Submitted to Statistical Science

# Multiple Imputation: A Review of Practical and Theoretical Findings

**Jared S. Murray***

University of Texas at Austin



*Abstract.* Multiple imputation is a straightforward method for handling missing data in a principled fashion. This paper presents an overview of multiple imputation, including important theoretical results and their practical implications for generating and using multiple imputations. A review of strategies for generating imputations follows, including recent developments in flexible joint modeling and sequential regression/chained equations/fully conditional specification approaches. Finally, we compare and contrast different methods for generating imputations on a range of criteria before identifying promising avenues for future research.

*Key words and phrases:* missing data, proper imputation, congeniality, chained equations, fully conditional specification, sequential regression multivariate imputation.

## 1. INTRODUCTION

Multiple imputation (MI) (Rubin, 1987) is a simple but powerful method for dealing with missing data. MI as originally conceived proceeds in two stages: A data disseminator creates a small number of completed datasets by filling in the missing values with samples from an imputation model. Analysts compute their estimates in each completed dataset and combine them using simple rules to get pooled estimates and standard errors that incorporate the additional variability due to the missing data.

MI was originally developed for settings in which statistical agencies or other data disseminators provide multiply imputed databases to distinct end-users. There are a number of benefits to MI in this setting: The disseminator can support approximately valid inference for a wide range of potential analyses with a small set of imputations, and the burden of dealing with the missing data is on the imputer rather than the analyst. All analyses conducted on the publicly available files can be based on the same set of imputations, ensuring that differences in results are not due to the handling of missing data.

*12110 Speedway B6500, Austin, Texas. (e-mail: jared.murray@mccombs.utexas.edu).*

*The author gratefully acknowledges support from the National Science Foundation under grant numbers SES-1130706, SES-1631970 and DMS-1043903. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the funding agencies.

With the introduction of easy-to-use software to generate imputations and combine estimates it has become increasingly common for users to create their own imputations prior to analysis. The set of methods available to generate imputations has also grown substantially, from simple parametric models and resampling methods to iterative classification and regression tree-based algorithms and flexible Bayesian nonparametric models. There are several textbook treatments of multiple imputation (e.g. Rubin (1987); Little and Rubin (2002); Van Buuren (2012); Carpenter and Kenward (2013)) but fewer recent reviews of the variety of methods available to create multiply imputed files.

This paper provides a review of MI, with a focus on methods for generating imputations and the theoretical results and empirical evidence available to guide the selection and critique of imputation procedures. We restrict attention to methods for imputing item missing data (imputing the subset of values that are missing for an incomplete observation) in settings with independent observations. Much of the discussion also applies to other data structures, and to problems other than item missing data where MI has proven useful (see Reiter and Raghunathan (2007) for some examples of other uses for multiple imputation).

The paper proceeds as follows: Section 2 briefly reviews the mechanics of multiple imputation for a scalar estimand. Section 3 reviews the conditions under which the usual MI rules give valid inference. Section 4 summarizes the practical implications of the theoretical results, particularly for choosing a method for generating imputations. Section 5 reviews methods for imputing a single variable subject to missingness. Section 6 reviews methods for imputing several variables. Section 7 discusses some of the considerations for choosing an imputation model. Section 8 concludes with discussion and directions for future work.

## 2. MULTIPLE IMPUTATION: HOW DOES IT WORK?

Let $Y_i = (Y_{i1}, Y_{i2}, \dots Y_{ip})$ denote a $p-$dimensional vector of values corresponding to the $i^{th}$ unit and $R_i = (R_{i1}, R_{i2}, \dots R_{ip})$ be a vector of indicator variables representing the response pattern, where $R_{ij} = 1$ if $Y_{ij}$ is observed and is zero otherwise. We will use lowercase letters to distinguish fixed values from random variables, and denote the realized values in a particular dataset with a tilde (e.g., $R_i$ is a random vector, $r_i$ is a particular value that might be taken by $R_i$, and $\tilde{r}_i$ is the observed response pattern for unit $i$ observed in a particular dataset).

Let $R = \{R_i : 1 \leq i \leq n\}$ with $r$ and $\tilde{r}$ defined similarly. The observed and missing values from a dataset of size $n$ with response pattern $R$ are denoted $Y_{obs}(R) = \{Y_{ij} : r_{ij} = 1,\ 1 \leq j \leq p,\ 1 \leq i \leq n\}$ and $Y_{mis}(R) = \{Y_{ij} : r_{ij} = 0,\ 1 \leq j \leq p,\ 1 \leq i \leq n\}$, respectively. Where the explicit dependence on the response pattern is a distraction we will drop the functional notation and simply refer to $Y_{mis}$ and $Y_{obs}$.

We assume throughout that the missing data are *missing at random (MAR)* (Rubin, 1987), that is,

$$\Pr(R = \tilde{r} \mid Y_{obs}(\tilde{r}) = \tilde{y}_{obs}, Y_{mis}(\tilde{r}) = y_{mis}, \phi) \tag{2.1}$$

takes the same value for all $y_{mis}$ and $\phi$, where $\phi$ parameterizes our model of the response mechanism (the distribution of $(R \mid Y)$). Under MAR we do not need to explicitly model the response process to impute the missing data. (Rubin, 1987, Result 2.3). MI may be used for missing data that are not MAR provided

we explicitly model the response mechanism or make other identifying assumptions (see Rubin (2003a) for related discussion and examples of MI for non-MAR missing data).

### 2.1 Multiple imputation for a scalar estimand

Let $Q$ be an estimand of interest, which may be a function of complete data in a finite population or a model parameter. Let $\hat{Q}(Y)$ be an estimator of $Q$ with sampling variance $U$ estimated by $\hat{U}(Y)$; where there is no ambiguity we refer to these as $\hat{Q}$ and $\hat{U}$. In order to fix ideas we focus on scalar $Q$. Inference for vector $Q$ is similar in spirit; see (Rubin, 1987, Chapter 3), also (Schafer, 1997, Chapter 4, Section 3) or the review in (Reiter and Raghunathan, 2007, Section 2.1).

Assume $Y_{mis}^{(1)}, Y_{mis}^{(2)}, \ldots, Y_{mis}^{(M)}$ are $M$ imputations for $Y_{mis}$. Define $\hat{Q}^{(m)} = \hat{Q}(Y_{obs}, Y_{mis}^{(m)})$, the estimator computed using the $m^{th}$ completed dataset (with $\hat{U}^{(m)}$ defined similarly), and

$$\bar{Q}_M = \sum_{m=1}^{M} \frac{\hat{Q}^{(m)}}{M}, \quad \bar{U}_M = \sum_{m=1}^{M} \frac{\hat{U}^{(m)}}{M}, \quad B_M = \sum_{m=1}^{M} \frac{(\hat{Q}^{(m)} - \bar{Q}_M)^2}{M-1}. \tag{2.2}$$

These statistics form the basis for inference under MI: $\bar{Q}_M$ averages the estimate computed in each imputed dataset to obtain an estimate of $Q$. The variance estimator of $\bar{Q}_M$ has an ANOVA style decomposition:

$$T_M = \bar{U}_M + \left(1 + \frac{1}{M}\right) B_M, \tag{2.3}$$

where $\bar{U}_M$ is an estimate of the variance of $\hat{Q}$ if we had the complete data ("within-imputation" variance), and $B_M$ estimates the excess variance due to the missing values ("between-imputation" variance). The factor $(1 + 1/M)$ is a bias adjustment for small $M$, as explained in (Rubin, 1987, Chapter 3.3).

MI was originally derived under Bayesian considerations. The Bayesian derivation of MI begins with the identities

$$P(Q \mid Y_{obs}) = \int P(Q \mid Y_{mis}, Y_{obs}) P(Y_{mis} \mid Y_{obs})\, dY_{mis} \tag{2.4}$$

$$\mathrm{E}(Q \mid Y_{obs}) = \mathrm{E}(\mathrm{E}(Q \mid Y_{mis}, Y_{obs}) \mid Y_{obs}) \tag{2.5}$$

$$\begin{aligned} \mathrm{Var}(Q \mid Y_{obs}) = {} & \mathrm{E}(\mathrm{Var}(Q \mid Y_{mis}, Y_{obs}) \mid Y_{obs}) \\ & + \mathrm{Var}(\mathrm{E}(Q \mid Y_{mis}, Y_{obs}) \mid Y_{obs}) \end{aligned} \tag{2.6}$$

When imputations are generated from $P(Y_{mis} \mid Y_{obs})$, the MI statistics are Monte Carlo estimates of the relevant quantities:

$$\bar{Q}_M \approx \mathrm{E}(\mathrm{E}(Q \mid Y_{mis}, Y_{obs}) \mid Y_{obs}) = \mathrm{E}(Q \mid Y_{obs}) \tag{2.7}$$

$$\bar{U}_M \approx \mathrm{E}(\mathrm{Var}(Q \mid Y_{mis}, Y_{obs}) \mid Y_{obs}), \tag{2.8}$$

$$(1 + 1/M) B_M \approx \mathrm{Var}(\mathrm{E}(Q \mid Y_{mis}, Y_{obs}) \mid Y_{obs}) \tag{2.9}$$

$$T_M \approx \mathrm{Var}(\mathrm{E}(Q \mid Y_{obs})). \tag{2.10}$$

Rubin (1987) proposed constructing confidence intervals for $Q$ based on an asymptotic normal approximation to the posterior distribution (2.4): Taking $M$

to infinity, $(\bar{Q}_\infty - Q) \sim N(0, T_\infty)$ approximately in large samples. In large samples with finite $M$ interval estimation for $Q$ proceeds using a reference $t-$distribution for $\bar{Q}_M$: $(\bar{Q}_M - Q) \sim t_{\nu_M}(0, T_M)$. Rubin (1987) computed an approximate value for $\nu_M$ using a moment matching argument, obtaining $\nu_M = (M-1)\,(1 + 1/r_M)^2$ where $r_M = (1 + 1/M)B_M/\bar{U}_M$ is a measure of the relative increase in variance due to nonresponse. Barnard and Rubin (1999) proposed an alternative degrees of freedom estimate with better behavior in moderate samples, suggesting it for general use. See Reiter and Raghunathan (2007) for a review of combining rules for more general estimands.

## 3. MULTIPLE IMPUTATION: WHEN DOES IT WORK?

In this section we give a high-level review of some of the justifications for using MI and the estimators given above. Special consideration is given to results that can inform the selection of an imputation model.

### 3.1 Bayesian (in)validity under MI

Since the MI estimators were derived under Bayesian arguments we might hope that MI yields valid Bayesian inference. In general it does not. Suppose the analyst has specified a Bayesian model as $P_A(Y, Q) = P_A(Y \mid Q)P_A(Q)$. The analyst's inference is based on the posterior distribution

$$P_A(Q \mid Y_{obs}) = \int P_A(Q \mid Y_{mis}, Y_{obs})P_A(Y_{mis} \mid Y_{obs})\, dY_{mis}. \tag{3.1}$$

Now suppose the imputer has generated imputations according to $Y_{mis}^{(m)} \sim P_I(Y_{mis} \mid Y_{obs})$. On computing $\hat{Q}(Y_{obs}, Y_{mis}^{(m)})$ the analyst has a draw from the hybrid model

$$P_H(Q \mid Y_{obs}) = \int P_A(Q \mid Y_{mis}, Y_{obs})P_I(Y_{mis} \mid Y_{obs})\, dY_{mis} \tag{3.2}$$

If $P_A(Y_{mis} \mid Y_{obs}) = P_I(Y_{mis} \mid Y_{obs})$, then MI delivers the analyst's posterior inference in the sense that $\hat{Q}^{(m)}$ is a draw from (3.1). If the posterior distribution for $Q$ is approximately normal and $M$ is not too small the MI statistics will give a reasonable approximation to the posterior.

However, in practice the imputer and the analyst will likely have different models for $(Y_{mis} \mid Y_{obs})$. Even if one analyst should happen to share the same model as the imputer, the next analyst may have a different set of beliefs encoded in their model, resulting in $P_{A'}(Y_{mis} \mid Y_{obs}) \neq P_A(Y_{mis} \mid Y_{obs})$. In this case the imputer cannot deliver valid Bayesian inference to both analysts with a single set of imputations. Since Bayesian validity is generally unattainable (and good repeated sampling behavior is desirable in its own right), MI is usually evaluated based on its frequentist properties. The remaining subsections explore conditions under which MI yields valid frequentist inference.

### 3.2 Frequentist Validity: Conditions on complete data inference

We will follow Rubin (1996) and assume that the complete data inference is at least *confidence valid*, meaning that a nominal $100(1-\alpha)\%$ confidence interval has actual coverage at least $100(1-\alpha)\%$. (The stronger condition of *randomization validity* requires that the nominal and actual coverage rates agree.) We also

assume that the sampling distribution of $\hat{Q}$ is normal, so that valid confidence intervals can be obtained from $\hat{Q}$ and $\hat{U}$. In this case confidence validity requires that

$$\mathrm{E}(\hat{Q}) = Q \tag{3.3}$$

$$\mathrm{E}(\hat{U}) \geq \mathrm{Var}(\hat{Q}), \tag{3.4}$$

where the expectation and variance are over repeated sampling. Randomization validity obtains when $\mathrm{E}(\hat{U}) = \mathrm{Var}(\hat{Q})$. We depart slightly from Rubin (1996, 1987) in omitting any conditioning on fixed values in a finite population.

In practice normality and (3.3)-(3.4) may only hold asymptotically, or when particular modeling assumptions are correct. Whether this is plausible for a particular analysis will depend on the nature of $\hat{Q}$. For our purposes we will assume that any necessary conditions for confidence validity with completely observed data are satisfied, since our primary consideration is the impact of missingness and imputation. Of course, if the complete data inference is not valid it would be unreasonable to expect MI or any other missing data procedure to remedy the issue.

### 3.3 Proper imputation for valid inference

Chapter 4, Section 4.2 in Rubin (1987) outlines conditions under which MI inferences are randomization or confidence valid when $M = \infty$. Imputations satisfying these conditions for a particular estimand $Q$ and posited response mechanism are known as *proper* imputations. Proper imputation coupled with valid complete data inference yields valid MI inference (Rubin, 1987, Result 4.1). It is important to remember that imputations are only proper with respect to a particular estimand $Q$ and a posited response mechanism.

We focus on three essential conditions necessary for an imputation procedure to be proper for an estimand $Q$. (The other conditions are somewhat technical and generally not the source of improper imputations and invalid inference in practice.)

*3.3.1 Three essential conditions for proper imputation.* Rubin (1996) distilled the formal definition of proper imputation given in (Rubin, 1987, Section 4.2) into three conditions that generally ensure imputations are proper. They concern the behavior of the MI statistics under repeated realizations of the response mechanism, holding the sample values $Y$ fixed (that is, under repeated sampling from $P(R \mid Y)$). The first two conditions require that $\bar{Q}_\infty$ and $\bar{U}_\infty$ be approximately unbiased for $\hat{Q}$ and $\hat{U}$:

$$\mathrm{E}(\bar{Q}_\infty \mid Y) \approx \hat{Q}(Y) \tag{3.5}$$

$$\mathrm{E}(\bar{U}_\infty \mid Y) \approx \hat{U}(Y), \tag{3.6}$$

where the expectations are with respect to $P(R \mid Y)$.

Naturally (3.5)-(3.6) will hold if $P(Y_{mis} \mid Y_{obs})$ is correctly specified by the imputer. However, imputations made under misspecified models can still satisfy (3.5)-(3.6) so long as they broadly capture the features of the predictive distribution that are relevant for computing $Q$ and $U$ and the proportion of missing

data is not extreme. To see this more clearly we can write

$$\mathrm{E}(\bar{Q}_\infty \mid Y) = \sum_{m=1}^{\infty} \mathrm{E}\left(\hat{Q}(Y_{obs}(R), Y_{mis}^{(m)}(R)) \mid Y\right). \tag{3.7}$$

With no missing data the expectations inside the sum are all $\hat{Q}(Y)$. With modest amounts of missing data, the imputed values need to be sufficiently poor to overwhelm the influence of the observed data in computing $Q$. (What constitutes "sufficiently poor" naturally depends on $Q$.) Similar logic applies to $\bar{U}_\infty$.

The third condition for proper imputation is more subtle: It requires that the between-imputation variability $B_\infty$ be approximately unbiased for the variance of $\bar{Q}_\infty$:

$$\mathrm{E}(B_\infty \mid Y) \approx \mathrm{Var}(\bar{Q}_\infty \mid Y). \tag{3.8}$$

Satisfying this condition generally requires that we account for uncertainty *in the imputation model itself* (or equivalently uncertainty in the parameters indexing a model class), since the observed data used to estimate the model, $Y_{obs}(R)$, varies over samples from the response mechanism. (Recall that the variance in (3.8) is with respect to $P(R \mid Y)$.)

Many seemingly reasonable stochastic imputation procedures fail to be proper because they do not satisfy (3.8); these include imputing from a model by plugging in the MLE or drawing imputations from the empirical distribution of observed cases (Rubin, 1987, Ch. 4). Accounting for uncertainty in the imputation model can be achieved (or approximated) in a variety of ways, such as sampling the parameters indexing a particular model class from their posterior under a Bayesian model or through small adjustments to the bootstrap (as described in Section 5.2). See Section 4.1 for further discussion.

### 3.4 Congeniality and confidence validity

It is well-known that the MI estimate $T_\infty$ can be inconsistent for certain choices of $Q$ (Wang and Robins, 1998; Robins and Wang, 2000; Kim, 2002; Nielsen, 2003; Kim et al., 2006). The bias is typically positive and tends to have limited influence on coverage rates for common estimands when the amount of missingness is not extreme (Rubin, 2003a). Rubin (1996) reviewed early examples of inconsistency and gave sufficient conditions for MI inference to be confidence proper (i.e., for $T_\infty$ to conservatively estimate $Var(\bar{Q}_\infty)$); they are similar to the conditions in Section 3.3.1, averaged over repeated sampling of $Y$ in addition to the response mechanism.

Meng (1994) introduced the concept of *congeniality* for understanding the inconsistency of the MI variance estimate. Roughly, an analysis procedure is congenial to an imputation model $P_I(Y_{mis} \mid Y_{obs})$ if we can take the complete data analysis and embed it into a Bayesian model $P_A(Y \mid Q)P_A(Q)$ such that

1. Its posterior $P_A(Q \mid Y)$ recapitulates the desired analysis in the sense that

$$\mathrm{E}_A(Q \mid Y) = \hat{Q}(Y), \quad \mathrm{Var}_A(Q \mid Y) = \hat{U}(Y). \tag{3.9}$$

2. It matches the imputation model, i.e.,

$$P_A(Y_{mis} \mid Y_{obs}) = P_I(Y_{mis} \mid Y_{obs}). \tag{3.10}$$

Under congeniality, MI delivers samples from $P_A(Q \mid Y_{obs})$ (Section 3.1), which we have constructed to yield confidence valid inference. Unless the analyst is the imputer, congeniality is less a condition we should try to satisfy than one we should try to fail gracefully – uncongeniality is generally "the rule not the exception" (Xie and Meng, 2017), for the same reasons discussed in Section 3.1.

Xie and Meng (2017) revisited the behavior of MI inferences under uncongeniality and provided a host of new results. At a high level their findings affirm and generalize common rules of thumb originating with Meng (1994): Even if the "true" model is nested within the imputer's and the analyst's models (e.g., if the imputation model includes both relevant and irrelevant covariates in an otherwise correctly specified regression model for the missing data), standard MI inference may be invalid. However, if the analyst's procedure is *self-efficient* (meaning essentially that their estimator cannot be improved by ignoring relevant data (Meng, 1994; Meng and Romero, 2003)), then:

1. When the imputer's model is more saturated than the analyst's, the usual MI inference is confidence valid and generally robust.
2. When the imputer's model is less saturated than the analyst's, confidence validity is not guaranteed.

It is generally safer to conduct an uncongenial analyses under (1) than under (2), since conservative inferences will obtain. Xie and Meng (2017) also provide remarkably simple and broadly applicable (if somewhat exacting) alternative variance estimates that are valid under uncongeniality: Use $T_M^* = 2T_M$ for a vector $Q$, or sum and square the standard errors for a univariate $Q$: $T_M^* = (\sqrt{U_M} + \sqrt{B_M})^2 + (1/M)B_M$.

Like most strong theoretical results, Xie and Meng (2017)'s results depend on a number of assumptions. One of these assumptions is that the true model ("God's model") is nested within the imputation model class. In his discussion of the paper, Reiter (2017) notes that "[I]n my experience, very low coverage rates in MI confidence intervals arise more often from the imputation procedure generating bias in $[\bar{Q}_\infty]$ than from bias in the MI variance estimator," often due to rote application of default imputation procedures. This has been in part a shared experience (Murray and Reiter (2016)), motivating the focus of this review on the specification of imputation models.

## 4. PRACTICAL IMPLICATIONS OF THEORETICAL RESULTS FOR IMPUTATION MODELING

The theoretical results summarized above suggest a number of practical considerations for generating imputations. These are reviewed below; for more detailed discussion and examples, see e.g. Rubin (1987); Little (1988); Rubin (1996); Van Buuren (2012). Throughout this section and the rest of the paper we will continue to refer to procedures that generate imputations as "imputation models", regardless of whether they are completely specified probability models.

### 4.1 Imputations should reflect uncertainty about missing values and about the imputation model.

The goal in multiple imputation is to account for uncertainty due to the missing values in subsequent inference. This is a different objective than estimating

or predicting the missing values, which could generally be achieved via simpler means. The situation in MI is similar to the more familiar task of constructing valid predictive intervals with a regression model, where we need to account for uncertainty in the unobserved response as well as uncertainty in the regression fit.

Suppose we have a single variable subject to missingness, to be imputed using a regression model. If we were only concerned with reconstructing the missing values, we would just impute the fitted values. This would clearly lead to invalid MI inferences. Instead, MI propagates the intrinsic uncertainty about the missing values via some stochastic mechanism, for example, by adding a randomly generated residual to the regression prediction. However, to achieve at least approximately proper imputations we also need to account for uncertainty about the imputation model itself – that is, uncertainty in the fitted values of the regression model. Methods that do not appropriately reflect both sources of uncertainty tend to violate (3.8) and underestimate the between-imputation variance, yielding standard errors that are too small and anti-conservative inferences (Rubin, 1987, 1996).

Bayesian imputation procedures provide a natural mechanism to account for model uncertainty. Imputations are generated from

$$P(Y_{mis} \mid Y_{obs}) = \int P(Y_{mis} \mid \theta, Y_{obs}) P(\theta \mid Y_{obs}) d\theta. \tag{4.1}$$

where $\theta$ is a parameter indexing a model for $Y$ (or a model for $Y_{mis}$ given $Y_{obs}$). To see how model uncertainty propagates, observe that imputations can be sampled compositionally: For $1 \leq m \leq M$, first draw a value $\theta^{(m)} \sim P(\theta \mid Y_{obs})$ and then sample $Y_{mis}^{(m)} \sim P(Y_{mis} \mid \theta^{(m)}, Y_{obs})$. Model uncertainty is represented by $P(\theta \mid Y_{obs})$, and the intrinsic uncertainty about the missing values is represented by $P(Y_{mis} \mid \theta, Y_{obs})$. Approximations to full Bayesian inference have also proven useful: Rubin and Schenker (1986)'s approximate Bayesian bootstrap for proper hot deck imputation is one early example (Section 5.2). Chapter 10 of Little and Rubin (2002) reviews several others.

Of course, Bayesian modeling is not magic – if $\theta$ indexes a class of misspecified models then we should expect our imputations and inferences to suffer, at least for estimands that are sensitive to this misspecification. For example, when $Y_{mis}$ contains variables with significant skew a multivariate normal imputation model would likely yield approximately valid inference for marginal means but invalid inference for some marginal quantiles, since (3.5) can be violated when $Q$ is an extreme quantile.

From a coverage perspective, model misspecification becomes increasingly consequential in large samples where the complete data standard errors are small and $P(\theta \mid Y_{obs})$ will tend to concentrate on the parameters of the "best" misspecified model. Even small biases due to misspecification in the imputation model can become large relative to the pooled standard errors. Enlarging the imputation model class $P(Y \mid \theta)$ via non- and semiparametric Bayesian modeling can guard against misspecification and also mitigate the artificial certainty implied by fixing a regular parametric model and only considering uncertainty in its parameters. Section 6.1.1 explores recent promising developments in this area.

### 4.2 Imputation models should generally include as many variables as possible.

There are multiple reasons for entertaining the largest possible imputation model: The missing at random assumption tends to be more tenable as more completely-observed variables are added to the imputation model. In addition, if variables predictive of the missing values are left out of the imputation model but used to compute $Q$ or $U$, then the imputations will be improper – the imputed values will be incorrectly independent of the omitted variables, leading to bias over repeated imputations (violations of (3.5) or (3.6)) (Rubin, 1996). In this case the analysis and imputation models are uncongenial in the "wrong" way – the imputer's model is less-saturated than the analysis model. In sum, the cost of excluding a relevant variable (invalid inference) is often greater than the cost of including an irrelevant variable (roughly, additional variance). This is particularly relevant when the analyst and imputer are not the same, and the imputations must support many unspecified analyses. Even when the imputer and the analyst are the same it would be useful to generate one set of imputations that can support the usual process of iterative model building and refinement, rather than generating a new set of imputations for each analysis model that is considered. See Collins, Schafer and Kam (2001) and Schafer (2003) for further discussion of the tradeoffs involved.

These points are particularly relevant for design variables in complex surveys. Design-based estimators will typically use stratum and cluster information to compute $U$. Reiter, Raghunathan and Kinney (2006) show empirically that failing to account for an informative sampling design can lead to invalid inference. They suggest including indicator variables for strata and cluster membership in the imputation model, or including stratum fixed effects and cluster random effects in imputation models. It may be useful to include estimated response propensities or final adjusted survey weights (sampling weights with e.g. calibration and post-stratification adjustments) as well, especially if complete design information is not available to the imputer (Rubin, 1996).

### 4.3 Imputation models should be as flexible as possible.

Finally, imputation models should try to "track the data" (Rubin, 1996) by modeling relevant features of the joint distribution of the missing values. Loosely, a feature of the joint distribution is relevant if it is a possible target of inference itself, or more generally if it yields a more accurate predictive distribution for the missing data. Interactions, nonlinearities, and non-standard distributional forms are all potentially relevant features.

As Meng (1994) succinctly put it, "Sensible imputation models should not only use all available information to increase predictive power, but should also be as general and objective as practical in order to accommodate a potentially large number of different data analyses." We would add that where possible, imputation models should have some capacity to *adapt* to unanticipated features of the data (such as interactions, nonlinearities, and complex distributions), especially when the imputer has limited time and resources to spend on iteratively improving the imputation model.

## 5. GENERATING IMPUTATIONS FOR A SINGLE VARIABLE

We begin by cataloging some of the more common approaches to generating imputations for a single variable subject to missingness, conditional on other fully observed variables. In the next section we consider how these can be extended to generate imputations for several variables.

### 5.1 Regression Modeling

Imputation by sampling from univariate regression models is conceptually straightforward. Generalized linear models and extensions to deal with complications such as zero-inflation and truncation are popular options; these are not reviewed in depth here but see e.g. Van Buuren and Oudshoorn (1999), Raghunathan et al. (2001), Su et al. (2011), or Van Buuren (2012) (Chapter 3). These methods are quite common in practice, but since most readers will be familiar and they are well-reviewed elsewhere we will not enumerate them here.

To generate proper imputations some method should be used to account for parameter uncertainty – simple strategies like sampling from the regression model with parameters fixed at the observed data MLE are generally improper. Posterior sampling under a non- or weakly informative prior tends to be proper when the model fits well. Prior distributions can also ease problems like separation in logistic regression and apply helpful regularization in conditional models with many variables in the conditioning set (Su et al., 2011).

### 5.2 Hot Deck/Nearest Neighbor Methods.

The hot deck and other nearest-neighbor methods (Chen and Shao, 2000; Andridge and Little, 2010) begin by defining a distance metric between cases in terms of the observed covariates. Imputations for a missing value are borrowed from a nearby completely observed case (the "donor"). These methods tend to be simpler to implement than fully specified regression models and often make fewer assumptions. However, these methods are far from assumption free – the choice of distance metric, the definition of the donor pool, and how to sample from the donor pool all influence the quality of imputations.

The hot deck (Andridge and Little, 2010) defines distance via cross-classifications of fully observed variables which determine adjustment cells. Missing values are imputed by sampling with replacement from the pool of donors within the same cell. This strategy ensures that all imputations are plausible values, which is an appealing feature relative to regression imputation. Complications arise when there are many fully observed variables to incorporate into the cross-classification or when the sample size is low, leading to many small or empty adjustment cells.

MI with the hot deck is also known to be improper for simple estimands like a population mean (Rubin and Schenker, 1986). The hot deck effectively assumes that the distribution of missing values within an adjustment cell is *exactly* the empirical distribution of the observed values within that cell, which leads to $B$ having downward bias (due to ignoring uncertainty in the implicit imputation model). Rubin and Schenker (1986) propose a simple modification that makes the hot deck proper, based on an approximation to the Bayesian bootstrap (Rubin, 1981). Instead of sampling the $n_m$ missing values from the empirical distribution of the $n_o$ observed values within an adjustment cell, the approximate Bayesian bootstrap (ABB) first samples a set of $n_o$ values with replacement from the

observed data and then samples $n_m$ imputed values with replacement from this set. This simple adjustment yields proper imputations for the population mean of the adjustment cell (Rubin and Schenker, 1986). (See also Kim (2002) for a more accurate variance estimate in small samples.)

Predictive mean matching (PMM) (Little, 1988) instead measures the distance between cases by the distance between their predicted means for the variable subject to missingness (traditionally estimated using a linear regression, although in principle any method could be used to make the prediction). PMM generalizes the hot deck, which is a special case of PMM using saturated models with categorical predictors. By avoiding the discretization and making some assumptions about the relationships between the predictors and the response (such as linearity) PMM can handle more variables than the hot deck, but may be sensitive to the predictive model specification.

To define the donor pool Heitjan and Little (1991) proposed sampling from a window of $k$ nearby potential donors in PMM in the hope of making the method approximately proper. The donor's value may be imputed, or its residual can be added to the predicted mean of the missing value to generate an imputation. Schenker and Taylor (1996) found these two approaches to perform similarly in simulations; the former will always impute a previously realized value, which may be desirable. See Vink et al. (2014) for an approach to semi-continuous variables. Morris, White and Royston (2014) compared newer developments and current implementations of these techniques, cautioning in particular against the imputation of a single nearest neighbor (which appears to be common in software implementations of PMM) as it is improper.

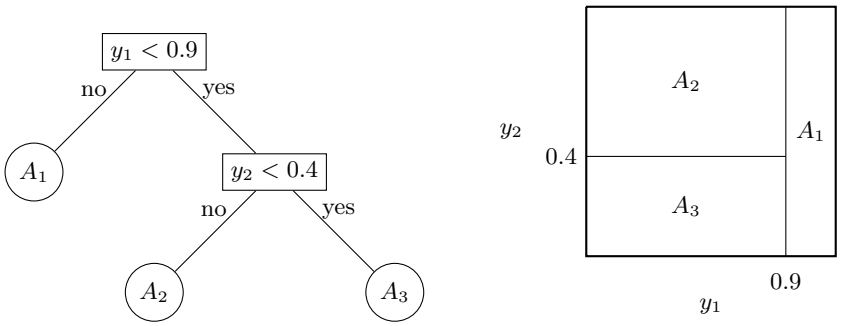


FIG 1. *(Left) An example CART tree, with internal nodes labeled by their splitting rules and terminal nodes given labels $A_h$. (Right) The corresponding partition of $(Y_1, Y_2)$.*

PMM and the hot deck can be made more adaptive using recursive partitioning. Reiter (2005) and Burgette and Reiter (2010) proposed imputation via classification and regression trees (CART, Breiman et al. (1984)). A tree is grown using fully observed data to predict the variable subject to missingness. Then each incomplete case is assigned to its corresponding leaf, and an imputation is sampled from donors within in the same leaf. The imputer can control the size of the donor pool by growing the tree down to a specified minimum leaf size. This is a special case of PMM using CART to generate predictions; we could also think of it as an adaptive hot deck that leverages the most predictive variables and balances the size of the adjustment cells. Figure 1 shows an example tree grown on two variables $(Y_1, Y_2)$ to impute a third $(Y_3)$, along with the corresponding partition

which forms the adjustment cells.

Reiter (2005) and Burgette and Reiter (2010) drew ABB samples from within the leaves in an effort to generate proper imputations. Van Buuren (2012) (Algorithm 3.6) suggested also accounting for uncertainty in the tree itself by growing it on a different bootstrap sample for each imputed dataset. Doove, Van Buuren and Dusseldorp (2014) proposed imputation by growing a random forest (an ensemble of trees) (Breiman, 2001) of size $k$ by bootstrapping the complete cases and (optionally) sub-sampling the variables, as in traditional applications of random forests. An imputed value is generated by sampling from the $k$ trees and then following the procedure to generate a CART imputation. Shah et al. (2014) proposed fitting a random forest, estimating its predictive error variance, and generating imputations as the random forest prediction plus a normally distributed residual.

Limited results exist comparing these different recursive partitioning methods, and there is similarly limited guidance as to how they should be tuned. But they can be fast and effective imputation engines, particularly for large sets of categorical variables that take a relatively limited set of levels (see e.g. Akande, Li and Reiter (2017)).

## 6. GENERATING IMPUTATIONS FOR MULTIPLE VARIABLES

There are two basic strategies for imputing multivariate missing data: Jointly modeling the variables subject to missingness, or specifying a collection of univariate conditional imputation models that condition on all the other variables (this approach goes under various names including sequential regression multivariate imputation (Raghunathan et al., 2001) and multiple imputation by chained equations (Van Buuren and Oudshoorn, 1999), but we will use "fully conditional specification" (FCS) as in Van Buuren et al. (2006)). Joint models can be further classified into "simultaneous" approaches that define a multivariate distribution $f(Y)$ directly or "sequential" approaches that build up a multivariate distribution using a ladder of conditional distributions, where the model for each variable conditions only on those earlier in the sequence. Appendix A has pointers to software implementations of many methods described in this section.

To describe the different approaches we need some new notation: Let $Y_{j,obs}$ and $Y_{j,mis}$ denote the set of observed and missing values for the $j^{th}$ variable. Let $Y_{imp}$ denote an imputed dataset, and $Y_{j,imp}$ denote a set of imputations for $Y_{j,mis}$. We will use the subscript $(-j)$ to denote the same quantities for all but the $j^{th}$ variable.

### 6.1 Joint specification: Simultaneous approaches

Early simultaneous joint modeling approaches were based on the multivariate normal (MVN) or $t$ distribution; these are reviewed in Schafer (1997) and Little and Rubin (2002). For high dimensional continuous observations low-rank structure can be imposed on the covariance matrix (Audigier, Husson and Josse, 2016). Various authors have proposed imputing categorical data under a misspecified MVN model, either leaving the continuous imputations for discrete variables as-is or rounding them based on some thresholds (Horton, Lipsitz and Parzen, 2003; Bernaards, Belin and Schafer, 2007). This is naturally more complicated when the discrete variables are not ordinal, particularly if they take many levels.

Additionally, end users may not trust imputations from a data disseminator if the imputed data appear invalid. Therefore it is often preferable to use models that are appropriate for the types of variables at hand.

For small numbers of strictly discrete variables a simple multinomial model may be feasible. However, with a large number of discrete variables it is impossible to fit saturated multinomial models and further restrictions are necessary. Options include log-linear models (Schafer, 1997), latent class models (Vermunt et al., 2008; Gebregziabher and DeSantis, 2010; Vidotto, Vermunt and Kaptein, 2015), or multiple correspondence analysis (Audigier, Husson and Josse, 2017) (which is closely related to a certain class of multivariate logit models (Fithian and Josse, 2017)).

Joint models for mixed continuous and categorical data are also available. For the remainder of Section 6.1, suppose we have collected the continuous variables into a vector $Y$ and the discrete variables into another vector $X$. The general location model (GLOM) (Olkin and Tate, 1961; Little and Schluchter, 1985; Schafer, 1997) assumes that $(Y \mid X = x) \sim N(\mu_x, \Sigma_x)$ and $X \sim \pi$. (Liu and Rubin (1998) generalized the $(Y \mid X)$ model to the larger class of elliptically symmetric distributions.) The number of parameters in this saturated model grows rapidly with the sample space of $X$, so imputers typically impose further constraints. Examples include common covariance structure ($\Sigma_x \equiv \Sigma$ for all $x$), removing higher-order effects from the conditional means by specifying $\mu_x = D(x)B$ for a matrix of regression coefficients $B$ and design vector $D(x)$, and imposing log-linear constraints on $\pi$ to rule our higher-order interactions in the marginal model for $X$.

*6.1.1 Mixtures and Nonparametric Bayesian Models.* Even without additional parameter constraints, most parametric joint models make restrictive assumptions. Mixture models provide a simple and expressive way to enrich a parametric model class. For example, latent class models for categorical data are mixtures of independence models (log-linear models with only main effects) which have proven useful in multiple imputation (e.g., Vermunt et al., 2008; Gebregziabher and DeSantis, 2010). Mixtures of multivariate normal distributions can model complex features of joint continuous distributions (Böhning et al., 2007; Elliott and Stettler, 2007).

Several Bayesian nonparametric models have recently been proposed for multiple imputation. Most of these are based on infinite mixture models or their truncated approximations (but see Paddock (2002) for an early exception based on Polya trees, and also the sequential regression approach in Xu, Daniels and Winterstein (2016)). Relative to parametric Bayesian approaches these models are appealing for their ability to grow in complexity with increasing sample size. Under some circumstances this can allow the model to capture unanticipated structure like interactions and nonlinear relationships or nonstandard distributions, reflecting these in the imputed values.

Recall that we have separated the data into vectors of categorical variables $X$ and continuous variables $Y$. For imputing multivariate categorical data, Si and Reiter (2013) adopt a truncated version of the Dirichlet process mixture of product multinomials (DP-MPMN) proposed by Dunson and Xing (2009). This is a latent class model with a large number of classes (say $k^{\mathcal{X}}$) and a particular prior over the class distribution.

Suppose the $j^{th}$ categorical variable takes (possibly unordered) values indexed by $1, 2, \ldots, d_j$ and let $H_i^{\mathcal{X}} \in \{1, \ldots, k^{\mathcal{X}}\}$ be a latent mixture component index for observation $i$. Let $\Pr(X_{ij} = x_{ij} \mid H_i^{\mathcal{X}} = s) = \psi_{sx_{ij}}^{(j)}$. The DP-MPMN model assumes that

$$\Pr(H_i^{\mathcal{X}} = s) = \phi_s^{\mathcal{X}} \tag{6.1}$$

$$\Pr(X_i = x_i \mid H_i^{\mathcal{X}} = s, \Psi) = \prod_{j=1}^{p} \psi_{sx_{ij}}^{(j)}, \tag{6.2}$$

so that the elements of $X$ are conditionally independent given the latent class membership. The prior on $\phi$ is a truncated version of the stick-breaking construction for the Dirichlet process (DP) (Sethuraman, 1994), introduced in Ishwaran and James (2001) to simplify Gibbs sampling in DP mixture models:

$$\phi_s^{\mathcal{X}} = \xi_s \prod_{l<s}(1 - \xi_l), \quad \{\xi_s\}_{s=1}^{k^{\mathcal{X}}} \overset{iid}{\sim} Beta(1, \alpha), \quad \xi_{k^{\mathcal{X}}} \equiv 1. \tag{6.3}$$

The model is completed with prior distributions on $\Psi$ and $\alpha$ (see Si and Reiter (2013) for a complete specification). Manrique-Vallier and Reiter (2014a,b) extended this model to assign zero probability to impossible values of $X$, such as cells that are logically impossible (pregnant men or children collecting retirement benefits) or necessarily empty due to skip patterns. Manrique-Vallier and Reiter (2016) introduced a variant of this model for edit-imputation that simultaneously accounts for missing values and observed values that are logically impossible but present due to measurement error. Hu, Reiter and Wang (2017) extended this model to nested data structures (i.e., hierarchical structures like individuals nested within households) in the presence of structural zeros.

For imputing continuous data Kim et al. (2014) suggested a truncated DP mixture of multivariate normal distributions. Let $H_i^{\mathcal{Y}}$ be the mixture component index for record $i$. This model assumes that

$$\Pr(H_i^{\mathcal{Y}} = r) = \phi_r^{\mathcal{Y}} \tag{6.4}$$

$$\left(Y_i \mid H_i^{\mathcal{Y}} = r, -\right) \sim N(\mu_r, \Sigma_r), \tag{6.5}$$

with a prior on $\phi_r^{\mathcal{Y}}$ defined via a stick-breaking process similar to (6.3). Kim et al. (2014) modified the model in (6.5) to constrain the support of $Y$ to a set $\mathcal{A}$ with bounds determined by a set of linear inequalities, so that $\Pr(Y \notin \mathcal{A}) = 0$ under the prior. Kim et al. (2015) extended this approach to simultaneous edit-imputation, generating imputed values for observations outside of $\mathcal{A}$ via a measurement error model.

Murray and Reiter (2016) built a hierarchical mixture model for mixed continuous and categorical observations by combining the models in (6.1)-(6.2) and (6.4)-(6.5), with two important adjustments. First, (6.5) is modified to include a regression on $X$ with component-specific coefficients:

$$(Y_i \mid X_i = x_i, H_i^{\mathcal{Y}} = r, -) \sim N(D(x_i)B_r, \Sigma_r). \tag{6.6}$$

By default the design matrix $D(x_i)$ encodes main effects. Allowing the component means to depend on $X$ greatly reduces the number of mixture components

necessary to capture $X-Y$ relationships. Second, the mixture component indices in each model are given a hierarchical prior introduced by Banerjee, Murray and Dunson (2013):

$$\Pr(H_i^{\mathcal{X}} = s, H_i^{\mathcal{Y}} = r \mid Z_i = z) = \phi_{zs}^{\mathcal{X}}\phi_{zr}^{\mathcal{Y}} \tag{6.7}$$

$$\Pr(Z_i = z) = \lambda_z, \tag{6.8}$$

Here $\lambda_z$ is assigned a stick-breaking prior, Each pair $\phi_z^{\mathcal{X}} = \left(\phi_{z1}^{\mathcal{X}}, \ldots, \phi_{zk^{\mathcal{X}}}^{\mathcal{X}}\right)'$ and $\phi_z^{\mathcal{Y}} = \left(\phi_{z1}^{\mathcal{Y}}, \ldots, \phi_{zk^{\mathcal{Y}}}^{\mathcal{Y}}\right)'$ are probability vectors also assigned independent truncated stick breaking priors. This is a "mixture of mixtures" model; marginalizing over the latent variables the joint density is

$$f(X_i, Y_i) = \sum_{z=1}^{k^{\mathcal{Z}}} \lambda_z \left(\sum_{r=1}^{k^{\mathcal{Y}}} \phi_{zr}^{\mathcal{Y}} N(Y_i; D(X_i)B_r, \Sigma_r)\right) \left(\sum_{s=1}^{k^{\mathcal{X}}} \phi_{zs}^{\mathcal{X}} \prod_{j=1}^{p} \psi_{sX_{ij}}^{(j)}\right). \tag{6.9}$$

Each mixture component is itself composed of two mixture models, one for $(Y \mid X)$ and one for $X$. These lower-level mixtures share some parameters ($B, \Sigma$, and $\Psi$), enforcing a degree of parsimony.

DeYoreo, Reiter and Hillygus (2016) used a similar hierarchical mixture model constructed based on different considerations, splitting the variables into sets based on their type (ordinal or nominal) and high or low rates of missing values. An expressive model class is specified for the variables with high rates of missing values, and a simpler model class is utilized for variables with low rates of missingness. Ordinal variables are explicitly modeled as such by thresholding mixtures similar to (6.6).

Further extensions, combinations, and enhancements of these models are possible. Despite their complexity, all of these models have been shown to perform well for MI with real, complicated data and little or no tuning.

## 6.2 Fully Conditional Specification

FCS avoids explicit joint probability models by specifying a collection of univariate conditional imputation models instead (Van Buuren and Oudshoorn, 1999; Raghunathan et al., 2001). Each univariate model typically conditions on all the remaining variables. In FCS the missing values are imputed by iteratively sampling from these conditional models:

1. Begin by filling in $Y_{mis}$ with plausible values to generate an initial completed dataset, stored in $Y_{imp}$
2. For $1 \leq j \leq p$, use a univariate imputation method to sample new imputed values for $Y_{j,mis}$ from a distribution $P(Y_{j,mis} \mid Y_{j,obs}, Y_{(-j),imp})$, and store them in $Y_{j,imp}$.
3. Iterate the previous step until apparent convergence and return the final value of $Y_{imp}$

This process is repeated $M$ times, saving the returned value as one of the $M$ imputations. Any of the univariate imputation methods in the previous section could be used. This lends FCS some flexibility relative to the joint-simultaneous approaches described above.

But this flexibility comes at a cost: Even if each $g_j$ is a completely specified probability model, taken together they often do not correspond to a proper joint

distribution for $Y$ (Arnold and Press, 1989; Arnold, Castillo and Sarabia, 2001). A set of full conditional distributions that do not correspond to any joint distribution is said to be *incompatible*. Simple adjustments like adding polynomial terms or interactions to univariate regression models can induce incompatibility (Liu et al., 2014).

While the algorithm above looks like a standard Gibbs sampler, if the conditional models are incompatible the behavior of the FCS imputation algorithm is unclear: The imputations from the FCS algorithm given above may converge to a unique limiting distribution, or fail to converge to any unique limiting distribution, or converge to different distributions depending on the initial values and/or order of the updates. Li, Yu and Rubin (2012) give examples of incompatible FCS models with fixed parameters whose imputations either diverge or converge to different stationary distributions depending on the order of their updates. This phenomenon seems to be rare in real data, and Zhu and Raghunathan (2015) note that estimating rather than fixing parameters ameliorates at least some of the problems in Li, Yu and Rubin (2012)'s examples.

There are some limited convergence results available when the fully conditional specification comprises univariate Bayesian regression models. Liu et al. (2014) study an iterative FCS imputation procedure that uses a set of Bayesian regression models $g_j(Y_{ij} \mid, Y_{(-j)}, \theta_j)$ with prior distributions $\pi_j(\theta_j)$. With a slight abuse of notation, define

$$g_j(Y_{j,obs} \mid Y_{(-j),imp}, \theta_j) = \prod_{i=1}^{n} g_j(Y_{ij} \mid Y_{j,imp}, \theta_j)^{R_{ij}} \tag{6.10}$$

$$g_j(Y_{j,imp} \mid Y_{j,obs}, Y_{(-j),imp}, \theta_j) = \prod_{i=1}^{n} g_j(Y_{ij} \mid Y_{(-j),imp}, \theta_j)^{1-R_{ij}}. \tag{6.11}$$

Algorithm 1 gives one iteration of an iterative FCS sampler under these models.

**Algorithm 1** Iterative FCS Sampler from Liu et al. (2014)

For $1 \leq j \leq p$,

1. Sample $\theta_j \sim \pi_j(\theta_j \mid Y_{j,obs}, Y_{(-j),imp}) \propto g_j(Y_{j,obs} \mid Y_{(-j),imp}, \theta_j)\pi_j(\theta_j)$
2. Sample $Y_{j,imp} \sim g_j(Y_{j,imp} \mid Y_{j,obs}, Y_{(-j),imp}, \theta_j)$

We can compare this approach to a proper MCMC algorithm under a joint model. Specifically we consider a collapsed Gibbs sampler (Liu, 1994) that targets $P(Y_{mis} \mid Y_{obs}) = \int P(Y_{mis}, \theta \mid Y_{obs})d\theta$ directly, by jointly sampling $(Y_{j,mis}, \theta \mid Y_{j,obs}, Y_{(-j),imp})$ at each step. It is impractical to use directly, but it is helpful to make comparisons with Algorithm 1.

Let the joint model be given by $f(Y_i \mid \theta)$, with full conditionals $f_j(Y_{ij} \mid Y_{(-j)}, \theta)$ and joint prior distribution $\pi(\theta)$ (where $\theta = (\theta_1, \theta_2, \ldots, \theta_p)$). Define $f_j(Y_{j,obs} \mid Y_{(-j),imp}, \theta)$ and $f_j(Y_{j,imp} \mid Y_{j,obs}, Y_{(-j),imp}, \theta)$ as in equations (6.10)-(6.11). Algorithm 2 gives one iteration of the collapsed Gibbs sampler.

Under some regularity conditions the two algorithms are equivalent in finite samples if we can write $\pi(\theta) = \pi_j(\theta_j)\pi_{(-j)}(\theta_1, \theta_2, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_p)$ for any $j$ and the set of $g_j$'s are compatible and correspond to the full conditionals of $f$ (Hughes et al., 2014). This is sufficient to ensure that the conditional distributions in both steps of each algorithm agree.

**Algorithm 2** Collapsed Gibbs Sampler for a Joint Model

For $1 \leq j \leq p$,

1. Sample $\theta \sim \pi(\theta \mid Y_{j,obs}, Y_{(-j),imp}) \propto f_j(Y_{j,obs} \mid Y_{(-j),imp}, \theta)\pi(\theta)$
2. Sample $Y_{j,imp} \sim f(Y_{j,mis} \mid, Y_{j,obs}, Y_{(-j),imp}, \theta)$

If $\pi(\theta) \neq \pi_j(\theta_j)\pi_{(-j)}(\theta_1, \theta_2, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_p)$ for some $j$ but the conditional models are compatible and correspond to the full conditionals of $f$, the two algorithms agree as $n \to \infty$ provided the FCS algorithm has a unique stationary distribution (Liu et al., 2014). Intuitively, in this case the data in $Y^{(-j)}$ influence $\theta_j$ indirectly through the other parameters, but the FCS algorithm ignores this information. Asymptotically the priors become irrelevant in regular parametric models, but in finite samples inference based on the FCS imputations may be inefficient in this regime (Seaman and Hughes, 2016).

Finally, Liu et al. (2014) show that if the FCS algorithm uses an inconsistent set of models but has a unique stationary distribution then MI estimates computed using imputations from Algorithm 1 are consistent provided that the following conditions hold:

1. The collection of conditional models are incompatible, but become compatible with a joint model $f$ after constraining $\theta$.
2. The model class defined by $f$ contains the true distribution that generated the data.

These are rather restrictive; verifying a unique stationary distribution is challenging, as is checking condition 1 above. It also seems unlikely that condition 2 will hold exactly for the simple parametric models in common use. Zhu and Raghunathan (2015) provide some further convergence results for FCS algorithms where each observation is missing at most one value, but without assuming a unique stationary distribution for the FCS chain.

### 6.3 Joint specifications: Sequential approach

Sequential approaches to imputation modeling fix a permutation of $1, 2, \ldots, p$ and build up a joint distribution from a series of univariate models. For example, if the variables are already in the desired order we would have

$$f(Y) = f_1(Y_1)f_2(Y_2 \mid Y_1)f_3(Y_3 \mid Y_2, Y_1) \ldots f_p(Y_p \mid Y_{p-1}, \ldots, Y_1). \tag{6.12}$$

Examples of this approach include (Lipsitz and Ibrahim, 1996; Ibrahim, Lipsitz and Chen, 1999; Ibrahim et al., 2005; Lee and Mitra, 2016; Xu, Daniels and Winterstein, 2016), among others.

Provided that each $f_j$ is a proper univariate probability model, a sequential specification always defines a coherent joint model, unlike FCS approaches. However, different orderings will generally lead to different joint distributions and potentially different fits. Heuristics have been proposed for selecting the order, for example ordering variables by their types (e.g. Ibrahim, Lipsitz and Chen (1999)) or percentage of missing values (e.g., Rubin and Schafer (1990)). The latter is particularly well-motivated when the missing data are monotone (when there is an ordering such that $R_{ij} = 0 \Rightarrow R_{ij'} = 0$ for $j' > j$. ). If the missing data are not exactly monotone one can identify a permutation that is nearly

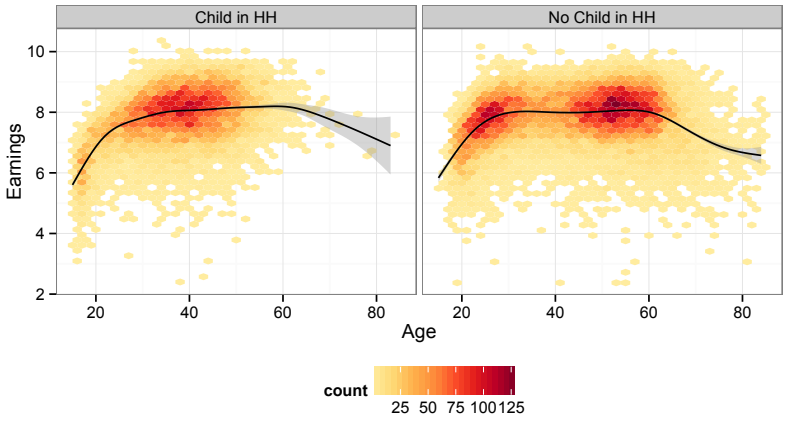


FIG 2. *Joint distribution of householder age and log total earnings, stratified on whether the household includes one of the householder's own children, using the population Murray and Reiter (2016) constructed from complete cases in the first wave of the Survey of Income and Program Participation's 2008 panel.*

monotone and use FCS or delete observed values to "monotonize" the missing data pattern, so that proper sequential techniques can be used for the majority of missing values (as in Rubin (2003b) and extended in Li et al. (2014)).

Another consideration in joint-sequential modeling is that variables early in the sequence may have complex distributions because they are marginalized over many related covariates. For example, Figure 2 shows the joint distribution of householder earnings and age, conditional on whether the householder has any children living in the same household (the data are from complete cases in wave one of the Survey of Income and Program Participation's 2008 panel). The distributions are quite complicated, and it would be difficult to capture them well with simple parametric regression models in any order.

## 7. CHOOSING AND ASSESSING AN IMPUTATION STRATEGY

### 7.1 Comparing FCS and Joint approaches

FCS and joint approaches have competing strengths. FCS models are relatively simple to implement and widely available in software, especially compared to joint-sequential approaches. Joint-simultaneous models including the multivariate normal, log-linear models, and the GLOM are also easy to set up and widely available, but inflexible in practice even relative to simple FCS procedures (e.g. Van Buuren (2007); Stuart et al. (2009); He et al. (2010); Drechsler (2010); Kropko et al. (2014)).

More sophisticated joint models can be challenging to implement, although this is changing – many of the nonparametric Bayesian methods have publicly available implementations (Appendix A). However, even with a good implementation the nonparametric Bayesian models are generally more computationally expensive than simpler joint models (especially those based on low-rank methods, e.g. Audigier, Husson and Josse (2016, 2017)) or FCS methods. Joint-sequential approaches currently take more effort to set up, but they inherit many of the positive features of FCS and joint-simultaneous approaches (univariate models

that are readily assessed and modified but also consistent with joint models).

The convergence properties of FCS in general settings is still mostly an open question. The behavior of FCS algorithms under non- or quasi-Bayesian imputation procedures like PMM is entirely an open question. While the lack of a coherent joint distribution does undermine the theoretical justifications for MI inference detailed in Rubin (1987), experience with FCS in simulations and real applications does not seem to suggest that either lack of convergence or compatibility with a joint model are necessarily overriding concerns.

In fact, under the current theoretical results ensuring that the imputations generated by FCS converge to the imputations under a proper joint model requires using restrictive (implicit) joint models and there is strong empirical evidence that these joint models can be too simple to perform well with realistic data (e.g. Murray and Reiter (2016); Akande, Li and Reiter (2017)). Therefore at this point it would probably be a mistake to choose the models in an FCS imputation routine to try to ensure convergence; it seems much more important to use flexible, adaptive imputation models wherever possible, whether using a joint or FCS imputation strategy.

Imputers who do choose to use FCS should use flexible univariate models wherever possible and take care to assess apparent convergence of the algorithm, for example by computing traces of pooled estimates or other statistics and using standard MCMC diagnostics (Gelman et al., 2013, Chpater 11). It may also be helpful to examine the results of many independent runs of the algorithm with different initializations and to use random scans over the $p$ variables to try to identify any convergence issues and mitigate possible order dependence.

### 7.2 Practical considerations derived from MI theory

We can also compare methods on the practical considerations derived from theoretical results as summarized in Section 4:

*7.2.1 Accounting for uncertainty.* Most of the methods reviewed above include some mechanism for reflecting imputation model uncertainty. Bayesian or approximately Bayesian methods (including the approximate Bayesian bootstrap) do this naturally, whether part of a joint modeling or FCS imputation routine. Their behavior is not well understood in the FCS setting, however. Tree-based methods seem promising for some applications, but more work is required to find parameter settings and resampling strategies that make them reliably proper.

*7.2.2 Include as many variables as possible.* Joint-sequential models may be easier to fit than FCS with many covariates, since all but one univariate model will include fewer than $p$ predictors. Simultaneous joint models somewhat lag behind sequential and FCS approaches here. This is particularly true with mixed data types and many fully observed covariates - most of these models are not easily adapted to condition on additional covariates, so fully observed variables must be included as additional variables in the joint model. Modeling fully observed variables instead of conditioning on them can waste "degrees of freedom" and lead to poorer model fit for the conditional distribution of the missing data. Carefully constructed models can help (DeYoreo, Reiter and Hillygus, 2016), but seem to only go so far.

*7.2.3 Use flexible imputation models.* Non- and semiparametric methods (Bayesian and otherwise, such as sequential tree-based methods) are flexible in their ability to capture certain unanticipated features of the data. Empirically these methods can outperform existing default MI procedures in simulations, particularly when the simulations are not built around simple parametric models themselves. More of these realistic evaluations are needed, as discussed in Sections 7.3 and 8.

However, with flexible imputation models it can be challenging to manually adjust the imputation model to incorporate prior information or address model misfit. Incorporating meaningful prior information into nonparametric Bayesian imputation models is challenging but not impossible; see e.g. Schifeling and Reiter (2016) for a strategy to include prior information in DP-MPMN models. While iterative imputation model refinement and assessment is ideal, it is not always possible. Empirical evidence suggests that flexible imputation models are much better as defaults than simple parametric models or PMM using linear models.

### 7.3 Empirical comparisons between methods

Empirical comparisons of several different imputation models on realistic datasets are relatively rare. Most papers introducing a new imputation model evaluate it using synthetic data generated from a researcher-specified multivariate probability model. The new imputation model is typically compared to a small number of competitors. These simulation studies can be informative – for example, both Burgette and Reiter (2010) and Doove, Van Buuren and Dusseldorp (2014) found evidence that imputations for continuous values generated via recursive partitioning can preserve interactions but underestimate main effects. However, models that are easy to simulate from and present in a paper will naturally be gross simplifications of the distribution of data in real populations.

Simulations based on repeated sampling from realistic populations can be more informative. In these studies a population is compiled from existing data. Random samples are taken from these populations and values are "blanked out" via a known stochastic nonresponse mechanism. Each of the resulting incomplete datasets are multiply imputed and used to compute a range of estimates and confidence intervals, assessing the bias, coverage and efficiency of the MI estimates under the imputation model. Since the missing values are known, these can all be compared against the frequentist operating characteristics of the complete data procedure without appeal to asymptotic theory or other approximations. While the results are specific to a particular population and a set of estimands, this framework is much closer to reality than fully synthetic examples.

There are several recent examples of this kind of evaluation: Akande, Li and Reiter (2017) compared FCS with CART, the DP-MPMN model described in 6.1.1, and a default application of FCS with main effects multinomial logistic regression in a large repeated-sampling study of imputation using categorical data from the American Community Survey. The DP-MPMN imputations tended to yield better coverage than FCS-CART overall, but had much worse coverage for a small number of estimands. Manrique-Vallier and Reiter (2014b) also demonstrated the utility of accounting for structural zeros in this model with a population constructed from publicly available data from the U.S. Census. A default version of Murray and Reiter (2016)'s joint model for mixed data types outperformed FCS using the default settings in R's `mice` package (Van Buuren and Groothuis-

Oudshoorn, 2011) in a large repeated-sampling study with data from the Survey of Income and Program Participation. Evidence suggested that misspecification bias was primarily to blame for FCS's poor performance.

### 7.4 Imputation model diagnostics

A more obvious way to choose between imputation models is by fitting multiple and choosing the one that appears to fit the data best. Checking the fit of imputation models is challenging, but some approaches have been proposed. For methods that employ univariate regressions, imputers can examine standard diagnostics for those models (Abayomi, Gelman and Levy, 2008; Su et al., 2011). Abayomi, Gelman and Levy (2008) suggested other diagnostic plots comparing imputed and observed values, primarily comparing marginal and bivariate distributions. Under MAR the distribution of missing values may be different than the distribution of observed values; Bondarenko and Raghunathan (2016) used estimated response propensities to adjust for this and make diagnostic plots more comparable. He and Zaslavsky (2012) proposed posterior predictive checks, comparing the distribution of estimands computed on the multiply imputed datasets to the distribution of those estimands computed on entirely synthetic datasets generated by the imputation method (see also Nguyen, Lee and Carlin (2015)). These checks require the imputer to choose relevant estimands and generate many samples from posterior predictive distributions, which can be computationally expensive.

## 8. CONCLUSION

Over thirty years after Rubin's extensive treatment of MI (Rubin, 1987), experience with the method has cemented its reputation as a principled and practical solution to missing data problems. MI remains an active and fertile research area. While the behavior of the MI estimates have been the subject of intense scrutiny, relatively little is known about the comparative merits of various imputation models that have been proposed in recent years. Considerations based on theoretical findings suggest the use of more flexible imputation models where possible. Empirical evidence also suggests that simple defaults (MVN/log-linear models, or default FCS imputation using simple imputation models such as PMM with linear mean functions or regression models including only main effects) should be avoided, or at least carefully scrutinized.

Nonparametric Bayesian methods for generating imputations have recently emerged as a promising technique for generating imputations. In addition to new model development, more work is needed on scalable posterior computation with these models. In addition, the heuristic justification for why Bayesian MI "tends to be proper" is based on the asymptotic behavior of parametric Bayesian models (Rubin, 1987). It would be interesting to revisit this argument from the perspective of Bayesian nonparametric models, where the asymptotics are more involved (see Rousseau (2016) for a recent review). For example, can semiparametric Bernstein von-Mises results be derived for likely targets of MI inference under Bayesian nonparametric models used for imputation?

Joint-sequential approaches appear understudied and underutilized in the literature, perhaps because they currently require more intervention to set up. More research is needed on the implications of choosing different permutations of the

variables in joint-sequential approaches. Further development of algorithmic approaches for selecting good joint-sequential variable orderings in the same vein as Li et al. (2014) would also be welcome. There remains considerable work to be done in characterizing the behavior of FCS approaches to generating imputations; while some theoretical results exist, they are limited in scope and do not address some of the most effective variants of these algorithms (including PMM and CART).

More empirical comparisons of imputation methods and models are also needed. The field would benefit greatly from a repository of ready-to-use synthetic populations constructed from real data files. A common set of samples from these populations complete with missing values already generated would allow for easy comparisons across methods. A forward-thinking statistical agency could kickstart this repository, providing a public good (and possibly improving the state of their own missing data imputation routines) by sponsoring an imputation challenge in the spirit of a Kaggle competition.

The applications of MI have grown far beyond imputing item missing data in public use files: MI is used with synthetic data for disclosure limitation (Rubin, 1993; Reiter, 2002; Raghunathan, Reiter and Rubin, 2003), to adjust for measurement error (Cole, Chu and Greenland, 2006; Blackwell, Honaker and King, 2015), and to perform statistical matching/data fusion (Rässler, 2004; Reiter, 2012; Fosdick, DeYoreo and Reiter, 2016). In these new settings the amount of missing data can be much greater than typical applications of MI for item missing data, and imputation model development, selection, and assessment is even more consequential. We expect that new models and methods for multiple imputation will be an active research area for the foreseeable future.

## REFERENCES


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

Akande, O., Li, F. and Reiter, J. (2017). An Empirical Comparison of Multiple Imputation Methods for Categorical Data. *The American statistician* 0–0.

Andridge, R. R. and Little, R. J. A. (2010). A Review of Hot Deck Imputation for Survey Non-response. *International statistical review = Revue internationale de statistique* **78** 40–64.

Arnold, B. C., Castillo, E. and Sarabia, J. M. (2001). Conditionally Specified Distributions: An Introduction. *Statistical science: a review journal of the Institute of Mathematical Statistics* **16** 249–265.

Arnold, B. C. and Press, J. S. (1989). Compatible Conditional Distributions. *Journal of the American Statistical Association* **84** 152–156.

Audigier, V., Husson, F. and Josse, J. (2016). Multiple imputation for continuous variables using a Bayesian principal component analysis. *Journal of statistical computation and simulation* **86** 2140–2156.

Audigier, V., Husson, F. and Josse, J. (2017). MIMCA: multiple imputation for categorical variables with multiple correspondence analysis. *Statistics and computing* **27** 501–518.

Banerjee, A., Murray, J. and Dunson, D. B. (2013). Bayesian Learning of Joint Distributions of Objects. In *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Barnard, J. and Rubin, D. B. (1999). Miscellanea. Small-sample degrees of freedom with multiple imputation. *Biometrika* **86** 948–955.

Bernaards, C. A., Belin, T. R. and Schafer, J. L. (2007). Robustness of a multivariate normal approximation for imputation of incomplete binary data. *Statistics in medicine* **26** 1368–1382.

Blackwell, M., Honaker, J. and King, G. (2015). A Unified Approach to Measurement Error and Missing Data. *Sociological methods & research* 0049124115585360.

Böhning, D., Seidel, W., Alfó, M., Garel, B., Patilea, V., Walther, G., Di Zio, M., Guarnera, U. and Luzi, O. (2007). Imputation Through Finite Gaussian Mixture Models. *Computational Statistics & Data Analysis* **51** 5305–5316.

Bondarenko, I. and Raghunathan, T. (2016). Graphical and numerical diagnostic tools to assess suitability of multiple imputations and imputation models. *Statistics in medicine* **35** 3007–3020.

Breiman, L. (2001). Random forests. *Machine learning* **45** 5–32.

Breiman, L., Friedman, J., Stone, C. J. and Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Burgette, L. F. and Reiter, J. P. (2010). Multiple Imputation for Missing Data via Sequential Regression Trees. *American journal of epidemiology* **172** 1070–1076.

Van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software* **45** 1–67.

Carpenter, J. and Kenward, M. (2013). *Multiple Imputation and its Application*, 1 ed. Wiley.

Chen, J. and Shao, J. (2000). Nearest neighbor imputation for survey data. *Journal of official statistics* **16** 113.

Cole, S. R., Chu, H. and Greenland, S. (2006). Multiple-imputation for measurement-error correction. *International journal of epidemiology* **35** 1074–1081.

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

DeYoreo, M., Reiter, J. P. and Hillygus, D. S. (2016). Bayesian Mixture Models with Focused Clustering for Mixed Ordinal and Nominal Data. *Bayesian Analysis*.

Doove, L. L., Van Buuren, S. and Dusseldorp, E. (2014). Recursive partitioning for missing data imputation in the presence of interaction effects. *Computational statistics & data analysis* **72** 92–104.

Drechsler, J. (2010). Multiple imputation of missing values in the wave 2007 of the IAB Establishment Panel. *IAB Discussion Paper*.

Dunson, D. B. and Xing, C. (2009). Nonparametric Bayes Modeling of Multivariate Categorical Data. *Journal of the American Statistical Association* **104** 1042–1051.

Elliott, M. R. and Stettler, N. (2007). Using a Mixture Model for Multiple Imputation in the Presence of Outliers: the "Healthy for Life" Project. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **56** 63–78.

Fithian, W. and Josse, J. (2017). Multiple correspondence analysis and the multilogit bilinear model. *Journal of multivariate analysis* **157** 87–102.

Fosdick, B. K., DeYoreo, M. and Reiter, J. P. (2016). Categorical data fusion using auxiliary information. *The annals of applied statistics* **10** 1907–1929.

Gebregziabher, M. and DeSantis, S. M. (2010). Latent Class Based Multiple Imputation Approach for Missing Categorical Data. *Journal of Statistical Planning and Inference* **140** 3252–3262.

Gelman, A., Carlin, J. B., Rubin, D. B., Vehtari, A., Dunson, D. B. and Stern, H. S. (2013). Bayesian data analysis.

He, Y. and Zaslavsky, A. M. (2012). Diagnosing imputation models by applying target analyses to posterior replicates of completed data. *Statistics in medicine* **31** 1–18.

He, Y., Zaslavsky, A. M., Landrum, M. B., Harrington, D. P. and Catalano, P. (2010). Multiple imputation in a large-scale complex survey: a practical guide. *Statistical methods in medical research* **19** 653–70.

Heitjan, D. F. and Little, R. J. A. (1991). Multiple Imputation for the Fatal Accident Reporting System. *Journal of the Royal Statistical Society. Series C, Applied statistics* **40** 13–29.

Horton, N. J., Lipsitz, S. R. and Parzen, M. (2003). A Potential for Bias When Rounding in Multiple Imputation. *The American statistician* **57** 229–232.

Hu, J., Reiter, J. P. and Wang, Q. (2017). Dirichlet process mixture models for modeling and generating synthetic versions of nested categorical data. *Bayesian analysis* **(to appear)**.

Hughes, R. A., White, I. R., Seaman, S. R., Carpenter, J. R., Tilling, K. and Sterne, J. A. C. (2014). Joint modelling rationale for chained equations. *BMC medical research methodology* **14** 28.

Ibrahim, J. G., Lipsitz, S. R. and Chen, M. H. (1999). Missing Covariates in Generalized Linear Models when the Missing Data Mechanism is Non-Ignorable. *Journal of the Royal Statistical Society, Series B* **61** 173–190.

Ibrahim, J. G., Chen, M. H., Lipsitz, S. R. and Herring, A. H. (2005). Missing Data Methods for Generalized Linear Models: A Comparative Review. *Journal of the American Statistical Association* **100** 332–346.

Ishwaran, H. and James, L. F. (2001). Gibbs Sampling Methods for Stick-Breaking Priors. *Journal of the American Statistical Association* **96** 161–173.

Kim, J. K. (2002). A note on approximate Bayesian bootstrap imputation. *Biometrika* **89** 470–477.

Kim, J. K., Michael Brick, J., Fuller, W. A. and Kalton, G. (2006). On the bias of the multiple-imputation variance estimator in survey sampling. *Journal of the Royal Statistical Society. Series B, Statistical methodology* **68** 509–521.

Kim, H. J., Reiter, J. P., Wang, Q., Cox, L. H. and Karr, A. F. (2014). Multiple Imputation of Missing or Faulty Values Under Linear Constraints. *Journal of business & economic statistics: a publication of the American Statistical Association* **32** 375–386.

Kim, H. J., Cox, L. H., Karr, A. F., Reiter, J. P. and Wang, Q. (2015). Simultaneous Edit-Imputation for Continuous Microdata. *Journal of the American Statistical Association* **110** 987–999.

Kropko, J., Goodrich, B., Gelman, A. and Hill, J. (2014). Multiple imputation for continuous and categorical data: Comparing joint multivariate normal and conditional approaches. *Political Analysis* **22** 497–519.

Lee, M. C. and Mitra, R. (2016). Multiply imputing missing values in data sets with mixed measurement scales using a sequence of generalised linear models. *Computational statistics & data analysis* **95** 24–38.

Li, F., Yu, Y. and Rubin, D. B. (2012). Imputing missing data by fully conditional models: Some cautionary examples and guidelines. *Duke University Department of Statistical ...*.

Li, F., Baccini, M., Mealli, F., Zell, E. R., Frangakis, C. E. and Rubin, D. B. (2014). Multiple Imputation by Ordered Monotone Blocks With Application to the Anthrax Vaccine Research Program. *Journal of computational and graphical statistics: a joint publication of American Statistical Association, Institute of Mathematical Statistics, Interface Foundation of North America* **23** 877–892.

Lipsitz, S. R. and Ibrahim, J. G. (1996). A Conditional Model for Incomplete Covariates in Parametric Regression Models. *Biometrika* **83** 916–922.

Little, R. J. A. (1988). Missing-Data Adjustments in Large Surveys. *Journal of Business & Economic Statistics* **6** 287–296.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*, 2 ed. Wiley-Interscience.

Little, R. J. A. and Schluchter, M. D. (1985). Maximum Likelihood Estimation for Mixed Continuous and Categorical Data with Missing Values. *Biometrika* **72** 497–512.

Liu, J. S. (1994). The Collapsed Gibbs Sampler in Bayesian Computations with Applications to a Gene Regulation Problem. *Journal of the American Statistical Association* **89** 958–966.

Liu, C. and Rubin, D. B. (1998). Ellipsoidally Symmetric Extensions of the General Location Model for Mixed Categorical and Continuous Data. *Biometrika* **85** 673–688.

Liu, J., Gelman, A., Hill, J., Su, Y.-S. and Kropko, J. (2014). On the stationary distribution of iterative imputations. *Biometrika* **101** 155–173.

Manrique-Vallier, D. and Reiter, J. P. (2014a). Bayesian Estimation of Discrete Multivariate Latent Structure Models With Structural Zeros. *Journal of computational and graphical statistics: a joint publication of American Statistical Association, Institute of Mathematical Statistics, Interface Foundation of North America* **23** 1061–1079.

Manrique-Vallier, D. and Reiter, J. P. (2014b). Bayesian multiple imputation for large-scale categorical data with structural zeros. *Survey Methodology* **40** 125-134.

Manrique-Vallier, D. and Reiter, J. P. (2016). Bayesian Simultaneous Edit and Imputation for Multivariate Categorical Data. *Journal of the American Statistical Association* 0–0.

Meng, X.-L. (1994). Multiple-Imputation Inferences with Uncongenial Sources of Input. *Statistical Science* **9** 538–558.

Meng, X.-L. and Romero, M. (2003). Discussion: Efficiency and Self-Efficiency with Multiple Imputation Inference. *International statistical review = Revue internationale de statistique* **71** 607–618.

Morris, T. P., White, I. R. and Royston, P. (2014). Tuning multiple imputation by predictive mean matching and local residual draws. *BMC medical research methodology* **14** 75.

Murray, J. S. and Reiter, J. P. (2016). Multiple Imputation of Missing Categorical and

Continuous Values via Bayesian Mixture Models With Local Dependence. *Journal of the American Statistical Association* **111** 1466–1479.

Nguyen, C. D., Lee, K. J. and Carlin, J. B. (2015). Posterior predictive checking of multiple imputation models. *Biometrical journal. Biometrische Zeitschrift* **57** 676–694.

Nielsen, S. F. (2003). Proper and Improper Multiple Imputation. *International statistical review = Revue internationale de statistique* **71** 593–607.

Olkin, I. and Tate, R. F. (1961). Multivariate Correlation Models with Mixed Discrete and Continuous Variables. *The Annals of Mathematical Statistics* **32** 448–465.

Paddock, S. M. (2002). Bayesian nonparametric multiple imputation of partially observed data with ignorable nonresponse. *Biometrika* **89** 529–538.

Raghunathan, T. E., Reiter, J. P. and Rubin, D. B. (2003). Multiple imputation for statistical disclosure limitation. *Journal of official statistics* **19** 1.

Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J. and Solenberger, P. (2001). A Multivariate Technique for Multiply Imputing Missing Values using a Sequence of Regression Models. *Survey methodology* **27** 85–96.

Rässler, S. (2004). Data Fusion: Identification Problems, Validity, and Multiple Imputation. *Austrian Journal of Statistics* **33** 153–171.

Reiter, J. P. (2002). Satisfying disclosure restrictions with synthetic data sets. *Journal of official statistics* **18** 531.

Reiter, J. P. (2005). Using CART to generate partially synthetic public use microdata. *Journal of Official Statistics* **21** 441.

Reiter, J. P. (2012). BAYESIAN FINITE POPULATION IMPUTATION FOR DATA FUSION. *Statistica Sinica* **22** 795–811.

Reiter, J. (2017). Discussion: Dissecting multiple imputation from a multi-phase inference perspective: what happens when God's, imputer's and analyst's models are uncongenial? *Statistica Sinica*.

Reiter, J. P., Raghunathan, T. E. and Kinney, S. K. (2006). The importance of modeling the sampling design in multiple imputation for missing data. *Survey methodology* **32** 143.

Reiter, J. P. and Raghunathan, T. E. (2007). The Multiple Adaptations of Multiple Imputation. *Journal of the American Statistical Association* **102** 1462–1471.

Robins, J. M. and Wang, N. (2000). Inference for imputation estimators. *Biometrika* **87** 113–124.

Rousseau, J. (2016). On the Frequentist Properties of Bayesian Nonparametric Methods. *Annual Review of Statistics and Its Application* **3** 211–231.

Rubin, D. B. (1981). The Bayesian Bootstrap. *Annals of statistics* **9** 130–134.

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

Rubin, D. B. (1993). Discussion: statistical disclosure limitation. *Journal of official statistics* **9** 461–468.

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

Rubin, D. B. (2003a). Discussion on Multiple Imputation. *International statistical review = Revue internationale de statistique* **71** 619–625.

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

Rubin, D. B. and Schafer, J. L. (1990). Efficiently creating multiple imputations for incomplete multivariate normal data. In *Proceedings of the Statistical Computing Section of the American Statistical Association* **83** 88.

Rubin, D. B. and Schenker, N. (1986). Multiple Imputation for Interval Estimation from Simple Random Samples with Ignorable Nonresponse. *Journal of the American Statistical Association* **81** 366–374.

Schafer, J. (1997). *Analysis of Incomplete Multivariate Data*. CRC press.

Schafer, J. L. (2003). Multiple Imputation in Multivariate Problems When the Imputation and Analysis Models Differ. *Statistica Neerlandica* **57** 19–35.

Schenker, N. and Taylor, J. M. G. (1996). Partially parametric techniques for multiple imputation. *Computational statistics & data analysis* **22** 425–446.

Schifeling, T. A. and Reiter, J. P. (2016). Incorporating Marginal Prior Information in Latent Class Models. *Bayesian analysis* **11** 499–518.

Seaman, S. R. and Hughes, R. A. (2016). Relative efficiency of joint-model and full-conditional-specification multiple imputation when conditional models are compatible: The

general location model. *Statistical methods in medical research* 0962280216665872.
Sethuraman, J. (1994). A Constructive Definition of Dirichlet Priors. *Statistica Sinica* **4** 639–650.
Shah, A. D., Bartlett, J. W., Carpenter, J., Nicholas, O. and Hemingway, H. (2014). Comparison of Random Forest and Parametric Imputation Models for Imputing Missing Data Using MICE: A CALIBER Study. *American journal of epidemiology* **179** 764–774.
Si, Y. and Reiter, J. P. (2013). Nonparametric Bayesian Multiple Imputation for Incomplete Categorical Variables in Large-Scale Assessment Surveys. *Journal of Educational and Behavioral Statistics* **38** 499-521.
Stuart, E. A., Azur, M., Frangakis, C. and Leaf, P. (2009). Multiple imputation with large data sets: a case study of the Children's Mental Health Initiative. *American journal of epidemiology* **169** 1133–9.
Su, Y.-S., Gelman, A., Hill, J., Yajima, M. et al. (2011). Multiple imputation with diagnostics (mi) in R: Opening windows into the black box. *Journal of Statistical Software* **45** 1–31.
Van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical methods in medical research* **16** 219–42.
Van Buuren, S. (2012). Flexible imputation of missing data.
Van Buuren, S. and Oudshoorn, K. (1999). Flexible Multivariate Imputation by MICE. *Leiden, The Netherlands: TNO Prevention Center*.
Van Buuren, S., Brand, J. P. L., Groothuis-Oudshoorn, C. G. M. and Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation* **76** 1049–1064.
Vermunt, J. K., Van Ginkel, J. R., Van Der Ark, L. A. and Sijtsma, K. (2008). Multiple Imputation of Incomplete Categorial Data using Latent Class Analysis. *Sociological Methodology* **38** 369–397.
Vidotto, D., Vermunt, J. K. and Kaptein, M. C. (2015). Multiple imputation of missing categorical data using latent class models: State of art. *Psychological test and assessment modeling* **57** 542–576.
Vink, G., Frank, L. E., Pannekoek, J. and van Buuren, S. (2014). Predictive mean matching imputation of semicontinuous variables. *Statistica Neerlandica* **68** 61–90.
Wang, N. and Robins, J. M. (1998). Large-sample theory for parametric multiple imputation procedures. *Biometrika* **85** 935–948.
Xie, X. and Meng, X.-L. (2017). Dissecting multiple imputation from a multi-phase inference perspective: what happens when God's, imputer's and analyst's models are uncongenial? *Statistica Sinica*.
Xu, D., Daniels, M. J. and Winterstein, A. G. (2016). Sequential BART for imputation of missing covariates. *Biostatistics*.
Zhu, J. and Raghunathan, T. E. (2015). Convergence Properties of a Sequential Regression Multiple Imputation Algorithm. *Journal of the American Statistical Association* **110** 1112–1124.

## APPENDIX A: SOFTWARE FOR MULTIPLE IMPUTATION

Pointers to many software implementations of MI methods are available at `http://www.stefvanbuuren.nl/mi/Software.html`, an updated version of Appendix A of Van Buuren (2012). As of December 2017, it is missing links to R packages for several nonparametric Bayesian joint models: These include the R packages MixedDataImpute (imputation for mixed continuous and categorical missing values using the model in Murray and Reiter (2016)), NPBayesImpute (imputation for multivariate categorical data, possibly with structural zeros, as presented in Si and Reiter (2013); Manrique-Vallier and Reiter (2014a,b)), and NestedCategBayesImpute (imputation got multivariate categorical data with hierarchical data structures, as described in Hu, Reiter and Wang (2017)).