Garret M. Fitzmaurice, Michael G. Kenward, Geert Molenberghs, Anastasios A. Tsiatis, Geert Verbeke

Handbook of Missing Data

Contents

Symbol Description

α	To solve the generator maintenance scheduling, in the past, several mathematical techniques have been applied.
σ^2	These include integer programming, integer linear programming, dynamic programming, branch and bound etc.
$\sum$	Several heuristic search algorithms have also been developed. In recent years expert systems,
abc	fuzzy approaches, simulated annealing and genetic algorithms have also been tested.
$\theta\sqrt{abc}$	This paper presents a survey of the literature
ζ	over the past fifteen years in the generator
∂	maintenance scheduling. The objective is to
sdf	present a clear picture of the available recent literature
ewq	of the problem, the constraints and the other aspects of
bvcn	the generator maintenance schedule.

Part I

Multiple Imputation

1

Fully conditional specification

Stef van Buuren

1. *Netherlands Organisation for Applied Scientific Research TNO, Leiden*
2. *Department of Methodology & Statistics, Faculty of Social and Behavioural Sciences, University of Utrecht*

CONTENTS

1.1 Introduction

1.1.1 Overview

The term *fully conditional specification* (FCS) refers to a class of imputation models for non-monotone multivariate missing data. Other names for this class of models include *sequential regression multivariate imputation* and *chained equations*. As non-monotone missing data frequently occur in practice, FCS covers a wide range of missing data problems.

The material presented here builds upon the foundations of multiple imputation as laid out in Chapter 13, and draws heavily on Chapters 4 and 5 of my book *Flexible Imputation of Missing Data* (van Buuren, 2012). Additional background, computer code to apply FCS in practice, and an overview of current software can be found on www.multiple-imputation.com.

The present chapter focusses on fully conditional specification, an approach for multivariate multiple imputation that has become very popular with practitioners thanks to its ease of use and flexibility. Section 1.2 outlines a number of practical problems that appear when trying to impute multivariate missing data. Section 1.3 distinguishes various multivariate missing data patterns, and introduces four linkage measures that aid in setting up multivariate imputation models. Section 1.4 briefly review three general strategies to impute multivariate missing data. Section 1.5 describes the FCS approach, its assumptions, its history, the Multivariate Imputation by Chained Equations (MICE) algorithm, and discusses issues surrounding compatibility and performance. Section 1.6 provides a systematic account of seven choices that needs to be made when applying FCS in practice. Section 1.7 highlights the role of diagnostics in imputation.

1.1.2 Notation

Let Y denote the $N \times p$ matrix containing the data values on p variables for all N units in the sample. The response indicator R is an $N \times p$ binary matrix. Elements in R are denoted by r_{ij} with $i = 1, \ldots, N$ and $j = 1, \ldots, p$. Element $r_{ij} = 1$ if the corresponding data value in Y is observed, and $r_{ij} = 0$ if it is missing. We assume that we know where the missing data are, so R is always completely observed. The observed data in Y are collectively denoted by Y^o. The missing data are collectively denoted as Y^m, and contain all Y-values that we do not see because they are missing. Notation Y_j denotes the j-th column in Y, and Y_{-j} indicates the complement of Y_j, that is, all columns in Y except Y_j. When taken together $Y = (Y^o, Y^m)$ contains the hypothetically complete data values. However, the values of the part Y^m are unknown to us, and the data are thus incomplete. Notation Y_j^o and Y_j^m stand for the observed and missing data in Y_j, respectively. Symbol $\dot{Y}_j$ stands for imputations of Y_j^m.

1.2 Practical problems in multivariate imputation

There are various practical problems that may occur when one tries to impute multivariate missing data. Many imputation models for Y_j use the remaining columns Y_{-j} as predictors. The rationale is that conditioning on Y_{-j} preserves the relations among the Y_j in the imputed data. This section considers some potential difficulties in setting up imputation models of this type.

Suppose that we want to impute variable Y_j given other predictors in the data Y_{-j}. An obvious difficulty is that any of the predictors Y_{-j} may also contain missing data. In that case, it is not possible to calculate a linear predictor for cases that have missing data, and consequently such cases cannot be imputed.

A second difficulty is that circular dependence can occur, where Y_j^m depends on Y_h^m and Y_h^m depends on Y_j^m with $h \neq j$. In general, Y_j and Y_h are correlated, even given other variables. The limiting case occurs if Y_h^m is a function of Y_j^m for example, a transformation. When ignored, such circularities may lead to inconsistencies in the imputed data, or to solutions with absurd values.

Third, variables may have different measurement levels, e.g., binary, unordered categorical, ordered categorical, continuous, or censored data. Properly accounting for such features of the data is not possible using the application of theoretically convenient models, such as the multivariate normal, potentially leading to impossible values, e.g. negative counts. Distributions can take many forms. If the scientific interest focusses on extreme quantiles of the distribution, the imputation model should fairly represent the shape of the entire distribution.

Many datasets consist of hundreds, or even thousands, of variables. This creates problems in setting up imputation models. If the number of incomplete variables is large, problems with collinearity, unstable regression weights and empty cells occur. The general advice is to condition on as many variables as possible, but this may easily lead to imputation models that have more parameters than data points. A good selection of predictors and a judicious choice of constraints will often substantially improve imputations.

The ordering of rows and columns can be meaningful, e.g., as in longitudinal or spatial data. Data closer in space or time are typically more useful as predictors. With monotone missing data, imputation needs to progress from the most complete to least complete variable, so one may wish to regulate the sequence in which variables are imputed. Also, modeling could be done efficiently if it respects the known ordering in the data.

The relation between Y_j and predictors Y_{-j} can be complex, e.g., nonlinear, subject to censoring processes, or functionally dependent. For example, if the complete-data model requires a linear and a quadratic term, then both terms should be present in the imputation model. Also, the contribution of a given predictor may depend on the value of another one. If such relations exist in the observed data, it makes sense to preserve these in the imputed data. Taking care of complex relations is not automatic and requires careful treatment on behalf of the imputer.

Imputation of multivariate missing data can create impossible combinations, such as pregnant grand-fathers, or ”quality of life” of the deceased. We would generally like to avoid combinations of data values that can never occur in reality (i.e., if the values we would have been observed), but achieving this requires a careful analysis and setup of the imputation model.

In practice, it can also happen that the total of a set of variables is known (e.g. a budget total), but that some of the components are missing. If two components are missing, then imputation of one implies imputation of the other, since both values should add up to a known total. A more complex problem surfaces when two or more components are missing.

Finally, there are often different causes for the missing data in the same dataset. For example, the missing data could result from a failure to submit the questionnaire, from non-contact, from the fact that the respondent skipped the question, and so on. Depending on the subject matter, each of these multiple causes could require its own imputation model.

Other complexities may appear in real life. Properly addressing such issues is not only challenging, but also vital to creating high quality and believable imputations.

1.3 Missing data patterns

1.3.1 Overview

It is useful to study the missing data pattern for several reasons. For monotone missing data, we have convergence in one step, so there is no need to iterate. Also, the missing data pattern informs us which variables can contain information for imputation, and hence plays an important role in the setup of the imputation model.

Figure 1.1 illustrates four missing data patterns. The simplest type is the monotone pattern, which can result from drop-out in longitudinal studies. If a pattern is monotone, the variables can be sorted conveniently according to the percentage of missing data. Imputation can then proceed variable by variable from left to right with one pass through the data (Little and Rubin, 2002).

The patterns displayed in Figure 1.1 are connected since it is possible to travel to all dark cells by horizontal or vertical moves, just like the moves of the rook in chess. Connected patterns are needed to estimate parameters. For example, in order to be able to estimate a correlation coefficient between two variables, they need to be connected, either directly by a set of cases that

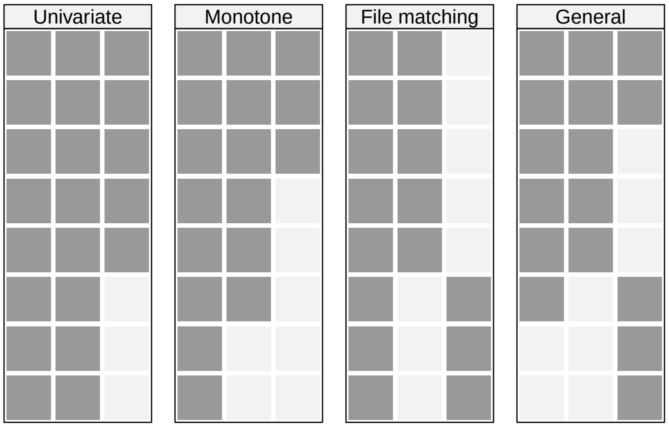

FIGURE 1.1
Some missing data patterns in multivariate data. Dark is observed, light is missing.

have scores on both, or indirectly through their relation with a third set of connected data. Unconnected patterns may arise in particular data collection designs, like data combination of different variables and samples, or potential outcomes.

More intricate missing data patterns can occur for data organised in the 'long' format, where different visits of the same subject form different rows in the data. Van Buuren (2011) contains examples for hierarchical data.

1.3.2 Ways to quantify the linkage pattern

The missing data pattern influences the amount of information that can be transferred between variables. Imputation can be more precise if other variables are present for those cases that are to be imputed. By contrast, predictors are potentially more powerful if they are present in rows that are very incomplete in other columns. This section introduces four measures of linkage of the missing data pattern. Note that degree of missingness is only one factor to consider, so the material presented is very much a partial guide to decisions faced by the imputer.

The *proportion of usable cases* (van Buuren et al., 1999) for imputing variable Y_j from variable Y_k is defined as

$$I_{jk} = \frac{\sum_i^n (1 - r_{ij}) r_{ik}}{\sum_i^n (1 - r_{ij})}. \tag{1.1}$$

This quantity can be interpreted as the number of pairs (Y_j, Y_k) with Y_j

missing and Y_k observed, divided by the total number of missing cases in Y_j. The proportion of usable cases I_{jk} equals 1 if variable Y_k is observed in all records where Y_j is missing. The statistic can be used to quickly select potential predictors Y_k for imputing Y_j based on the missing data pattern. High values of I_{jk} are preferred.

Reversely, we can also measure how well observed values in variable Y_j connect to missing data in other variables as

$$O_{jk} = \frac{\sum_i^n r_{ij}(1 - r_{ik})}{\sum_i^n r_{ij}}. \tag{1.2}$$

This quantity is the number of observed pairs (Y_j, Y_k) with Y_j observed and Y_k missing, divided by the total number of observed cases in Y_j. The quantity O_{jk} equals 1 if variable Y_j is observed in all records where Y_k is missing. The statistic can be used to evaluate whether Y_j is a potential predictors for imputing Y_k.

The statistics in equations 1.1 and 1.2 are specific for the variable pair (Y_j, Y_k). We can define overall measures of how variable Y_j connects to all other variables Y_{-j} by aggregating over the variable pairs.

The *influx coefficient* I_j is defined as

$$I_j = \frac{\sum_k^p \sum_i^n (1 - r_{ij}) r_{ik}}{\sum_k^p \sum_i^n r_{ik}} \tag{1.3}$$

The coefficient is equal to the number of variable pairs (Y_j, Y_k) with Y_j missing and Y_k observed, divided by the total number of observed data cells. The value of I_j depends on the proportion of missing data of the variable. Influx of a completely observed variable is equal to 0, whereas for completely missing variables we have $I_j = 1$. For two variables with the same proportion of missing data, the variable with higher influx is better connected to the observed data, and might thus be easier to impute.

The *outflux coefficient* O_j is defined in an analogous way as

$$O_j = \frac{\sum_k^p \sum_i^n r_{ij}(1 - r_{ik})}{\sum_k^p \sum_i^n (1 - r_{ik})} \tag{1.4}$$

The quantity O_j is the number of variable pairs with Y_j observed and Y_k missing, divided by the total number of incomplete data cells. Outflux is an indicator of the potential usefulness of Y_j for imputing other variables. Outflux depends on the proportion of missing data of the variable. Outflux of a completely observed variable is equal to 1, whereas outflux of a completely missing variable is equal to 0. For two variables having the same proportion of missing data, the variable with higher outflux is better connected to the missing data, and thus potentially more useful for imputing other variables. Note the word 'potentially', since the actual usefulness will also depend on the amount of association between the variables.

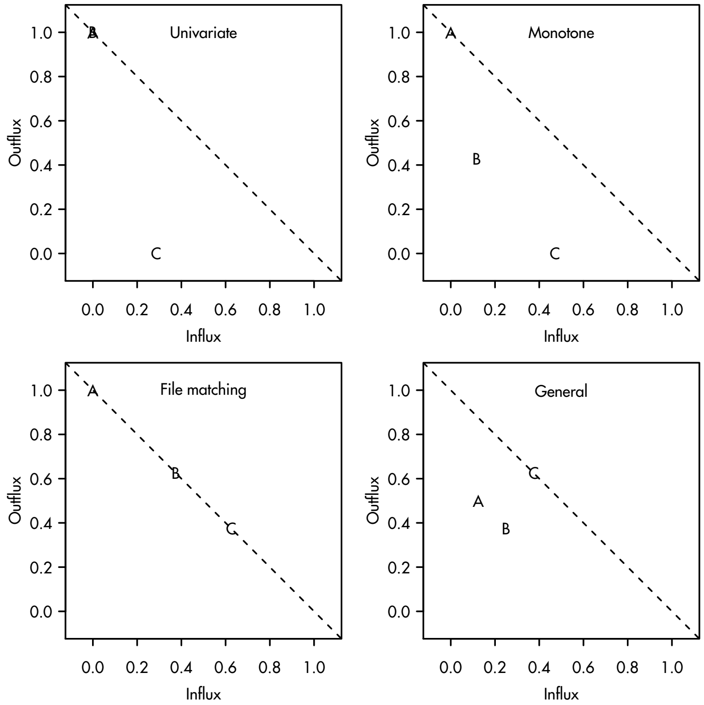

FIGURE 1.2
Fluxplot: Outflux versus influx in the four missing data patterns from Figure 1.1.

The influx of a variable quantifies how well its missing data connect to the observed data on other variables. The outflux of a variable quantifies how well its observed data connect to the missing data on other variables. Higher influx and outflux values are preferred. Figure 1.2 plots outflux against influx. In general, variables that are located higher up in the display are more complete and thus potentially more useful for imputation. In practice, variables closer to the subdiagonal are better connected than those further away. The fluxplot can be used to spot variables that clutter the imputation model. Variables that are located in the lower regions (especially near the left-lower corner) *and* that are uninteresting for later analysis are better removed from the data prior to imputation.

Influx and outflux are summaries of the missing data pattern intended to aid in the construction of imputation models. Keeping everything else constant, variables with high influx and outflux are preferred. Realize that outflux indicates the potential (and not actual) contribution to impute other variables. A variable with high O_j may turn out to be useless for imputation if it is fully unrelated to the incomplete variables, e.g., an administrative person identifier.

On the other hand, the usefulness of a highly predictive variable is severely limited by a low O_j.

1.4 Multivariate imputation models

1.4.1 Overview

Rubin (1987, pp. 160–166) distinguished three tasks for creating imputations: the modeling task, the imputation task, and the estimation task. The modeling task is to provide a specification for the joint distribution $P(Y) = P(Y^o, Y^m)$ of the hypothetically complete data. The issues that arise with incomplete data are essentially the same as for complete data, but in imputation the emphasis is on getting accurate predictions of the missing values. The imputation task is to specify the posterior predictive distribution $P(Y^m|Y^o)$ of the missing values given the observed data and given the assumed model $P(Y)$. The estimation task consists of calculating the posterior distribution of the parameters of this distribution given the observed data, so that random draws can be made from it.

In Rubin's framework, the imputations follow from the specification of the joint distribution $P(Y)$. Van Buuren (2007) distinguished three strategies to specify the model used to impute multivariate missing data.

- *Monotone data imputation.* Given a monotone missing data pattern, imputations are created by drawing for a sequence of univariate conditional distributions $P(Y_j|Y_1, \ldots, Y_{j-1})$ for $j = 1, \ldots, p$;

- *Fully conditional specification (FCS).* For general patterns, the user specifies a conditional distribution $P(Y_j|Y_{-j})$ directly for each variable Y_j, and assumes this distribution to be the same for the observed and missing Y_j (ignorability assumption). Imputations are created by iterative drawing from these conditional distributions. The multivariate model $P(Y)$ is implicitly specified by the given sets of conditional models;

- *Joint modeling.* For general patterns, imputations are drawn from a multivariate model $P(Y)$ fitted to the data, usually per missing data pattern, from the derived conditional distributions.

Chapter 13 reviews methods for multiple imputation for monotone missing data, whereas Chapter 15 covers joint modeling in great detail. The present chapter concentrates on FCS. The remainder of this section briefly reviews monotone data imputation and joint modeling.

1.4.2 Imputation of monotone missing data

If the missing data pattern in Y is monotone, then the variables can be ordered as $Y_1, \dots, Y_j, \dots, Y_p$ according to their missingness. The joint distribution $P(Y) = (Y^o, Y^m)$ decomposes as (Rubin, 1987, pp. 174)

$$P(Y|\phi) = P(Y_1|\phi_1)P(Y_2|Y_1, \phi_2)\dots\ P(Y_p|Y_1, \dots, Y_{p-1}, \phi_p) \qquad (1.5)$$

where $\phi_1, \dots, \phi_j, \dots, \phi_p$ are the parameters of the model to describes the distribution of Y. The ϕ_j parameters only serve to create imputations, and are generally not of any scientific interest or relevance. The decomposition requires that the missing data pattern is monotone. In addition, there is a second, more technical requirement: the parameters of the imputation models should be *distinct* (Rubin, 1987, pp. 174–178).

Monotone data imputation is fast and provides great modeling flexibility. Depending on the data, Y_1 can be imputed by a logistic model, Y_2 by a linear model, Y_3 by a proportional odds model, and so on. In practice, a dataset may be near-monotone, and may become monotone if a small fraction of the missing data were imputed (Li, 1988; Rubin and Schafer, 1990; Schafer, 1997; Rubin, 2003b). See Chapter 13 for more detail.

1.4.3 Imputation by joint modeling

Joint modeling starts from the assumption that the hypothetically complete data can be described by a multivariate distribution. Assuming ignorability, imputations are created as draws under the assumed model. Joint modeling describes the data Y by the multivariate distribution $P(Y|\theta)$, where θ is a vector of unknown parameters of the distribution. The model for $P(Y|\theta)$ can be any multivariate distribution, but the multivariate normal distribution, with $\theta = (\mu, \Sigma)$ for the mean μ and covariance Σ, is a convenient and popular choice.

Within the joint modeling framework, the parameters of scientific interest are functions of θ (Schafer, 1997, Ch. 4). Observe that the θ parameters are conceptually different from the ϕ_j parameters used in Section 1.4.2. The θ parameters derive from the multivariate model specification, whereas the ϕ_j parameters are just unknown parameters of the imputation model, and have no scientific relevance.

When the assumptions hold, joint modeling is elegant and efficient. For example, under multivariate normality, the *sweep operator* and *reverse sweep operator* are highly efficient computational tools for converting outcomes into predictors and vice versa. See Little and Rubin (2002, pp. 148–156) and Schafer (1997, p. 157–163) for details.

The major limitation of joint modeling is that the specified multivariate model may not be a good fit to the data. For example, if the data are skewed or if dependencies occur in the data, it could prove be difficult to find an appropriate multivariate model. Schafer (1997, p. 211–218) reported simulations that showed that imputations drawn under the multivariate normal model are generally robust to non-normal data. Joint modeling by a multivariate distribution can often be made more realistic through data transformations, or through the use of specific rounding techniques. Nevertheless, in many practical situations where imputations are desired (c.f. section 1.2), there will be no reasonable multivariate model for the data.

1.5 Fully conditional specification (FCS)

1.5.1 Overview

In contrast to joint modeling, FCS specifies the multivariate distribution $P(Y|\theta)$ through a set of conditional densities $P(Y_j|Y_{-j}, \phi_j)$, where ϕ_j are unknown parameters of the imputation model. As in section 1.4.2, the ϕ_j parameters are not of scientific interest, and only serve to model conditional relations used for imputation. The key assumption is that the conditional densities for Y_j^o and Y_j^m are the same (ignorability assumption). This conditional density is used to impute Y_j^m given the other information. Starting from simple random draws from the marginal distribution, imputations $\dot{Y}_1$ are drawn for Y_1^m given the information in the other columns. Then, Y_2 is imputed given the currently imputed data, and so on until all variables are imputed with one pass through the data. Then, Y_1^m is re-imputed during the second iteration using the imputation draw in iteration one, and so on. In practice, the iteration process can already be stopped after five or ten passes through the data. FCS is a generalization of univariate imputation for monotone data, and borrows the idea of Gibbs sampling from the joint modeling framework.

FCS bypasses task 1 of the procedure of section 1.4.1, the specification of the joint distribution $P(Y|\theta)$. Instead, the user specifies the conditional distribution $P(Y_j|Y_{-j})$ directly for each variable Y_j. Imputations are created by iterative draws from these conditional distributions. The multivariate model $P(Y|\theta)$ is only implicitly specified by the specified set of conditional models.

The idea of conditionally specified models is quite old. Conditional probability distributions follow naturally from the theory of stochastic Markov chains (Bartlett, 1978, pp. 231–236). For spatial data analysis, Besag preferred the use of conditional probability models over joint probability models, since “the conditional probability approach has greater intuitive appeal to the practising statistician” (Besag, 1974, p. 223). Buck (1960) proposed a procedure for calculating estimates for all missing entries by multiple regression. For example, to impute missing data in the first variable, Y_1 was regressed on $Y_2, \ldots, Y_p$, where the regression coefficients are computed using the complete cases. Buck’s method does not iterate and requires a large sized sample of complete cases. 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. Due to a lack of computational resources at that time, they were unable to evaluate the idea, but later work along these lines has been put forward by Finkbeiner (1979), Raymond and Roberts (1987), Jinn and Sedransk (1989), van Buuren and van Rijckevorsel (1992) and Gold and Bentler (2000).

Multiple imputation is different from this literature because it draws imputations from a distribution instead of calculating optimal predictive values. Ideas similar to FCS have surfaced under a variety of names: stochastic relaxation (Kennickell, 1991), variable-by-variable imputation (Brand, 1999), switching regressions (van Buuren et al., 1999), sequential regressions (Raghunathan et al., 2001), ordered pseudo-Gibbs sampler (Heckerman et al., 2001), partially incompatible MCMC (Rubin, 2003a), iterated univariate imputation (Gelman, 2004) and chained equations (van Buuren and Groothuis-Oudshoorn, 2000).

The main reasons for using FCS is increased flexibility and ease of use. Little (2013) explains the advantages of conditional modeling as follows:

> When modeling, it can be useful to factor a multivariate distribution into sequence of conditional distributions. Univariate regression is easier to understand, and a sequence of univariate conditional regressions is more easily elaborated, for example, by including interactions, polynomials, or splines, or modeling heteroscedasticity.

1.5.2 Chained equations: The MICE algorithm

There are several ways to implement imputation under conditionally specified models. Algorithm 1.1 describes one particular instance: the MICE algorithm (van Buuren and Groothuis-Oudshoorn, 2000, 2011) which divides the multivariate data in columns. The algorithm starts with a random draw from the observed data, and imputes the incomplete data in a variable-by-variable fashion. One iteration consists of one cycle through all Y_j. The number of iterations can often be low, say 5 or 10. The MICE algorithm generates multiple imputations by executing Algorithm 1.1 in parallel m times.

1.5.3 Properties

The MICE algorithm is a Markov chain Monte Carlo (MCMC) method, where the state space is the collection of all imputed values. If the conditionals are compatible, the MICE algorithm is a Gibbs sampler, a Bayesian simulation technique that samples from the conditional distributions in order to obtain

Algorithm 1.1: MICE algorithm for imputation of multivariate missing data.

```
1. Specify an imputation model $P(Y_j^m|Y_j^o, Y_{-j})$ for variable $Y_j$ with $j = 1, \dots, p$.
2. For each $j$, fill in starting imputations $\dot{Y}_j^0$ by random draws from $Y_j^o$.
3. Repeat for $t = 1, \dots, T$:
4. Repeat for $j = 1, \dots, p$:
5. Define $\dot{Y}_{-j}^t = (\dot{Y}_1^t, \dots, \dot{Y}_{j-1}^t, \dot{Y}_{j+1}^{t-1}, \dots, \dot{Y}_p^{t-1})$ as the currently complete data except $Y_j$.
6. Draw $\dot{\phi}_j^t \sim P(\phi_j^t|Y_j^o, \dot{Y}_{-j}^t)$.
7. Draw imputations $\dot{Y}_j^t \sim P(Y_j^m|Y_j^o, \dot{Y}_{-j}^t, \dot{\phi}_j^t)$.
8. End repeat $j$.
9. End repeat $t$.
```

samples from the joint distribution (Gelfand and Smith, 1990; Casella and George, 1992). In conventional applications of the Gibbs sampler the full conditional distributions are derived from the joint probability distribution (Gilks, 1996). In MICE however, the conditional distributions are directly specified by the user, and so the joint distribution is only implicitly known, and may not even exist. While the latter is clearly undesirable from a theoretical point of view (since we do not know the joint distribution to which the algorithm converges), in practice it does not seem to hinder useful applications of the method (cf. Section 1.5.4).

In order to converge to a stationary distribution, a Markov chain needs to satisfy three important properties (Roberts, 1996; Tierney, 1996):

- *irreducible*, the chain must be able to reach all interesting parts of the state space;
- *aperiodic*, the chain should not oscillate between states;
- *recurrence*, all interesting parts can be reached infinitely often, at least from almost all starting points.

With the MICE algorithm, irreducibility is generally not a problem since the user has large control over the interesting parts of the state space. This flexibility is actually the main rationale for FCS instead of a joint model.

Periodicity is a potential problem, and can arise in the situation where imputation models are clearly inconsistent. A rather artificial example of oscillatory behavior occurs when Y_1 is imputed by $Y_2\beta+\epsilon_1$ and Y_2 is imputed by $-Y_1\beta+\epsilon_2$ for some constant β. The sampler will oscillate between two qualitatively different states, so this is a periodic procedure. The problem with periodic sampler is that the result will depend on the stopping point. In general, we would like the statistical inferences to be independent of the stopping point. A way to diagnose the *ping-pong* problem is to stop the chain at different points. The stopping point should not affect the statistical inferences. The addition of noise to create imputations is a safeguard against periodicity, and allows the sampler to "break out" more easily.

Non-recurrence may also be a potential difficulty, manifesting itself as explosive or non-stationary behavior. For example, if imputations are made through deterministic functions, the Markov chain may lock up. Such cases can sometimes be diagnosed from the trace lines of the sampler. See van Buuren and Groothuis-Oudshoorn (2011) for examples and remedies. As long as the parameters of imputation models are estimated from the data, non-recurrence is likely to be mild or absent.

1.5.4 Compatibility

Gibbs sampling is based on the idea that knowledge of the conditional distributions is sufficient to determine a joint distribution, if it exists. Two conditional densities $P(Y_1|Y_2)$ and $P(Y_2|Y_1)$ are said to be *compatible* if a joint distribution $P(Y_1, Y_2)$ exists that has $P(Y_1|Y_2)$ and $P(Y_2|Y_1)$ as its conditional densities. More precisely, the two conditional densities are compatible if and only if their density ratio $P(Y_1|Y_2)/P(Y_2|Y_1)$ factorizes into the product $u(Y_1)v(Y_2)$ for some integrable functions u and v (Besag, 1974). So, the joint distribution either exists and is unique, or does not exist.

The MICE algorithm is ignorant of the non-existence of the joint distribution, and happily produces imputations whether the joint distribution exists or not. The question is whether the imputed data can be trusted when we cannot find a joint distribution $P(Y_1, Y_2)$ that has $P(Y_1|Y_2)$ and $P(Y_2|Y_1)$ as its conditionals.

For the trivariate case, the joint distribution $P(Y_1, Y_2, Y_3)$, if it exists, is uniquely specified by the following set of three conditionals: $P(Y_1|Y_2, Y_3)$, $P(Y_2|Y_3)$ and $P(Y_3|Y_1)$ (Gelman and Speed, 1993). Imputation under FCS typically specifies general forms for $P(Y_1|Y_2, Y_3)$, $P(Y_2|Y_1, Y_3)$ and $P(Y_3|Y_1, Y_2)$, which is different, and estimates the free parameters for these conditionals from the data. Typically, the number of parameters in imputation is much larger than needed to uniquely determine $P(Y_1, Y_2, Y_3)$. While perhaps inefficient as a parametrization, it is not easy to see why that in itself would introduce bias or affect the accuracy of the imputations.

Not much is known about the consequences of incompatibility on the quality of imputations. Simulations with strongly incompatible models found no adverse effects on the estimates after multiple imputation (van Buuren et al., 2006). Somewhat surprisingly, several methods based on deliberately specified incompatible methods outperformed complete case analysis. Imputation using the partially compatible Gibbs sampler seems to be robust against incompatible conditionals in terms of bias and precision, thus suggesting that incompatibility may be a relatively minor problem in multivariate imputation. More work is needed to verify such claims in more general and more realistic settings though.

In cases where the multivariate density is of genuine scientific interest, incompatibility clearly represents an issue because the data cannot be represented by a formal model. For example, incompatible conditionals could produce a ridge (or spike) in an otherwise smooth density, and the location of the ridge may actually depend on the stopping point. If such is the case, then we should have a reason to favor a particular stopping point. Alternatively, we might try to reformulate the imputation model so that the stopping point effect disappears. In imputation the objective is to make correct statistical inferences by augmenting the data and preserving the relations and uncertainty in the data. In that case, having a joint distribution may be convenient theoretically, but the price may be lack of fit. Gelman and Raghunathan (2001) remarked:

> One may argue that having a joint distribution in the imputation is less important than incorporating information from other variables and unique features of the dataset (e.g., zero/nonzero features in income components, bounds, skip patterns, nonlinearity, interactions.

In practice, incompatibility issues could arise in MICE if deterministic functions of the data are imputed along with their originals. For example, the imputation model may contain interaction terms, data summaries or nonlinear functions of the data. Such terms may introduce feedback loops and impossible combinations into the system, which can invalidate the imputations (van Buuren and Groothuis-Oudshoorn, 2011). It is important to diagnose this behavior, and eliminate feedback loops from the system. Section 1.6.5 describes the tools to do this. Apart from potential feedback problems, it appears that incompatibility is a relatively minor problem in practice, especially if the amount of missing data is modest.

Further theoretical work has been done by Arnold et al. (2002). The field has recently become active. Several methods for identifying compatibility from actual data have been developed in the last few years (Tian et al., 2009; Ip and Wang, 2009; Tan et al., 2010; Wang and Kuo, 2010; Kuo and Wang, 2011; Chen, 2011). It is not yet known what the added value of such methods will be in the context of missing data.

1.5.5 Number of iterations

When m sampling streams are calculated in parallel, monitoring convergence is done by plotting one or more statistics of interest in each stream against iteration number t. Common statistics to be plotted are the mean and standard deviation of the synthetic data, as well as the correlation between different variables. The pattern should be free of trend, and the variance within a chain should approximate the variance between chains.

In practice, a low number of iterations appears to be enough. Brand (1999) and (van Buuren et al., 1999) set the number of iterations T quite low, usually somewhere between 5 to 20 iterations. This number is much lower than in other applications of MCMC methods, which often require thousands of iterations.

The explanation for the pleasant property is that the imputed data $\dot{Y}^m$ form the only memory of the MICE algorithm. Imputations are created in a stepwise optimal fashion that adds a proper amount of random noise to the predicted values (depending on the strength of the relations between the variables), which helps to reduce the autocorrelation between successive draws. Hence, convergence will be rapid, and in fact immediate if all variables are independent. Thus, the incorporation of noise into the multiply-imputed data has the pleasant side effect of speeding up convergence. Situations to watch out for include:

- the correlations between the Y_js are high;
- the missing data rates are high;
- constraints on parameters across different variables exist.

The first two conditions directly affect the amount of autocorrelation in the system. The latter condition becomes relevant for customized imputation models. A useful trick for reducing the amount of autocorrelation in highly correlated repeated measures Y_1 and Y_2 is to draw imputations $\dot{\delta}$ for the increment $Y_2 - Y_1$ rather than for Y_2. Imputations are then calculated as the sum of the previous value and the increment, $Y_1 + \dot{\delta}$.

Simulation work suggests that FCS can work well using no more than just five iterations, but many more iterations might be needed in problems with high correlations and high proportions of missing data (van Buuren, 2007, 2012).

It is important to investigate convergence by inspecting traces lines of critical parameters the Gibbs samplers, as these might point to anomalies in the imputed data. (van Buuren and Groothuis-Oudshoorn, 2011) shows several cases with problematic convergence of the MICE algorithm, and may even be entirely stuck because of circularities. Also, imputing large blocks of correlated data may produce degenerate solutions van Buuren (2012, pp. 208). Such cases can often be prevented by simplifying the prediction model.

In general, we should be careful about convergence in missing data problems with high correlations and high missing data rates. On the other hand, we really have to push the MICE algorithm to its limits to see adverse effect. Of course, it never hurts to do a couple of extra iterations, but in most applications good results can be obtained with a small number of iterations.

1.5.6 Performance

Each conditional density has to be specified separately, so FCS requires (sometimes considerable) modeling effort on the part of the user. Most software provides reasonable defaults for standard situations, so the actual effort required may be small.

A number of simulation studies provide evidence that FCS generally yields estimates that are unbiased and that possess appropriate coverage (Brand, 1999; Raghunathan et al., 2001; Brand et al., 2003; Tang et al., 2005; van Buuren et al., 2006; Horton and Kleinman, 2007; Yu et al., 2007). FCS and joint modeling will often find similar estimates, especially for estimates that depend on the center of the distribution, like mean, median, regression estimates, and so on. Lee and Carlin (2010) contrasted the multivariate normal joint model to FCS using a simulation of a typical epidemiological set-up. They found that both the FCS and joint modeling provided substantial gains over CCA when estimating the binary intervention effect. The joint model appeared to perform well even in the presence of binary and ordinal variables. The regression estimates pertaining to a skewed variable were biased when normality was assumed. Transforming to normality (in joint modeling or FCS) or using predictive mean matching (in FCS) could resolve this problem.

The studies by van der Palm et al. (2012) and Gebregziabher (2012) compared various imputation methods for fully categorical data, based on both joint modeling and FCS. They reported some mild improvements of a more recent latent class model over a standard FCS model based on logistic regression. Both studies indicate that the number of latent classes needs to be large. Additional detail can be found in the original papers.

Although the substantive conclusions are generally robust to the precise form of the imputation model, the use of the multivariate normal model, whether rounded or not, is generally inappropriate for the imputation of categorical data (van Buuren, 2007). The problem is that the imputation model is more restrictive than the complete-data model, an undesirable situation known as uncongeniality (Meng, 1994; Schafer, 2003). More particularly, the multivariate normal model assumes that categories are equidistant, that the relations between all pairs of variables is linear, that the residual variance is the same at every predicted value, and that no interactions between variables exist. Without appropriate assessment of the validity of these assumptions, imputation may actually introduce systematic biases into the data that we may not be aware of. For example, Lee et al. (2012) demonstrated that imputing ordinal variables as continuous can lead to bias in the estimation of the exposure outcome association in the presence of a non-linear relationship. It may seem a trivial remark, but continuous data are best imputed by methods designed for continuous data, and categorical data are best imputed by methods designed for categorical data.

1.6 Modeling in FCS

1.6.1 Overview

The specification of the imputation model is the most challenging step in multiple imputation. The imputation model should

- account for the process that created the missing data,
- preserve the relations in the data, and
- preserve the uncertainty about these relations.

The idea is that adherence to these principles will yield proper imputations, and thus result in valid statistical inferences (Rubin, 1987, pp. 118-128). Van Buuren and Groothuis-Oudshoorn (2011) list the following seven choices:

1. First, we should decide whether the MAR assumption is plausible. Chained equations can handle both MAR and MNAR, but multiple imputation under MNAR requires additional modeling assumptions that influence the generated imputations.

2. The second choice refers to the form of the imputation model. The form encompasses both the structural part and the assumed error distribution. In FCS the form needs to be specified for each incomplete column in the data. The choice will be steered by the scale of the variable to be imputed, and preferably incorporates knowledge about the relation between the variables.

3. A third choice concerns the set of variables to include as predictors in the imputation model. The general advice is to include as many relevant variables as possible, including their interactions. This may, however, lead to unwieldy model specifications.

4. The fourth choice is whether we should impute variables that are functions of other (incomplete) variables. Many datasets contain derived variables, sum scores, ratios and so on. It can be useful to incorporate the transformed variables into the multiple imputation algorithm.

5. The fifth choice concerns the order in which variables should be imputed. The visit sequence may affect the convergence of the algorithm and the synchronization between derived variables.

6. The sixth choice concerns the setup of the starting imputations and the number of iterations. The convergence of the MICE algorithm can be monitored by trace lines.

7. The seventh choice is m, the number of multiply imputed datasets. Setting m too low may result in large simulation error and statistical inefficiency, especially if the fraction of missing information is high.

The above points are by no means exhaustive. Much sensible advice on modeling in multiple imputation in an FCS context also can be found in Sterne et al. (2009), White et al. (2011b) and Carpenter and Kenward (2013). The remainder of this section discusses points 1–5. Point 6 was already addressed in section 1.5.5, while point 7 was discussed in Chapter 13.

1.6.2 MAR or MNAR?

The most important decision in setting up an imputation model is to determine whether the available data are enough to solve the missing data problem at hand. The MAR assumption is essentially the belief that the available data are sufficient to correct for the missing data. Unfortunately, the distinction between MAR and MNAR cannot, in general, be made from the data. In practice, 99% of the analysts assume MAR, sometimes explicitly, but often more so implicitly. While MAR is often useful as a starting point, the actual causes of the missingness may be related to the quantities of scientific interest, even after accounting for the data. An incorrect MAR assumption may then produce biased estimates.

Collins et al. (2001) investigated the role of 'lurking' variables Z that are correlated with the variables of interest Y and with the missingness of Y. For linear regression, they found that if the missing data rate did not exceed 25% and if the correlation between the Z and Y was 0.4, omitting Z from the imputation model had a negligible effect. For more extreme situations (50% missing data and/or a correlation of 0.9) the effect depended strongly on the form of the missing data mechanism. When the probability of being missing was linear in Z, then omitting Z from the imputation model only affected the intercept, whereas the regression weights and variance estimates were unaffected. When more missing data were created in the extremes, the reverse occurred: omitting Z biased the regression coefficients and variance estimates, but the intercept was unbiased with the correct confidence interval. In summary, they found that all estimates under multiple imputation appeared robust against MNAR. Beyond a correlation of 0.4, or for a missing data rate over 25%, it is the form of the missing data mechanism that determines which parameters will be biased.

While these results are generally comforting, there are three main strategies that we might pursue if the response mechanism is nonignorable:

- Expand the data in the imputation model in the hope of making the missing data mechanism closer to MAR;

- Formulate an explicit nonresponse model in combination with a complete-data model, and estimate the parameters of interest;
- Formulate and fit a series of nonignorable imputation models, and perform sensitivity analysis on the critical parameters.

In the first strategy, the MAR assumption is the natural starting point. MAR could be made more plausible by finding additional data that are strongly predictive of the missingness, and include these into the imputation model. The fact that these are included is more important than the precise form in which that is done (Jolani et al., 2013).

There is a large literature on the second option, often starting from the Heckman model (Heckman, 1979). There has been some recent work to generate multiple imputations under this model, as well as generalizations thereof. The key idea is the extend the imputation model with a model for the missing data process, where the probability of being missing depends on the variable Y_j to be imputed. The FCS framework can be used to generate imputations under the combined model by drawing imputations for Y_j and R_j. See Jolani (2012) for details.

Finally, one might perform a concise simulation study as in Collins et al. (2001) customized for the problem at hand with the goal of finding out how extreme the MNAR mechanism needs to be to influence the parameters of scientific interest. More generally, the use of sensitivity is advocated by the Panel on Handling Missing Data in Clinical Trials of the National Research Council (Council, 2010). Chapter 20 of the Handbook deals with sensitivity analysis using multiple imputation.

1.6.3 Model form

The MICE algorithm requires a specification of a univariate imputation method separately for each incomplete variable. It is important to select univariate imputation methods that have correct statistical coverage for the scientific parameters of interest, and that yield sensible imputed values. The measurement level of a variable largely determines the form of the univariate imputation model. There are special methods for continuous, dichotomous, ordered categories, unordered categories, count data, semi-continuous data, censored data, truncated data and rounded data. In addition, there are regression tree imputation methods aimed at preserving interactions. Chapter 3 of van Buuren (2012) contains an in-depth treatment of many univariate imputation methods. Predictive mean matching (Little, 1988) is an allround imputation method that works well in many cases. It is the default method in MICE for imputing continuous data (van Buuren and Groothuis-Oudshoorn, 2011).

Model specification is straightforward when the data are cross-sectional or longitudinal, where in the longitudinal setting, different time points are coded as different columns, i.e. as a *broad* matrix. Genuinely hierarchical data are typically coded as a *long* matrix, with time points or nested observations are coded as distinct rows. Van Buuren (2011) discusses and compares three imputation methods for hierarchical data:

- Ignore any clustering structure in the data, and use standard imputation techniques tools for nonhierarchical data;
- Add the cluster allocation as a fixed factor, thus allowing for between class variation by a fixed effects model;
- Draw imputations under a linear mixed-effect model by a Markov Chain Monte Carlo algorithm.

Ignoring the multilevel structure when in fact it is present will bias the intra-class correlation downwards, adding a fixed factor will bias it upwards, while the linear mixed-effects model is about right. In general, smaller class sizes complicate the imputation problem. Overall, imputation under the linear mixed-effects model is superior to the two other methods, but it is not yet ideal as the coverage may fail to achieve the nominal level. Computational details can be found in van Buuren (2012, pp. 84–87). Moreover, Chapter 9 of that book contains two applications on longitudinal data, one using a broad matrix in a repeated measured problem, and the other using the linear mixed-effects imputation model on the long matrix.

1.6.4 Predictors

The general advice is to include as many variables in the imputation model as possible (Meng, 1994; Collins et al., 2001), but there are necessarily computational limitations that must be taken into account. Conditioning on all other data is often reasonable for small to medium datasets, containing up to, say, 20–30 variables, without derived variables, interactions effects and other complexities. Including as many predictors as possible tends to make the MAR assumption more plausible, thus reducing the need to make special adjustments for MNAR.

For datasets containing hundreds or thousands of variables, using all predictors may not be feasible (because of multicollinearity and computational problems) to include all these variables. It is also not necessary. In practice, 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. Hardt et al. (2012) suggested that the number of complete rows in the imputation model should be at least three times the number of variables. Van Buuren et al. (1999) provide the following strategy for selecting predictor variables from a large database:

1. Include all variables that appear in the complete data model, i.e., the model that will be applied to the data after imputation, including the

outcome (Little, 1992; Moons et al., 2006). Failure to include the outcome will bias the complete data analysis, especially if the complete data model contains strong predictive relations. Note that this step is somewhat counter-intuitive, as it may seem that imputation would artificially strengthen the relations of the complete data model, which would be clearly undesirable. If done properly however, this is not the case. On the contrary, not including the complete data model variables will tend to bias the results toward zero. Note that interactions of scientific interest also need to be included in the imputation model.

2. In addition, include the variables that are related to the nonresponse. Factors that are known to have influenced the occurrence of missing data (stratification, reasons for nonresponse) are to be included on substantive grounds. Other variables of interest are those for which the distributions differ between the response and nonresponse groups. These can be found by inspecting their correlations with the response indicator of the variable to be imputed. If the magnitude of this correlation exceeds a certain level, then the variable should be included.

3. In addition, include variables that explain a considerable amount of variance. Such predictors help reduce the uncertainty of the imputations. They are basically identified by their correlation with the target variable. Only include predictors with a relatively high outflux coefficient (cf. section 1.3).

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. A simple indicator is the percentage of observed cases within this subgroup, the percentage of usable cases (cf. section 1.3).

Most predictors used for imputation are 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 and Groothuis-Oudshoorn (1999), but it may miss important predictors of predictors. A safer and more efficient, though more laborious, strategy is to perform the modeling steps also for the predictors of predictors of key variables. This is done in Groothuis-Oudshoorn et al. (1999). At the terminal node, one can apply a simple method, like sampling from the marginal, that does not need any predictors for itself.

By default, most computer programs impute a variable Y_j from all other variables Y_{-j} in the data. Some programs, however, have the ability to specify the set of predictors to be used per incomplete variable (Su et al., 2011; van

Buuren and Groothuis-Oudshoorn, 2011; Royston and White, 2011). These facilities are highly useful for refining the imputation model and for customizing the imputations to the data. Ridge regression (Hoerl and Kennard, 1970; Tibshirani, 1996) provides a alternative way to control the estimation process, making the algorithm more robust at the expense of bias.

Although it might seem somewhat laborious, the quality of the imputed values can be enhanced considerably by a judicious specification of the set of predictors that enter imputation model. It is generally worthwhile to set apart some time to set up the imputation model, often in combination with the use of suitable diagnostics (c.f. section 1.7).

1.6.5 Derived variables

Derived variables (transformations, recodes, interaction terms, and so on) pose special challenges for the imputation model. There are three general strategies to impute derived variables:

1. Leave derived variables out of the imputation, and calculate them afterwards from the multiply-imputed data;

2. Calculate derived variables before imputation, and impute them as usual;

3. Update derived variables within the imputation algorithm as soon as one of the original variables is imputed.

Method 1 is easy, but the generated imputations do not account for the relationship between the derived variables and other variables in the data, potentially resulting is biased estimates in the complete-data analysis. Method 2 repairs this deficit, but at the expense of creating inconsistencies between the imputations of the originals and of the derived versions. Method 3 can address both problems, but some care is needed in setting up the predictor matrix. This section looks briefly at various types of derived variables.

In practice, there is often extra knowledge about the data that is not modeled explicitly. For example, consider the weight/height ratio, defined as weight divided by height (kg/m). If any one of the triplet height, weight or weight/height ratio is missing, then the missing value can be calculated with certainty by a simple deterministic rule. Unless we specify otherwise, the default imputation model is however unaware of the relation between the three variables, and will produce imputations that are inconsistent with the rule. Inconsistent imputations are undesirable since they yield combinations of data values that are impossible had the data been observed.

The easiest way to deal with the problem is to leave any derived data outside the imputation process (Method 1). For example, we may impute any missing height and weight data, and append weight/height ratio to the imputed data afterward. The disadvantage is this post-imputation method is that the derived variable is not available for imputation, potentially resulting in incorrect statistical inferences.

Another possibility is to calculate the weight/height ratio before imputation, and impute it as "just another variable," an approach known as JAV (Method 2). Although JAV may yield valid statistical inferences in particular cases (for linear regression weights under MCAR, see von Hippel (2009)), it invariably produces impossible combinations of imputed values, and may be biased under MAR.

A solution for this problem is *passive imputation*, a method that calculates the derived variable on the fly once one of its components is imputed (Method 3). Passive imputation maintains the consistency among different transformations of the same data (thus solving the problem of JAV) and makes the derived variable available for imputation (thus solving the problem of the post-imputation method).

Care is needed in setting up the predictor matrix when using passive imputation. In particular, we may not use the derived variable as a predictor for its components, so feedback loops between the derived variables and their originals should be broken. In the above example, we would thus need to remove the weight/height ratio from the imputation models for height and weight. Failing to do so may result in absurd imputations and problematic convergence.

Figure 1.3 compares JAV to passive imputation on real data. The leftmost panel in Figure 1.3 shows the results of JAV. The imputations are far off any of the observed data, since JAV ignores the fact that the weight/height ratio is a function of height and weight. The middle panel shows that passive imputation represents an improvement over JAV. The values are generally similar to the real data and adhere to the derived rules. The rightmost panel shows that somewhat improved imputations can be obtained by preventing that the body mass index (BMI) and weight/height ratio (which have an exact nonlinear relationship) are simultaneous predictors.

The *sum score* is another type of derived variable. The sum score undefined if one of the original variables is missing. Sum scores of imputed variables are useful within the MICE algorithm to economize on the number of predictors. Van Buuren (2010) reports a simulation on sub scale scores from imputed questionnaire items that shows that plain multiple imputation using sum scores improves upon dedicated imputation methods. See sections 7.3 and 9.2 in van Buuren (2012) for applications on real data.

Interaction terms are also derived variables. The standard MICE algorithm only accommodates main effects. Sometimes the *interaction* between variables is of scientific interest. For example, in a longitudinal study we could be interested in assessing whether the rate of change differs between two treatment groups, in other words, the treatment-by-group interaction. The standard algorithm does not take interactions into account, so the interactions of interest should be added to the imputation model. Interactions can be added using passive imputation. An alternative is to impute the data in separate groups.

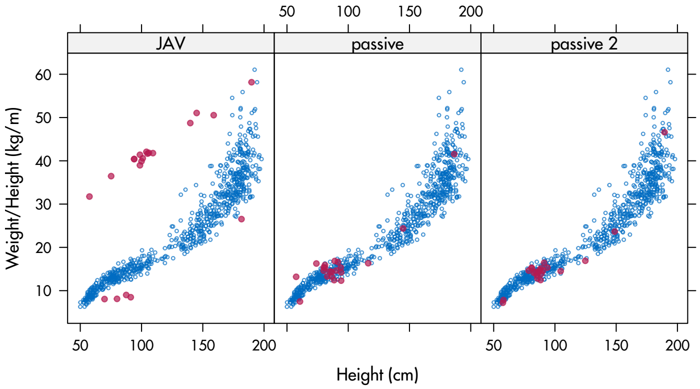

FIGURE 1.3
Three different imputation models to impute weight/height ratio. The relation between the weight/height ratio and height is not respected under "just another variable" (JAV). Both passive methods yield imputations that are close to the observed data. "Passive 2" does not allow for models in which weight/height ratio and BMI are simultaneous predictors.

In some cases it makes sense to restrict the imputations, possibly conditional on other data. For example, if we impute 'male', we can skip questions particular to females, e.g. about pregnancy. Such *conditional imputation* could reset the imputed data in the pregnancy block to missing, thus imputing only part depending on gender. Of course, appropriate care is needed when using the pregnancy variables are used later as a predictor to restrict to females. Such alterations to the imputations can be implemented easily within a FCS framework by *post-processing imputations* within the iterative algorithm.

Compositional data are another form of derived data, and often occur in household and business surveys. Sometimes we know that a set of variables should add up to a given total. If one of the additive terms is missing, we can directly calculate its value with certainty by deducting the known terms from the total. However, if two additive terms are missing, imputing one of these terms uses the available one degree of freedom, and hence implicitly determines the other term. Imputation of compositional data has only recently received attention (Tempelman, 2007; Hron et al., 2010; de Waal et al., 2011), and can be implemented conveniently with an FCS framework. See section 5.4.5 in van Buuren (2012) for an illustration of the main idea.

Nonlinear relations are often modeled using a linear model by adding quadratic or cubic terms of the explanatory variables. Creating imputed values that adhere to *quadratic relation* poses some challenges. Current imputation methodology either preserves the quadratic relation in the data and biases the estimates of interest, or provides unbiased estimates but does not preserve the quadratic relation (von Hippel, 2009; White et al., 2011a). An alternative approach that aims to define a *polynomial combination* Z as $Z = Y\beta_1 + Y^2\beta_2$ for some β_1 and β_2. The idea is to impute Z instead of Y and Y^2, followed by a decomposition of the imputed data Z into components Y and Y^2. Section 5.4.6 in van Buuren (2012) provided an algorithm that does these calculations. Simulations indicate that the quadratic method worked well in a variety of situations (Vink and van Buuren, 2013).

In all cases, feedback between different versions of the same variable should be prevented. Failing to do so may may lock up the MICE algorithm or produce erratic imputations.

1.6.6 Visit sequence

The MICE algorithm as described in section 1.5.2 imputes incomplete variables in the data from left to right. Theoretically, the visit sequence of the MICE algorithm is irrelevant as long as each column is visited often enough, though some schemes are more efficient than others. In practice, there are small order effects of the MICE algorithm, where the parameter estimates depend on the sequence of the variables. To date, there is little evidence that this matters in practice, even for clearly incompatible imputation models (van Buuren et al., 2006). For monotone missing data, convergence is immediate if variables are ordered according to their missing data rate. Rather than reordering the data itself, it is more convenient to change the visit sequence of the algorithm.

It may also be useful to visit a given variable more than once within the same iteration. For example, weight/height ratio can be recalculated immediately after the missing data in weight and after the missing data in height are imputed. This ensures that the weight/height ratio remains properly synchronized with both weight and height at all times.

1.7 Diagnostics

An important and unique advantage of multiple imputation over other statistical techniques is that we can easily infer the plausibility of the statistical (imputation) model. This is straightforward because the imputation model produces data, and we are very well equipped to look at data.

One of the best tools for assessing the plausibility of imputations is to study the discrepancy between the observed and imputed data. The idea is that high quality imputed data will have distributions similar to the observed data. Except under MCAR, the distributions do not need to be identical, as strong MAR mechanisms may induce systematic differences between the two distributions. However, any dramatic differences between the imputed and observed data (such as seeing a body mass index of 300 in the imputed data) should certainly alert us to the possibility that something is wrong with the imputation model. It is reassuring when the synthetic data could have been real values had they not been missing.

Suppose we compare density estimates of the observed and imputed data. Some type of discrepancies that are of interest are

- the points have different means;
- the points have different spreads;
- the points have different scales;
- the points have different relations;
- the points do not overlap and they defy common sense.

Such differences between the densities data may suggest a problem that needs to be further checked. Other useful graphic representation include the box plot, the stripplot, the histogram, and the scattergram of variables, stratified according to whether the data are real or imputed.

Figure 1.4 shows kernel density estimates of imputed and observed data. In this case, the distributions match up well. Other imputation diagnostics have been suggested by Gelman et al. (2005), Raghunathan and Bondarenko (2007), Abayomi et al. (2008) and Su et al. (2011).

Compared to diagnostic methods for conventional statistical models, imputation comes with the advantage that we can directly compare the observed and imputed values. Unfortunately, diagnostics are currently underused. One reason is that not all software properly supports diagnostics. Another reason is that the imputer may put too much trust into the appropriateness of the defaults of the software for the data at hand. Absurd imputations are however easy to spot by simple methods, and should be repaired before attempting complete-data analysis.

1.8 Conclusion

FCS has rapidly been adopted by applied researchers in many branches of science. FCS remains close to the data and is easy to apply. The relevant software is now widespread, and available in all major statistical packages. Appendix A of van Buuren (2012) is an overview of software for FCS.

The technology has now evolved into the standard way of creating multiple imputations. Of course, there are still open issues, and more experience is needed with practical application of FCS. Nevertheless, FCS is an open and modular technology that will continue to attract the attention of researchers who want to solve their missing data problems.

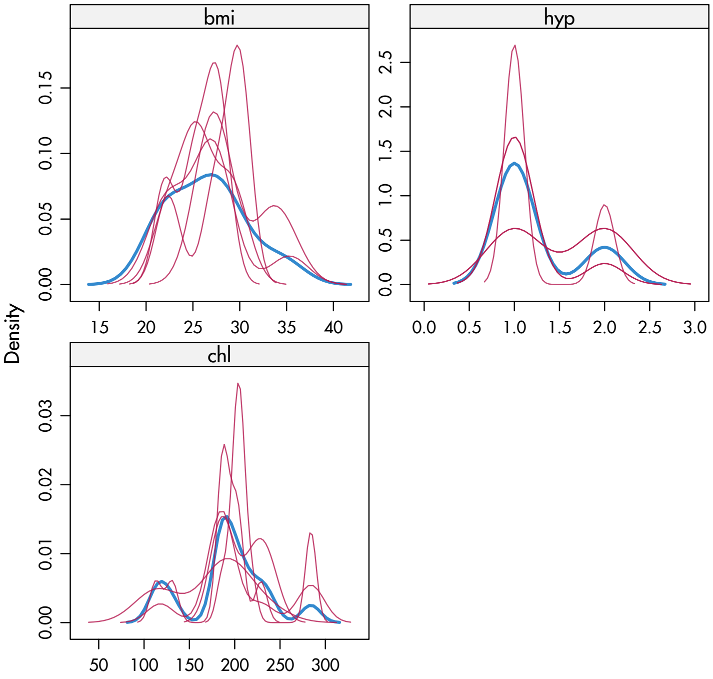

FIGURE 1.4
Kernel density estimates for the marginal distributions of the observed data (thick line) and the $m = 5$ densities per variable calculated from the imputed data (thin lines).

Bibliography

K. Abayomi, A. Gelman, and M. Levy. Diagnostics for multivariate imputations. *Journal of the Royal Statistical Society.Series C: Applied Statistics*, 57(3):273–291, 2008.

B. C. Arnold, E. Castillo, and J. M. Sarabia. Exact and near compatibility of discrete conditional distributions. *Computational Statistics and Data Analysis*, 40(2):231–252, 2002.

M. S. Bartlett. *An Introduction to Stochastic Processes.* Press Syndicate of the University of Cambridge, Cambridge, UK, 1978.

J. Besag. Spatial interaction and the statistical analysis of lattice systems. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 36(2):192–236, 1974.

J. P. L. Brand. *Development, Implementation and Evaluation of Multiple Imputation Strategies for the Statistical Analysis of Incomplete Data Sets.* PhD thesis, Erasmus University, Rotterdam, 1999.

J. P. L. Brand, S. van Buuren, C. G. M. Groothuis-Oudshoorn, and E. S. Gelsema. A toolkit in SAS for the evaluation of multiple imputation methods. *Statistica Neerlandica*, 57(1):36–45, 2003.

Samuel F Buck. A method of estimation of missing values in multivariate data suitable for use with an electronic computer. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 22(2):302–306, 1960.

James Carpenter and Michael Kenward. *Multiple imputation and its application.* John Wiley & Sons, 2013.

G. Casella and E. I. George. Explaining the Gibbs sampler. *The American Statistician*, 46(3):167–174, 1992.

H. Y. Chen. Compatibility of conditionally specified models. *Statistics and Probability Letters*, 80(7-8):670–677, 2011.

L. M. Collins, J. L. Schafer, and C. M. Kam. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychol. Methods*, 6(3):330–351, 2001.

Panel on Handling Missing Data in Clinical Trials; National Research Council. *The Prevention and Treatment of Missing Data in Clinical Trials.* The National Academies Press, Washington, D.C., 2010.

T. de Waal, J. Pannekoek, and S. Scholtus. *Handbook of Statistical Data Editing and Imputation*. John Wiley & Sons, Hoboken, NJ, 2011.

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

Mulugeta Gebregziabher. Lessons learned in dealing with missing race data: An empirical investigation. *Journal of Biometrics & Biostatistics*, 2012.

A. E. Gelfand and A. F. M. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410): 398–409, 1990.

A. Gelman. Parameterization and Bayesian modeling. *Journal of the American Statistical Association*, 99(466):537–545, 2004.

A. Gelman and T. E. Raghunathan. Discussion of Arnold et al. "conditionally specified distributions". *Statistical Science*, 16:249–274, 2001.

A. Gelman and T. P. Speed. Characterizing a joint probability distribution by conditionals. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 55(1):185–188, 1993.

A. Gelman, I. van Mechelen, M. Meulders, G. Verbeke, and D. F. Heitjan. Multiple imputation for model checking: Completed-data plots with missing and latent data. *Biometricss*, 61(1):74–85, 2005.

W. R. Gilks. Full conditional distributions. In R. Gilks, W., S. Richardson, and J. Spiegelhalter, D., editors, *Markov Chain Monte Carlo in Practice(5)*, pages 75–88. Chapman & Hall, London, 1996.

Terry C Gleason and Richard Staelin. A proposal for handling missing data. *Psychometrika*, 40(2):229–252, 1975.

M.S. Gold and P.M. Bentler. Treatments of missing data: A Monte Carlo comparison of RBHDI, iterative stochastic regression imputation, and expectation-maximization. *Structural Equation Modeling*, 7:319–355, 2000.

C. G. M. Groothuis-Oudshoorn, S. van Buuren, and J. L. A. van Rijckevorsel. *Flexible Multiple Imputation by Chained Equations of the AVO-95 Survey*, volume (PG/VGZ/00.045). TNO Prevention and Health, Leiden, 1999.

Jochen Hardt, Max Herke, and Rainer Leonhart. Auxiliary variables in multiple imputation in regression with missing X: A warning against including too many in small sample research. *BMC medical research methodology*, 12 (1):184, 2012.

D. Heckerman, D. M. Chickering, C. Meek, R. Rounthwaite, and C. Kadie. Dependency networks for inference, collaborative filtering, and data visualisation. *Journal of Machine Learning Research*, 1(1):49–75, 2001.

James J Heckman. Sample selection bias as a specification error. *Econometrica: Journal of the econometric society*, pages 153–161, 1979.

Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

N. J. Horton and K. P. Kleinman. Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *The American Statistician*, 61(1):79–90, 2007.

K. Hron, M. Templ, and P. Filzmoser. Imputation of missing values for compositional data using classical and robust methods. *Computational Statistics and Data Analysis*, 54(12):3095–3107, 2010.

E. H. Ip and Y. J. Wang. Canonical representation of conditionally specified multivariate discrete distributions. *Journal of Multivariate Analysis*, 100 (6):1282–1290, 2009.

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

S. Jolani. *Dual Imputation Strategies for Analyzing Incomplete Data*. PhD thesis, University of Utrecht, Utrecht, 2012.

S. Jolani, S. van Buuren, and L. E. Frank. Combining the complete-data and nonresponse models for drawing imputations under MAR. *Journal of Statistical Computation and Simulation*, 83(5):868–879, 2013.

A. B. Kennickell. Imputation of the 1989 survey of consumer finances: Stochastic relaxation and multiple imputation. In *Proceedings of the Section on Survey Research Methods. Joint Statistical Meeting 1991*, pages 1–10. ASA, Alexandria, VA, 1991.

K-L Kuo and Y. J. Wang. A simple algorithm for checking compatibility among discrete conditional distributions. *Computational Statistics and Data Analysis*, 55(8):2457–2462, 2011.

K. J. Lee and J. B. Carlin. Multiple imputation for missing data: Fully conditional specification versus multivariate normal imputation. *Am. J. Epidemiol.*, 171(5):624–632, 2010.

Katherine J Lee, John C Galati, Julie A Simpson, and John B Carlin. Comparison of methods for imputing ordinal data using multivariate normal imputation: a case study of nonâĂŘlinear effects in a large cohort study. *Stat. Med.*, 31(30):4164–4174, 2012.

K-H Li. Imputation using Markov chains. *Journal of Statistical Computation and Simulation*, 30(1):57–79, 1988.

R. J. A. Little. Missing-data adjustments in large surveys (with discussion). *Journal of Business Economics and Statistics*, 6(3):287–301, 1988.

R. J. A. Little. Regression with missing Xs: A review. *Journal of the American Statistical Association*, 87(420):1227–1237, 1992.

R. J. A. Little and D. B. Rubin. *Statistical Analysis with Missing Data.* John Wiley & Sons, New York, 2002.

Roderick J. Little. In praise of simplicity not mathematistry! ten simple powerful ideas for the statistical scientist. *Journal of the American Statistical Association*, 108(502):359–369, 2013.

X-L Meng. Multiple imputation with uncongenial sources of input (with discusson). *Statistical Science*, 9(4):538–573, 1994.

K. G. M. Moons, A. R. T. Donders, T. Stijnen, and F. E. Harrell. Using the outcome for imputation of missing predictor values was preferred. *J. Clin. Epidemiol.*, 59(10):1092–1101, 2006.

T. E. Raghunathan and I. Bondarenko. Diagnostics for multiple imputations. *Unpublished*, 2007.

T. E. Raghunathan, J. M. Lepkowski, J. van Hoewyk, and P. W. Solenberger. A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, 27(1):85–95, 2001.

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

G. O. Roberts. Markov chain concepts related to sampling algorithms. In R. Gilks, W., S. Richardson, and J. Spiegelhalter, D., editors, *Markov Chain Monte Carlo in Practice(3)*, pages 45–57. Chapman & Hall, London, 1996.

Patrick Royston and Ian R White. Multiple imputation by chained equations (MICE): Implementation in Stata. *J. Stat. Softw.*, 45(4):1–20, 2011.

D. B. Rubin. *Multiple Imputation for Nonresponse in Surveys.* John Wiley & Sons, New York, 1987.

D. B. Rubin. Discussion on multiple imputation. *International Statistical Review*, 71(3):619–623, 2003a.

D. B. Rubin. Nested multiple imputation of NMES via partially incompatible MCMC. *Statistica Neerlandica*, 57(1):3–18, 2003b.

D. B. Rubin and J. L. Schafer. Efficiently creating multiple imputations for incomplete multivariate normal data. In *ASA 1990 Proceedings of the Statistical Computing Section*, pages 83–88. ASA, Alexandria, VA, 1990.

J. L. Schafer. *Analysis of Incomplete Multivariate Data.* Chapman & Hall, London, 1997.

J. L. Schafer. Multiple imputation in multivariate problems when the imputation and analysis models differ. *Statistica Neerlandica*, 57(1):19–35, 2003.

J. A. Sterne, I. R. White, J. B. Carlin, M. Spratt, P. Royston, M. G. Kenward, A. M. Wood, and J. R. Carpenter. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *Brit. Med. J.*, 338:b2393, 2009.

Y. S. Su, A. Gelman, J. L. Hill, and M. Yajimi. Multiple imputation with diagnostics (mi) in R: Opening windows into the black box. *J. Stat. Softw.*, 45(2), 2011.

M. T. Tan, G-L Tian, and K. W. Ng. *Bayesian Missing Data Problem. EM, Data Augmentation and Noniterative Computation.* Chapman & Hall/CRC, Boca Raton, FL, 2010.

L. Tang, J. Song, T. R. Belin, and J. Ununtzer. A comparison of imputation methods in a longitudinal randomized clinical trial. *Stat. Med.*, 24(14): 2111–2128, 2005.

D. C. G. Tempelman. *Imputation of Restricted Data.* PhD thesis, University of Groningen, Groningen, 2007.

G-L Tian, M. T. Tan, K. W. Ng, and M-L Tang. A unified method for checking compatibility and uniqueness for finite discrete conditional distributions. *Communications in Statistics - Theory and Methods*, 38(1):115–129, 2009.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, pages 267–288, 1996.

L. Tierney. Introduction to general state-space Markov chain theory. In R. Gilks, W., S. Richardson, and J. Spiegelhalter, D., editors, *Markov Chain Monte Carlo in Practice(4)*, pages 59–74. Chapman & Hall, London, 1996.

S. van Buuren. Multiple imputation of discrete and continuous data by fully conditional specification. *Stat. Methods Med. Res.*, 16(3):219–242, 2007.

S. van Buuren. Item imputation without specifying scale structure. *Methodology*, 6(1):31–36, 2010.

S. van Buuren. Multiple imputation of multilevel data. In J. Hox, J. and K. Roberts, J., editors, *The Handbook of Advanced Multilevel Analysis(10)*, pages 173–196. Routledge, Milton Park, UK, 2011.

S. van Buuren. *Flexible Imputation of Missing Data.* Chapman & Hall/CRC Press, Boca Raton, FL, 2012.

S. van Buuren and C. G. M. Groothuis-Oudshoorn. *Flexible Multivariate Imputation by MICE*, volume (PG/VGZ/99.054). TNO Prevention and Health, Leiden, 1999.

S. van Buuren and C. G. M. Groothuis-Oudshoorn. *Multivariate Imputation by Chained Equations: MICE V1.0 User manual*, volume PG/VGZ/00.038. TNO Prevention and Health, Leiden, 2000.

S. van Buuren and K. Groothuis-Oudshoorn. MICE: Multivariate imputation by chained equations in R. *J. Stat. Softw.*, 45(3):1–67, 2011.

S. van Buuren and J. L. A. van Rijckevorsel. Imputation of missing categorical data by maximizing internal consistency. *Psychometrika*, 57(4):567–580, 1992.

S. van Buuren, H. C. Boshuizen, and D. L. Knook. Multiple imputation of missing blood pressure covariates in survival analysis. *Stat. Med.*, 18(6): 681–694, 1999.

S. van Buuren, J. P. L. Brand, C. G. M. Groothuis-Oudshoorn, and D. B. Rubin. Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, 76(12):1049–1064, 2006.

Daniël W van der Palm, L Andries van der Ark, and Jeroen K Vermunt. A comparison of incomplete-data methods for categorical data. *Stat. Methods Med. Res.*, 2012.

G. Vink and S. van Buuren. Multiple imputation of squared terms. *Sociological Methods and Research*, 42(4):598–607, 2013.

P. T. von Hippel. How to impute interactions, squares, and other transformed variables. *Sociological Methodology*, 39(1):265–291, 2009.

Y. J. Wang and K-L Kuo. Compatibility of discrete conditional distributions with structural zeros. *Journal of Multivariate Analysis*, 101(1):191–199, 2010.

I. R. White, N. J. Horton, J. R. Carpenter, and S. J. Pocock. Strategy for intention to treat analysis in randomised trials with missing outcome data. *Brit. Med. J.*, 342:d40, 2011a.

Ian R White, Patrick Royston, and Angela M Wood. Multiple imputation using chained equations: Issues and guidance for practice. *Stat. Med.*, 30 (4):377–399, 2011b.

L-M Yu, A. Burton, and O. Rivero-Arias. Evaluation of software for multiple imputation of semi-continuous data. *Stat. Methods Med. Res.*, 16(3):243–258, 2007.