitored, for example by the method of multiple sequences (Gelman and Rubin, 1992). A particular difficulty here is how to specify overdispersed starting values in the context of multivariate missing data, which is an area for further research. The costs of drawing a longer chain are only computational, so if the implementation is efficient, the benefits of faster convergence will be small.

The major advantage of FCS is increased flexibility in model building. It is easy to incorporate constraints on the imputed values, work with different transformations of the same variable, account for skip patterns, rounding, and so on. We concentrated on models containing main effects only. It is however straightforward to build imputation models that preserve higher order interactions between variables. Following imputation of the main effect, all interaction terms containing this effect can be immediately updated, thus preserving consistency across the data. It would be worthwhile to study how robust multiple imputation along these lines would be in preserving higher order interactions.

Throughout the paper, the imputation models assumed that MAR holds. It is relatively easy to adapt the method for models that are not MAR, but assumptions outside the data will be needed. Another extension is to impute vectors instead of scalars. This may be helpful if the relationships between the variables are difficult to model, or to speed up the method. Of course, we do not know whether our results will hold in such more complex cases, but the evidence obtained thus far suggests that FCS might also work well in such situations. Using FCS in concert with monotone missing data methods, as in Rubin (2003), appears to be particularly attractive because the potential for incompatibility is reduced relative to FCS. Of course, there is

always some point at which the technique breaks down, but conditional specification in multivariate imputation seems to be remarkably robust, and well worth investigating further.

ACKNOWLEDGEMENT

The authors thank the anonymous reviewer for providing extremely helpful comments.

Appendix A: Algorithms for univariate imputation

Depending on the distribution of y given x, concrete algorithms are as follows:

For normally distributed y with mean bx and variance s^2, $x=(x^{obs},x^{mis})$ is the $n \times p$ matrix of covariates and $n = n^{obs}+n^{mis}$ (Rubin, 1987, p. 167):

1. Estimate b by $b = (x^{obs\prime}x^{obs})^{-1}(x^{obs})' y^{obs}$.
2. A. Draw a random variable $g \sim c^2 (n^{obs} - p)$.

   B. Calculate $s*^2 = (y^{obs} - x^{obs} b)'(y^{obs} - x^{obs}b)/g$.
3. A. Draw $w_1 \sim N(0,I_p)$, i.e. p independent $N(0,1)$ variates where I_p is the identity matrix of order p.

   B. Calculate $b* = b + s* w_1 V^{1/2}$, where $V^{1/2}$ is the triangular square root of $V = (x^{obs\prime} x^{obs})^{-1}$ obtained by Cholesky factorization.
4. A. Draw $w_2 \sim N(0,I_n^{mis})$.

   B. Calculate $y* = x^{mis} b* + w_2 s*$.

Closely related algorithms that account for deviations from the normal distribution are:

Predictive mean matching: Replace step 4 by: 4. Calculate $y^{mis} = x^{mis} b*$. For each missing value $i = 1, \ldots, n^{mis}$ find the respondent whose $y^{obs} = x^{obs} b*$ is closest to y_i^{mis} and take y^{obs} of this case as the imputed value of i.

Hot-deck version: Replace in step 4 the draws of the n^{mis} normal deviates with draws of n^{mis} values with replacement from the set of n^{obs} observed standardized residuals $\{(y_i^{obs} - bx_i^{obs})(1-p/n^{obs})^{-1/2}/ s\}$ where $s = \Sigma_{obs} (y_i - x_i b)^2/(n^{obs}-p)$.

For dichotomous y we assume $P(y_i|x_i,b) = (\exp(x_i b)/(1+\exp(x_i b)))^{yi}\ (1-\exp(x_i b)/(1+\exp(x_i b))^{1-yi}$. Imputations of y^{mis} are obtained as follows*:*

1. Calculate by an iterative algorithm b, the MLE estimate of b and an estimate of the posterior variance of b (e.g. the Hessian matrix in $b=b$).
2. *A.* Draw $b^* \sim N(b,V(b))$.

   B. Calculate for $i=1,\ldots,n^{mis}$, $w_i=\exp(x_i b^*)/(1+\exp(x_i b^*)$.
3. Draw $u_i \sim \text{unif}(0,1)$, $i=1,\ldots, n^{mis}$. If $u_i > w_i$, impute $y_i=0$ otherwise impute $y_i=1$. For small samples, this procedure can be improve by SIR, as in Clogg *et al*. (1991).

A predictive mean matching version of this algorithm replaces step 3 by: Calculate $y^{mis} = x^{mis}\ b^*$. For each missing value $i = 1,\ldots,n^{mis}$ find the respondent whose $y^{obs} = x^{obs}\ b^*$ is closest to y_i^{mis} and take y^{obs} of this case as the imputed value of i.

For categorical y with unordered categories denoted by $0,\ldots,s\text{-}1$ suppose the distribution of y can be characterized as $\ln(P(y=j|x)/P(y=0|x)) = b_j x$, for $j=1,\ldots,s\text{-}1$, so the model for y is a series of separate logistic regression models of categories $1,\ldots,s\text{-}1$ against baseline category 0. An appropriate algorithm is for this model is:

1. Draw b^* from $N(b,V(b))$ where b is the MLE estimator of $b=(b_1,\ldots,b_{s-1})$ and $V(b)$ its estimated covariance matrix.
2. Calculate $p_{ij}^{mis} = \exp(-b_j * x_i^{mis})/(1+\sum_{v=1}^{s-1} \exp(-(b_v * x_i^{mis}))$ for $i=1,\ldots,n^{mis}$, $j=0,\ldots,s\text{-}1$ and $b_0=(0,\ldots,0)$.
3. Draw y_i^* from $\{0,\ldots,s\text{-}1\}$ with probabilities p_{ij}^{mis}, $i=1,\ldots,n^{mis}$, $j=0,\ldots,s\text{-}1$.

Appendix B: Generation of the missing data

This appendix describes the method developed by Brand (1999, pp. 110-113) for generating non-monotone multivariate missing entries in J variables $Y_1,\ldots,Y_J$ under MAR. We assume that $Y_1,\ldots,Y_J$ are initially completely known. Additional complete covariates $X_1,\ldots,X_L$ can be present for which no missing entries are sought. The method requires specification of the proportion of incomplete cases, the patterns of missing data that are allowed, the relative frequency of each pattern, and a specification of the way in which the observed information influence the response probability of each pattern.

More in particular, for a sample size n, let α ($0 < \alpha < 1$) denote the desired proportion of incomplete cases. Let there be P missing data patterns $R_1,\ldots,R_P$, chosen by the user, where $R_p = \{r_{p1},\ldots,r_{pJ}\}$ is a 0-1 response indicator vector of length J, with $r_{pj} = 0$ if variable Y_j is missing and $r_{pj} = 1$ otherwise. All response patterns except (0,0,…,0) or (1,1,…,1) may occur. Furthermore, let the vector $f = (f_1,\ldots,f_P)$ specify the relative frequencies of patterns $R_1,\ldots,R_P$, with $\Sigma_p f_p = 1$, also specified by the user.

Each case is randomly allocated to one of P candidate blocks with probability f_p. Within each candidate block, a subgroup of $\alpha n f_p$ cases is made incomplete according to pattern R_p using a probability model as follows. First calculate a linear score $s_i = \sum_j^J a_{pj} r_{pj} Y_{ij} + \sum_l^L b_{pl} X_{il}$ for each case in the block, where a_{pj} and b_{pl} are user weights specific to pattern p. A convenient choice for a_{pj} and b_{pl} is the set of regression weights from the linear regression of Y_j on $\{Y_{-j}, X_1,\ldots,X_L\}$ as computed from the initially complete data. Subsequently, divide the nf_p cases within the candidate block p into k_p subgroups using on their value s_i. The user can control the composition of each candidate subgroup by specifying k_p - 1 break points q_{pk} for k=1,…, k_p - 1 (in the form of

quantiles). In addition, specify for each subgroup h_k ($2 < k = k_p$) the odds w_{pk} of having response pattern R_p relative to the reference subgroup h_1. Together with α, these odds determine the probability on response pattern R_p for each case in the candidate block. For each case, a random draw from the uniform distribution is made. If this random draw does not exceed the probability on response pattern R_p, the data for that case are set to missing according to response pattern R_p. The procedure is repeated for every candidate block.

Choices for the subgroup size and the odds govern the properties of the incomplete data. For example, MARMID-like mechanisms have high missingness odds for cases with average s_i scores, while MARTAIL-like mechanisms are obtained by specifying higher missingness odds for extreme s_i scores. All of these are MAR, since the linear score s_i depends on the observed data only.

As an example, the simulations for Table 2 generated missing data in the multivariate data $\{Y_1,\ldots,Y_4,X_1,X_2\}$ using the following settings: P=4, R = {010111,001111,110011,101011}, $f_1 = f_2 = f_3 = f_4 = 0.25$, $a = 0.625$, $k_1 = k_2 = k_3 = k_4 = 2$, $q_{p1} = 0.5$ and $w_{pk} = 4$ with $p = 1,\ldots,$P and k=1,…, k_p-1.

REFERENCES

Arnold, B.C., Press, S.J. (1989). Compatible conditional distributions. *Journal of the American Statistical Association, 84,* 152-156.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussions). *Journal of the Royal Statistical Society, Series B*, 36:192-236.

Bhattacharryya, A. (1943). On some sets of sufficient conditions leading to the normal bivariate distribution. *Sankhya, 6,* 399-406.

Brand, J.P.L. (1999). *Development, implementation and evaluation of multiple imputation strategies for the statistical analysis of incomplete data sets*. Dissertation, Erasmus University Rotterdam.

Brand, J.P.L, van Buuren, S., Groothuis-Oudshoorn, C.G.M., Gelsema, E.S. (2003). A toolkit in SAS for the evaluation of multiple imputation methods. *Statistica Neerlandica*, *57*, 36-45.

Buck, S.F. (1960). A method of estimation of missing values in multivariate data suitable for use with an electronic computer. *Journal of the Royal Statistical Society, B, 22,* 302-306.

Clogg, C.C., Rubin, D.B., Schenker, N., Schultz, B. and Weidman, L. (1991). Multiple Imputation of Industry and Occupation Codes in Census Public-Use Samples Using Bayesian Logistic Regression. *Journal of the American Statistical Association*, 86, 413, 68-78.

Faris, P.D., Ghali, W.A., Brant, R., Norris, C.M., Galbraith, P.D., Knudtson, M.L. (2002). Multiple imputation versus data enhancement for dealing with missing data in observational health care outcome analyses. *Journal of Clinical Epidemiology, 55,* 184-191.

Finkbeiner, C. (1979). Estimation for the multiple factor model when data are missing. *Psychometrika, 44,* 409-420.

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

Gelman, A., Raghunathan, T.E. (2001). Discussion of Arnold *et al*. "Conditionally specified distributions". *Statistical Science*, *16*, 249-274.

Gleason, T.C., Staelin, R. (1975). A proposal for handling missing data. *Psychometrika*, *40*, 229-252.

Gold, M.S., Bentler, P.M. (2000). Treatments of missing data: A Monte Carlo comparison of RBHDI, Iterative Stochastic Regression Imputation, and Expectation-Maximization. *Structural Equation Modeling*, *7*, 319-355.

Hasslet, J., Raftery, A.E. (1989). Space-time modeling with long-memory dependence: Assessing Ireland's wind power resource. *Applied Statistics*, *38*, 1-50.

Heeringa, S.G., Little, R.J.A., Raghunathan, T.E. (2002). Multivariate imputation of coarsened survey data on household wealth. In R.M. Groves *et al*. (Eds.), *Survey Nonresponse* (pp. 357-371). New York: Wiley.

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

Hosmer, D.W., Lemeshow, S. (2000). *Applied logistic regression*. (Second edition). New York: John Wiley.

Jinn, J.-H., Sedransk, J. (1989). Effect on secondary data analysis of common imputation methods. *Sociological Methodology*, *19*, 213-241.

Kennickell, A.B. (1991). Imputation of the 1989 survey of consumer finances: Stochastic relaxation and multiple imputation, *ASA 1991 Proceedings of the Section on Survey Research Methods*, 1-10. Alexandria: ASA.

Kennickell, A. B. (1999), Multiple Imputation and Disclosure Control: The Case of the 1995 Survey of Consumer Finances. In *Record Linkage Techniques 1997*, Washington, DC: National Academy Press, 248-267.

Little, R.J.A., Rubin, D.B. (2002). *Statistical analysis with missing data. Second Edition*. New York: Wiley.

Oudshoorn, C.G.M., van Buuren, S., van Rijckevorsel, J.L.A. (1999). *Flexible multiple imputation by chained equations of the AVO-95 Survey*. Leiden: TNO Prevention and Health. Report PG/VGZ/99.045, 1999. http://www.multiple-imputation.com

Raghunathan, T.E., Siscovick, D.S. (1996). A multiple imputation analysis of a case-control study of the risk of primary cardiac arrest among pharmacologically treated hypertensives. *Applied Statistics, 45*, 335-352.

Raghunathan, T.E., Solenberger, P., van Hoewyk, J. (2000). *IVEware: Imputation and Variance Estimation Software: Installation Instructions and User Guide*. Survey Research Center, Institute of Social Research, University of Michigan. http://www.isr.umich.edu/src/smp/ive/

Raghunathan, T.E., Lepkowski, J.M., van Hoewyk, J., Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology, 27*, 85-95.

Raymond, M.R., Roberts, D.M. (1987). A comparison of methods for treating incomplete data in selection research. *Educational and Psychological Measurement, 47*, 13-26.

Royston, P. (2004). Multiple imputation of missing values. *The Stata Journal, 4*, 227-241.

Rubin, D.B. (1974). Characterizing the estimation of parameters in incomplete data problems, *Journal of the American Statistical Association, 69*, 467-474.

Rubin, D.B. (1976). Inference and missing data. *Biometrika, 63*, 581-592.

Rubin, D.B. (1987, 2004). *Multiple imputation for nonresponse in surveys*. New York: John Wiley. Reprinted as a Wiley Classic, 2004.

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

Rubin, D.B. (2003). Nested multiple imputation of NMES via partially incompatible MCMC. *Statistica Neerlandica, 57*, 3-18.

Schafer, J. (1997). *Analysis of incomplete multivariate data*. London: Chapman and Hall.

Van Buuren, S., van Rijckevorsel, J.L.A., Rubin, D.B. (1993). Multiple Imputation by splines. *Bulletin of the International Statistical Institute, Contributed Papers II*, 503-504.

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

Van Buuren, S., Oudshoorn C.G.M. (2000). *Multivariate imputation by chained equations: MICE V1.0 User's Manual*. Report PG/VGZ/00.038. Leiden: TNO Preventie en Gezondheid.

	Statistic	Pop	MCAR Fmi	MCAR AC (coverage)	MCAR MI (coverage)	MARRIGHT Fmi	MARRIGHT AC (coverage)	MARRIGHT MI (coverage)	MARTAIL Fmi	MARTAIL AC (coverage)	MARTAIL MI (coverage)	MARMID Fmi	MARMID AC (coverage)	MARMID MI (coverage)
LINEAR	E(y_1)	11.66	0.32	11.65 (95)	11.66 (95)	0.44	9.82 (00)	11.65 (95)	0.32	11.53 (91)	11.66 (95)	0.32	11.85 (96)	11.65 (96)
	P25(y_1)	8.16	0.35	8.16 (94)	8.15 (97)	0.23	6.77 (08)	8.13 (97)	0.36	8.50 (86)	8.15 (97)	0.36	7.60 (84)	8.13 (98)
	P50(y_1)	11.42	0.32	11.39 (95)	11.42 (97)	0.36	9.53 (01)	11.39 (97)	0.27	11.35 (92)	11.42 (97)	0.38	11.55 (98)	11.41 (98)
	P75(y_1)	14.90	0.30	14.89 (96)	14.92 (98)	0.48	12.53 (00)	14.92 (98)	0.31	14.35 (77)	14.91 (97)	0.31	15.89 (65)	14.92 (98)
	$r(y_1.x_1)$	0.72	0.43	0.72 (95)	0.72 (95)	0.54	0.65 (47)	0.72 (95)	0.51	0.65 (51)	0.72 (95)	0.35	0.79 (37)	0.72 (95)
	$r(y_1.x_2)$	0.59	0.39	0.59 (96)	0.59 (96)	0.48	0.50 (55)	0.58 (95)	0.42	0.50 (56)	0.59 (95)	0.36	0.67 (46)	0.59 (95)
	$r(y_1.x_3)$	0.66	0.41	0.66 (94)	0.66 (95)	0.52	0.59 (59)	0.66 (94)	0.47	0.58 (52)	0.66 (95)	0.36	0.74 (41)	0.66 (96)
	$r(y_1.x_4)$	0.61	0.40	0.61 (94)	0.60 (95)	0.48	0.52 (53)	0.60 (94)	0.43	0.51 (53)	0.60 (95)	0.37	0.70 (40)	0.61 (96)
LOGISTIC	P(y_1=0)	0.50	0.32	0.50 (95)	0.50 (94)	0.43	0.71 (00)	0.51 (95)	0.21	0.51 (91)	0.50 (95)	0.54	0.50 (99)	0.50 (95)
	P(y_1=1)	0.50	0.32	0.50 (95)	0.50 (94)	0.43	0.29 (00)	0.49 (95)	0.21	0.49 (91)	0.50 (95)	0.54	0.50 (99)	0.50 (95)
	E(x_1|y_1=0)	8.66	0.29	8.67 (96)	8.68 (96)	0.51	8.02 (27)	8.72 (97)	0.26	9.59 (19)	8.70 (96)	0.42	7.30 (02)	8.69 (94)
	E(x_1|y_1=1)	16.10	0.23	16.13 (95)	16.09 (95)	0.14	14.16 (20)	16.11 (95)	0.19	14.81 (21)	16.07 (96)	0.37	17.91 (04)	16.09 (95)
	E(x_2|y_1=0)	9.29	0.24	9.30 (95)	9.31 (96)	0.33	8.91 (75)	9.35 (97)	0.17	9.89 (73)	9.32 (95)	0.42	8.43 (38)	9.32 (95)
	E(x_2|y_1=1)	14.05	0.21	14.07 (96)	14.04 (94)	0.19	12.85 (53)	14.04 (95)	0.14	13.26 (60)	14.04 (95)	0.41	15.19 (28)	14.04 (95)
	E(x_3|y_1=0)	7.17	0.29	7.18 (96)	7.19 (95)	0.51	6.60 (25)	7.21 (97)	0.26	8.00 (16)	7.20 (96)	0.43	5.94 (02)	7.19 (95)
	E(x_3|y_1=1)	13.78	0.23	13.80 (96)	13.77 (96)	0.14	12.09 (20)	13.78 (95)	0.19	12.67 (21)	13.75 (96)	0.36	15.35 (04)	13.76 (95)
	E(x_4|y_1=0)	7.18	0.26	7.19 (95)	7.20 (97)	0.41	6.71 (60)	7.21 (96)	0.21	7.80 (58)	7.19 (96)	0.43	6.25 (23)	7.19 (95)
	E(x_4|y_1=1)	12.45	0.20	12.47 (93)	12.44 (94)	0.18	10.99 (39)	12.43 (96)	0.15	11.56 (51)	12.44 (96)	0.36	13.74 (23)	12.43 (95)
POLYTOME	P(y_1=0)	0.57	0.52	0.57 (96)	0.55 (96)	0.61	0.68 (08)	0.56 (95)	0.54	0.56 (95)	0.55 (95)	0.64	0.58 (96)	0.56 (95)
	P(y_1=1)	0.25	0.54	0.25 (96)	0.26 (97)	0.70	0.17 (13)	0.26 (96)	0.58	0.25 (95)	0.26 (97)	0.65	0.25 (95)	0.26 (95)
	P(y_1=2)	0.18	0.55	0.18 (95)	0.19 (97)	0.69	0.15 (75)	0.18 (96)	0.58	0.19 (95)	0.19 (96)	0.64	0.17 (93)	0.18 (96)
	E(x_1|y_1=0)	8.06	0.29	8.05 (98)	8.03 (99)	0.38	8.56 (09)	8.05 (99)	0.36	8.12 (97)	8.02 (100)	0.29	7.96 (98)	8.05 (100)
	E(x_1|y_1=1)	6.71	0.52	6.70 (95)	6.79 (97)	0.48	7.53 (30)	6.78 (98)	0.54	7.16 (56)	6.82 (98)	0.68	6.02 (11)	6.72 (95)
	E(x_1|y_1=2)	7.19	0.49	7.19 (96)	7.22 (97)	0.50	8.00 (41)	7.23 (96)	0.54	7.53 (82)	7.25 (96)	0.56	6.63 (58)	7.17 (96)

Table 1: Properties of multiple imputation in a univariate y_1. Given are the population value (Pop), the fraction of missing information (Fmi), the mean estimate under available case analysis and its 95% c.i. coverage (AC), the mean estimate for MI and its 95% c.i. coverage (MI) under four MAR missing data mechanisms: MCAR, MARRIGHT, MARTAIL and MARMID. Based on m=5 imputations. P25(y), P50(y) and P75(y) represent the first, second and third quartile of y respectively. Notation P(.) is the marginal probability of observing the argument, E(.|.) stands for the conditional expectation. Notation $r(y,x)$ is used for the Pearson correlation between y and x.

Statistic	Pop	Simulated data			Raw data		
		Fmi	AC (coverage)	MI (coverage)	Fmi	AC (coverage)	MI (coverage)
E(y_1)	12.36	0.15	11.35 (13)	12.36 (96)	0.15	11.36 (14)	12.37 (95)
P25(y_1)	8.15	0.11	7.37 (42)	8.18 (98)	0.11	7.39 (46)	8.19 (97)
P50(y_1)	11.72	0.17	10.53 (16)	11.78 (97)	0.17	10.54 (17)	11.79 (96)
P75(y_1)	15.88	0.22	14.47 (29)	15.91 (99)	0.22	14.45 (27)	15.93 (98)
E(y_2)	11.65	0.22	10.93 (31)	11.66 (97)	0.22	10.94 (31)	11.67 (95)
P25(y_2)	7.97	0.15	7.36 (45)	7.93 (97)	0.16	7.36 (43)	7.93 (97)
P50(y_2)	10.91	0.21	10.07 (38)	11.03 (98)	0.21	10.09 (40)	11.04 (97)
P75(y_2)	14.64	0.26	13.70 (54)	14.81 (98)	0.26	13.70 (54)	14.83 (97)
E(y_3)	10.45	0.12	9.53 (11)	10.44 (95)	0.13	9.54 (11)	10.45 (95)
P25(y_3)	6.76	0.11	6.07 (41)	6.79 (96)	0.10	6.09 (43)	6.81 (96)
P50(y_3)	9.94	0.14	8.88 (16)	9.96 (98)	0.14	8.89 (16)	9.98 (97)
P75(y_3)	13.52	0.21	12.21 (24)	13.54 (98)	0.21	12.21 (24)	13.54 (98)
E(y_4)	9.78	0.12	8.86 (10)	9.79 (95)	0.12	8.88 (12)	9.80 (95)
P25(y_4)	6.01	0.09	5.35 (43)	6.06 (97)	0.09	5.35 (45)	6.06 (98)
P50(y_4)	9.16	0.14	8.12 (15)	9.24 (98)	0.14	8.13 (15)	9.24 (96)
P75(y_4)	12.88	0.20	11.52 (25)	12.92 (97)	0.21	11.53 (24)	12.94 (98)
$r(y_1,y_2)$	0.73	0.44	0.69 (73)	0.73 (89)	0.43	0.69 (74)	0.74 (86)
$r(y_1,y_3)$	0.83	0.40	0.81 (79)	0.85 (88)	0.40	0.81 (79)	0.85 (86)
$r(y_1,y_4)$	0.74	0.53	0.38 (00)	0.77 (77)	0.53	0.38 (00)	0.77 (79)
$r(y_1,x_1)$	0.75	0.24	0.74 (90)	0.75 (92)	0.24	0.74 (91)	0.75 (93)
$r(y_1,x_2)$	0.62	0.20	0.61 (93)	0.63 (93)	0.20	0.61 (92)	0.63 (92)
$r(y_2,y_3)$	0.59	0.50	0.22 (00)	0.61 (91)	0.50	0.22 (00)	0.61 (90)
$r(y_2,y_4)$	0.66	0.39	0.62 (79)	0.68 (93)	0.39	0.61 (79)	0.68 (92)
$r(y_2,x_1)$	0.61	0.28	0.59 (88)	0.60 (94)	0.29	0.58 (87)	0.60 (93)
$r(y_2,x_2)$	0.48	0.26	0.46 (89)	0.47 (94)	0.25	0.46 (89)	0.47 (91)
$r(y_3,y_4)$	0.79	0.39	0.74 (61)	0.78 (92)	0.40	0.74 (60)	0.78 (93)
$r(y_3,x_1)$	0.82	0.25	0.82 (92)	0.83 (91)	0.25	0.81 (93)	0.82 (94)
$r(y_3,x_2)$	0.67	0.19	0.66 (93)	0.67 (94)	0.19	0.66 (90)	0.67 (93)
$r(y_4,x_1)$	0.84	0.24	0.84 (94)	0.85 (90)	0.24	0.84 (94)	0.85 (91)
$r(y_4,x_2)$	0.77	0.21	0.76 (92)	0.77 (91)	0.21	0.76 (90)	0.77 (90)

Table 2: Results for multiple imputation of multivariate missing data in $y_1,\ldots,y_4$ by means of iterated linear regressions (Gibbs sampling) for two data sets (simulated and raw) constructed from the Irish windspeed data. Given are the population value (Pop), the mean estimate under available case analysis and the coverage of its 95% c.i. (AC), the mean estimate under multiple imputation MI and its 95% c.i. (MI). Based on m=10 imputations. P25(y), P50(y) and P75(y) represent the first, second and third quartile of y respectively. Notation $r(y,x)$ stands for the Pearson correlation between y and x. Variables x_1 and x_2 are complete.

Statistic	Pop	Simulated data (n =412)			Raw data (n=412)		
		Fmi	AC (coverage)	MI (coverage)	Fmi	AC (coverage)	MI (coverage)
P(ME=0)	0.57	0.51	0.59 (87)	0.58 (95)	0.43	0.61 (65)	0.57 (95)
P(ME=1)	0.25	0.63	0.23 (83)	0.23 (95)	0.55	0.22 (75)	0.25 (98)
P(ME=2)	0.19	0.44	0.19 (96)	0.19 (96)	0.48	0.17 (88)	0.18 (95)
E[PB|SYMPT=0]	8.24	0.22	8.68 (61)	8.29 (95)	0.26	8.51 (76)	8.19 (96)
E[PB|SYMPT=1]	7.29	0.09	7.66 (22)	7.28 (96)	0.09	7.71 (20)	7.32 (95)
OR(SYMPT,HIST)	0.37	0.33	0.23 (93)	0.23 (93)	0.29	0.25 (94)	0.27 (95)
OR(SYMPT,BSE)	0.51	0.41	0.28 (93)	0.59 (96)	0.42	0.33 (78)	0.72 (96)
OR(HIST,BSE)	0.50	0.36	0.57 (96)	0.51 (96)	0.38	0.68 (97)	0.60 (98)

Table 3: Bias and coverage of multiple imputation (m=10) after Gibbs sampling applied to multivariate categorical data, compared with available case analysis. Mammography Experience Study (n=412), where SYMPT was recoded into two categories. Notation P(.) is the marginal probability of observing the argument, E(.|.) stands for the conditional expectation. OR(x,y) is the odds ratio of x and y.

Mechanism	Method	Compatibility statistics		Estimates			Fmi	Cov
		$E(b_1/s_1^2)$	$E(b_2/s_2^2)$	$E(Y_1)$	$E(b)$	$E(se(b))$		
	Theoretical values			5.00	0.600			95
MARRIGHT	Complete case analysis			4.84	0.597	0.051		93
	MI compatible linear	0.94	0.94	4.91	0.595	0.046	0.63	95
	MI incompatible quadratic	0.09	0.92	4.91	0.589	0.046	0.63	95
	MI incompatible log	4.12	0.90	4.91	0.582	0.047	0.64	95
MARMID	Complete case analysis			5.00	0.678	0.060		79
	MI compatible linear	0.94	0.96	5.00	0.613	0.057	0.75	94
	MI incompatible quadratic	0.09	0.92	5.00	0.601	0.058	0.75	94
	MI incompatible log	4.13	0.90	5.00	0.579	0.058	0.75	94
MARTAIL	Complete case analysis			5.00	0.556	0.040		78
	MI compatible linear	0.95	0.95	5.00	0.596	0.038	0.50	94
	MI incompatible quadratic	0.09	0.93	5.00	0.590	0.038	0.50	94
	MI incompatible log	4.35	0.93	5.00	0.590	0.037	0.50	95

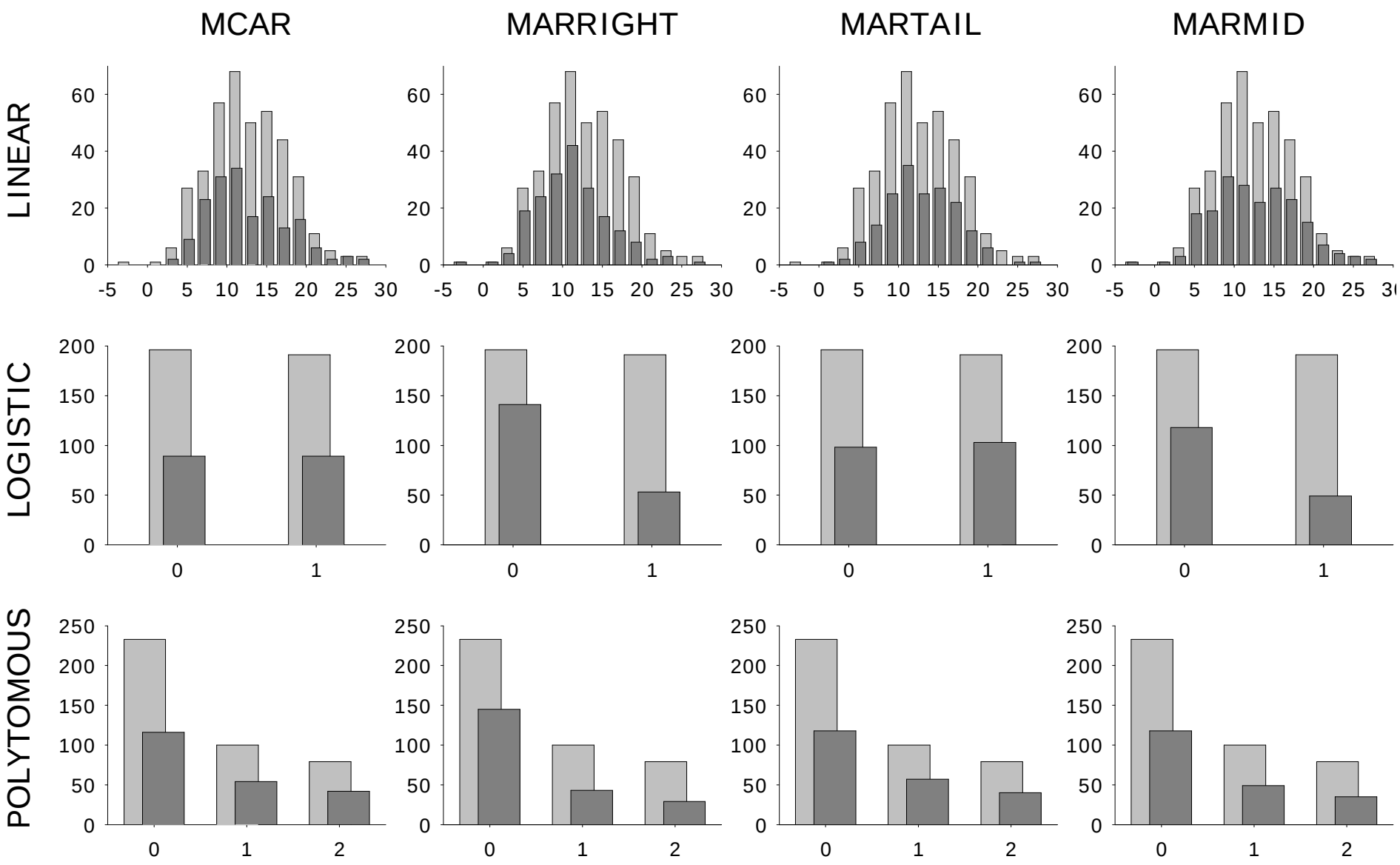
Table 4: Regression slopes, standard errors and coverages (95% c.i.) under one compatible and two incompatible multiple imputation models (bivariate normal data, r=0.6, n=1000, m=5, three symmetric missing data mechanisms, 500 replications) compared with complete case analysis. Fmi = fraction of missing information, Cov = 95% c.i. coverage.

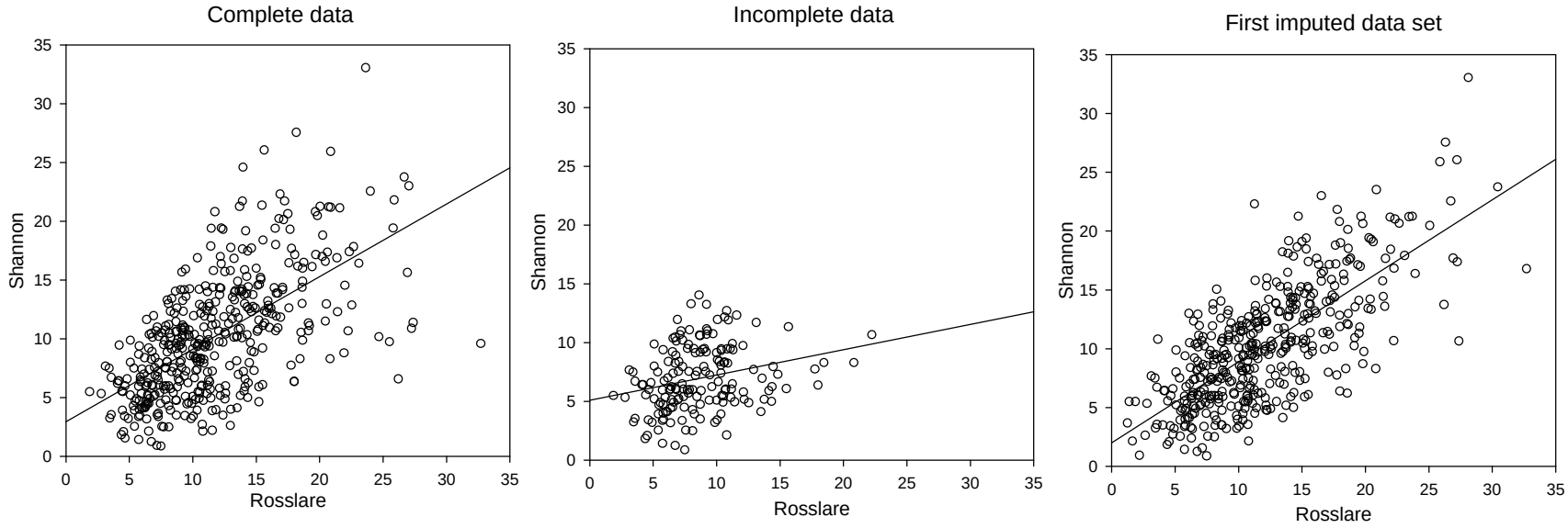
Figure captions

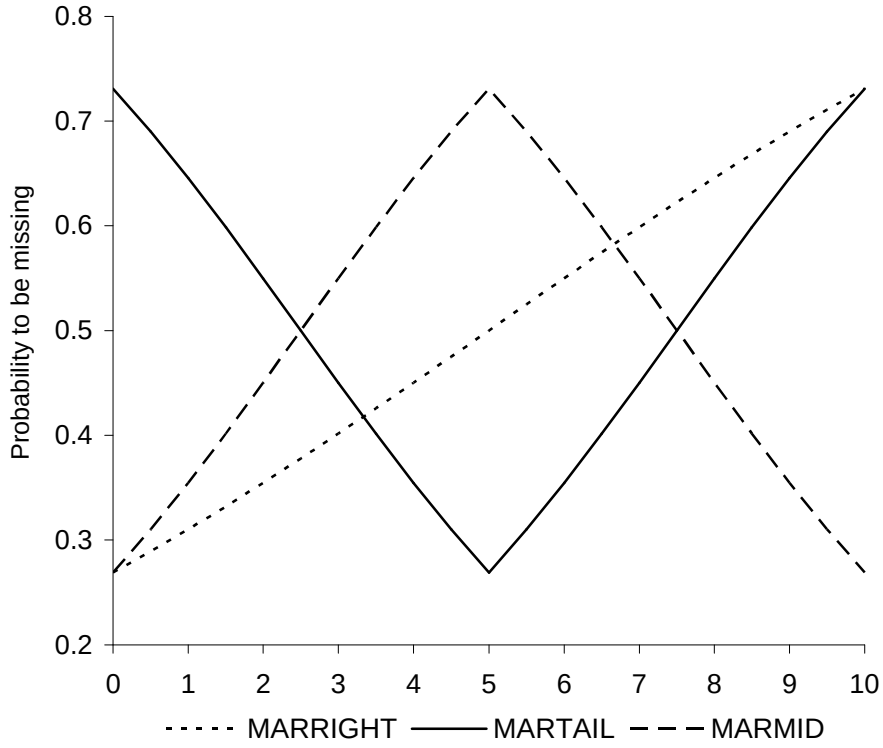
Figure 1: Distribution of the target variable before (light) and after (dark) missing data under four missing data mechanisms. MARRIGHT deletes more from the right tail, MARTAIL from both tails, and MARMID from the middle values. Data are the Irish Wind Speed sample (n=400, Haslett & Raftery, 1989) and Mammographic Experience data (Hosmer and Lemeshow, 2000, n=412).

Figure 2: Scatter plots of the locations Rosslare and Shannon from the Irish Wind Speed data (Haslett & Raftery, 1989). The incomplete data (middle panel) contains a subset of the complete data (left hand panel). Data are missing at random (MAR). The right hand panel illustrates the scatter of cases of the first multiply imputed data set.

Figure 3: Probability to be missing in the bivariate simulation as a function of the data value (i.e. Y_1 or Y_2) under three missing missing data mechanisms.


MCAR
MARRIGHT
MARTAIL
MARMID
LINEAR
LOGISTIC
POLYTOMOUS


Complete data
Incomplete data
First imputed data set
Shannon
Rosslare
0
5
10
15
20
25
30
35


0.8
0.7
0.6
0.5
0.4
0.3
0.2
Probability to be missing
0
1
2
3
4
5
6
7
8
9
10
MARRIGHT
MARTAIL
MARMID