

Imputing Missing Data by Fully Conditional Models: Some Cautionary Examples and Guidelines

Fan Li, Yaming Yu and Donald B. Rubin ¹

February 2, 2012

ABSTRACT

Missing data are pervasive in large public-use databases. Multiple imputation (MI) is an effective methodology to handle the problem. Current state-of-the-art procedures of MI often fit fully Bayesian models assuming some joint probability distribution for the underlying complete data. Though theoretically valid, joint modeling may not accurately capture the important relations between the variables that are outside that theoretical structure. Alternatively, a widely used strategy - multiple imputation using chained equations (MICE), first specifies a set of univariate conditional models and then iteratively imputes the missing data based on these conditional models. Though practically flexible, MICE defines a possibly incompatible Gibbs sampler (PIGS) when there is no joint distribution corresponding to the specified conditional distributions. We construct several examples to reveal some of the undesirable theoretical and algorithmic properties of a PIGS. We then propose a spectrum of imputation strategies, imputation by monotone blocks (IMB), which combines (1) sequential imputation for monotone missing data, (2) and a fully conditional strategy like MICE when (1) cannot be applied. The key is to partition an arbitrary missing data pattern into a series of monotone patterns. We further provide some general guidelines for choosing strategies within this spectrum in practice.

KEY WORDS: conditional imputation; Gibbs sampler; incompatibility; MICE; missing data; monotone block; multiple imputation; sequential.

¹Fan Li is assistant professor, Department of Statistical Science, Duke University, Durham, NC (email: fli@stat.duke.edu); Yaming Yu is associate professor, Department of Statistics, University of California, Irvine (email: yamingy@uci.edu); Donald Rubin is the John Loeb Professor, Department of Statistics, Harvard University, Cambridge, MA (email: rubin@stat.harvard.edu). We thank Fabrizia Mealli and Jerry Reiter for stimulating discussions. This research is partially funded by NSF-SES grant 11-31897.

1. MISSING DATA AND IMPUTATION IN LARGE SURVEYS

Missing data are pervasive in large databases, including many national surveys. Multiple imputation (MI) proposed by Rubin (1978) (see also Rubin, 1996, 2004) is an effective methodology to handle the problem. This methodology consists of three steps: (1) the imputer replaces each missing entry by a few plausible values, thereby creates completed datasets; (2) each resultant completed dataset is analyzed using standard statistical tools; and (3) the estimates or hypothesis tests from these multiple complete-data analyses are combined by simple “repeated imputation” procedures (also known as Rubin’s Rules) to yield inference for the original incomplete-data problem, which are valid under general conditions.

The most challenging step of the three is imputing the missing data. Analysts’ ease of drawing valid inferences by performing standard complete-data procedures must be traded off against the burden on the part of the imputer to impute sensibly. Current state-of-the-art procedures for imputing missing data either fit fully Bayesian models assuming some joint probability distribution for the underlying complete data, or fit possibly incompatible but flexible univariate conditional models. Other ad hoc methods include imputing the mean, imputing from regression estimates, hot-deck imputation, etc., but these have been shown to be generally unsatisfactory in practice.

Challenges to MI in large databases include (but not limited to):

1. *Arbitrary missing data patterns.* It is well-known that when missing data follow a monotone pattern, that is, when the variables can be ordered in a way such that the $j + 1$ th variable is missing whenever the j th is), inference and imputation can be performed easily in a principled fashion (see, for example, Rubin, 1974). However, missing data in practice rarely follow an exact monotone pattern, though they sometimes follow an approximate one.
2. *Many different types of variables.* Fully principled methods, which essentially require

a joint probability distribution on the underlying complete data, are not often used because large datasets usually consist of a mixture of discrete and continuous variables; often there are semi-continuous variables. For example, itemized income or expenditure data typically have a positive probability of being zero, with an approximately continuous complicated distribution when positive. Moreover, there are often special types of variables such as variables with very low variability or defined only on a subset of units (Li et. al, 2011) for which standard regression models not adequate. Given only a limited pool of flexible multivariate distributions, multivariate modeling is a real challenge.

3. *Observable differences between respondents and non-respondents.* A further complication is the complex response mechanisms that can generate missing data. Assumptions such as MCAR (missing completely at random), which justifies simple case-deletion, are usually inappropriate. Also, it is difficult to evaluate the quality of the imputations.

The multivariate modeling approach is theoretically valid, but fails to address at least some of the above complications. Though usually better than ad hoc methods in practice, it is limited by the paucity of available multivariate models that are both computationally tractable and provide good fits to real data. Some of these models currently used are (1) multivariate normal models (Schafer's NORM package), (2) normal models with random effects, (3) multivariate-t models, (4) multinomial log-linear models, and (5) general location models for mixed continuous and discrete variables (e.g., Liu and Rubin 1998). Tools for the normal, log-linear, and general location models as described in Schafer (1997) are available in S Plus and SAS (SAS Institute Inc., 2008).

Although the list is expanding, there is an inherent difficulty using theoretical models with real multivariate data. For example, it is difficult to incorporate semi-continuous variables in the model (Javaras and van Dyk, 2003). Jointly modeling multivariate data is complicated enough; taking into account missing data adds another level of complication. As a result, to

handle general missing data patterns, the joint modeling approach requires, for each model, specialized expertise in both statistical computing methodology and software development.

Another strategy - multiple imputation using chained equations (MICE) is based on iteratively sampling from separate univariate conditional models for each variable with missing data. Despite the fact that little is known about its theoretical properties, MICE is one of the most widely used imputation strategies based on fully conditional model. It has been applied to the imputation of many complex large scale databases in medical and social research (e.g., Kennickell, 1991; Raghunathan and Siscovick, 1996; Kennickell, 1999; van Buuren et al., 1999; Oudshoorn et al., 1999; Gelman and Raghunathan, 2001; Heeringa et al., 2002; Faris, et al. 2002; Schenker et al. 2006; Azur et al. 2009; He et al., 2011, to name a few); and implemented in many software systems, e.g., FRITZ (Kennickell 1991), HERMES missing data engine (Brand 1999), IVEWARE (Raghunathan, Solenberger and van Hoewyk, 2000), and MICE (van Buuren and Oudshoorn, 2000). A recent summary can be found in van Buuren (2007).

The contrast between the popularity in practice and the lack of understanding in theory of MICE motivates us to explore some of its theoretical and algorithmic properties, especially those related to incompatibility of the conditional models (defined in Section 2). Here we do not attempt to provide a complete solution to the challenging problem of incompatibility, but rather to generate interest in this important yet much open topic by illustrating the potential problems of MICE and proposing an intuitive but preliminary framework. The rest of the paper is organized as follows. In Section 2, we examine some of the theoretical and operational properties and potential problems of MICE through simple examples. In Section 3, we propose a general imputation framework “imputation by monotone blocks (IMB)” that combines the merits of the sequential imputation for monotone missing data and the fully conditional model based imputation approaches like MICE. Within this framework, one has the flexibility to find strategies that reduce incompatibility as much as possible and thus mit-

igates the incompatibility problem. Two simple measures of (in)compatibility are introduced to compare the IMB strategies. Section 4 concludes with a discussion.

2. MICE, PIGS AND THE INCOMPATIBILITY PROBLEM

In this section, we first review the MICE strategy and then examine some of its undesirable theoretical and algorithmic properties through simple examples. As an algorithm, MICE is a *Possibly Incompatible Gibbs Sampler* (PIGS), which may not converge to any distribution. Our examples serve as a signal of caution for using MICE as a universal imputation tool, although this method has displayed success in many practical examples (e.g., van Buuren et al., 2006).

2.1. Definition of MICE and PIGS

Suppose there are N ($i = 1, \dots, N$) units and J variables Y_j ($j = 1, \dots, J$). Denote the value of Y_j for unit i by y_{ij} , and the observed and missing values in Y_j by Y_j^{obs} and Y_j^{mis} , respectively. Let $Y_{-j} = \{Y_k : k \neq j\}$, $Y^{mis} = \{Y_1^{mis}, \dots, Y_J^{mis}\}$ and $Y^{obs} = \{Y_1^{obs}, \dots, Y_J^{obs}\}$. Define the response indicator $S_{ij} = \mathbf{1}\{y_{ij} \text{ is observed}\}$ for all i, j , and let $M_j = \{i : S_{ij} = 0\}$, i.e., the set of missing entries in variable j , and $M = \bigcup_{j=1}^J M_j$.

Throughout this paper we assume the units are exchangeable, and the missing data are missing at random (MAR) (Rubin, 1976). MICE, as an algorithm, is defined as follows:

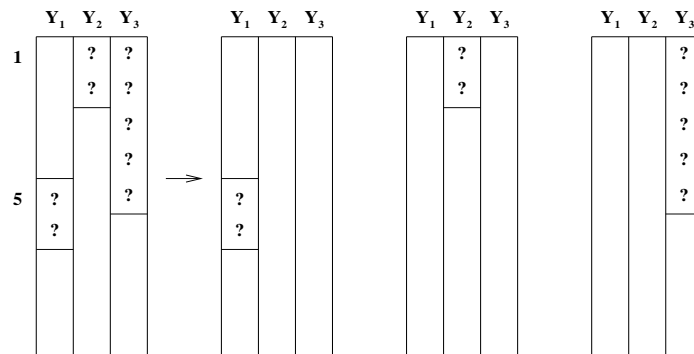
MICE For $j = 1, \dots, J$: (1) specify a model of Y_j conditional on Y_{-j} or a subset of Y_{-j} , (2) estimate the model parameters using only the units with observed Y_j and the most recently imputed values in Y_{-j} as observed, and (3) impute Y_j^{mis} from its predictive distribution based on this model and the parameters. Iterate until convergence under certain criterion.

Figure 1 illustrates a simple MICE with three variables, where missing data are marked by “?”. The main advantages of MICE include:

A.1 It is simple in that it reduces a multivariate incomplete data problem to a collection of univariate complete data problems. Consequently, it is highly flexible. For example, to help find a theoretical model that fits the empirical distribution of the data, transformations can be applied on both more observed covariates and the response variables, and one can specify non-Gaussian regression models with interactions and non-linear effects.

A.2 It is applicable to any pattern of missing data.

Figure 1: MICE with three variables. MICE imputes missing data in each variable in turn by fitting fully conditional models. For example, to impute Y_1^{mis} , MICE fits a model of Y_1^{obs} given Y_2, Y_3 (using both imputed and observed values of Y_2, Y_3), and then draws from the resulting predictive distribution.



MICE is usually implemented from a Bayesian perspective in practice. That is, we first specify a prior distribution for the parameters $\theta_{j|-j}$ of the conditional model of $Y_j|Y_{-j}$, then draw from the posterior predictive distribution of $\theta_{j|-j}$, and then draw Y_j^{mis} from its predictive distribution given the draw of $\theta_{j|-j}$ and Y^{obs} . Algorithmically, step j simply draws Y_j^{mis} conditional on Y_{-j}^{mis} (and Y^{obs}) according to certain probabilistic rules, which may be difficult to write down explicitly as a formula, especially for hot-deck procedures. From this perspective, MICE defines a PIGS, which is formally defined as follows:

PIGS Given a set of fully conditional densities $f_1(z_1|z_{-1}), f_2(z_2|z_{-2}), \dots, f_J(z_J|z_{-J})$,

and a starting value $z^{(0)} = (z_1^{(0)}, \dots, z_J^{(0)})$, iteratively draw z_1, \dots, z_J in turn according to these conditional distributions. That is, at iteration $t + 1$, we draw $z_1^{(t+1)}, \dots, z_J^{(t+1)}$ conditional on $z^{(t)}$ in the following fashion:

$$\begin{aligned} z_1^{(t+1)} &\sim f_1(\cdot | z_2^{(t)}, \dots, z_J^{(t)}), \\ z_2^{(t+1)} &\sim f_2(\cdot | z_1^{(t+1)}, z_3^{(t)}, \dots, z_J^{(t)}), \\ &\dots \\ z_J^{(t+1)} &\sim f_J(\cdot | z_1^{(t+1)}, z_2^{(t+1)}, \dots, z_{J-1}^{(t+1)}). \end{aligned}$$

We call the distributions f_1, \dots, f_J “conditional specifications”, which, when combined with an *update order*, define a PIGS algorithm. The update order above is denoted $[z_1, z_2, \dots, z_J]$. In the MICE, $z_j = Y_j^{mis}$, the missing data in variable j .

Notice that PIGS is simply the Gibbs sampler (Geman and Geman, 1984; Gelfand and Smith, 1990) except that the full conditionals $f_j(\cdot | \cdot)$, $j = 1, \dots, J$ may be *incompatible*, that is, there may exist no joint density $f(z_1, \dots, z_J)$ such that $f_j(\cdot | \cdot)$ are the corresponding full conditionals. The conditions under which a set of full conditionals are compatible have been studied, e.g., by Besag (1974), Arnold and Press (1989), and Hobert and Casella (1998). We note that incompatibility is a distinct concept from uncongeniality, which means a procedure for analyzing multiply imputed data sets cannot be derived from (is “uncongenial” to) the model adopted for MI (Meng, 1994); on the contrary, incompatibility is a concept involving solely the imputation models.

2.2. Examples of problems with PIGS

When imputing missing data using MICE, the heuristic hope is that by iterating between the fully conditional draws, the chain will converge to a single stationary distribution whose full conditionals are approximately $f_1(\cdot | \cdot), \dots, f_J(\cdot | \cdot)$. However, for a PIGS, this convergence

may not occur when the conditionals are incompatible; rather, different orders of update can give drastically different limiting distributions (if limiting distributions exist at all). In contrast, for a proper Gibbs sampler, different update orders may differ in convergence rates, but their stationary distributions are the same. We summarize this nontrivial problem of PIGS by the following mathematically trivial result.

Proposition 2.1 *If the J conditional specifications of a PIGS are incompatible, then the $J!$ deterministic update orders may give different stationary distribution.*

Below we show examples of PIGS such that (1) one order of update converges but another does not, (2) every deterministic order converges but random orders do not, and (3) all random orders converge, but the deterministic orders do not. Since MICE defines a PIGS, our examples illustrate the danger of using MICE, although in practice MICE may not be as extreme as these examples indicate.

Example 1. Here each fully conditional distribution is Gaussian whose mean is linear in the other variables.

$$\begin{aligned} f_1(z_1|z_2, z_3) &= N(-1.5z_2 - 0.5z_3, 1), \\ f_2(z_2|z_1, z_3) &= N(-0.5z_1 - 0.5z_3, 1), \\ f_3(z_3|z_1, z_2) &= N(-1.5z_1 - 1.5z_2, 1). \end{aligned}$$

When the update order is $[z_1, z_2, z_3]$, it can be shown that $\{z^{(t)}, t \geq 0\}$ form a transient Markov chain. In other words, $z^{(t)}$ does not converge to any stationary distribution. In contrast, the update order $[z_2, z_1, z_3]$ results in a convergent Markov chain (Appendix A contains a detailed proof).

Intuitively, we can focus on the mean $(w_1, w_2, w_3) = E(z_1, z_2, z_3)$, which propagates through the iterations according to a (deterministic) linear system:

- Step 1. set $w_1 = -1.5w_2 - 0.5w_3$,

- Step 2. set $w_2 = -0.5w_1 - 0.5w_3$,
- Step 3. set $w_3 = -1.5w_1 - 1.5w_2$,

where $(w_1, w_2, w_3) = (0, 0, 0)$ is the obvious fixed point, i.e., the only solution of all three steps. Except for certain special starting values, e.g., $(0, 0, 0)$ itself, if we iterate Step 1 \rightarrow Step 2 \rightarrow Step 3 \dots , then $(w_1^{(t)}, w_2^{(t)}, w_3^{(t)})$ diverges, i.e., $(0, 0, 0)$ is an *unstable fixed point* for this update strategy. If we iterate Step 2 \rightarrow Step 1 \rightarrow Step 3 \dots , however, $(w_1^{(t)}, w_2^{(t)}, w_3^{(t)})$ always converges to $(0, 0, 0)$. (See Table 1 for an illustration.) Because the three steps prescribe (in some sense) contradictory relationships between w_1, w_2, w_3 , convergence for one update order does not imply convergence for another; this is a peculiar feature of an iterative linear system. As a stochastic generalization of this linear system, a PIGS behaves in a similar fashion.

Table 1: The behavior of a linear iterative system depending on the order of update. Starting from $(1, 1, 1)$, one order converges but another diverges. The three steps for updating w_1, w_2, w_3 are given in Example 1.

iteration	$w_2 \rightarrow w_1 \rightarrow w_3$			$w_1 \rightarrow w_2 \rightarrow w_3$		
	w_1	w_2	w_3	w_1	w_2	w_3
0	1.00000	1.00000	1.00000	1.000	1.000	1.000
1	1.00000	-1.00000	0.00000	-2.000	0.500	2.250
2	0.75000	-0.50000	-0.37500	-1.875	-0.188	3.094
\vdots		\vdots			\vdots	
10	-0.01516	0.00899	0.00925	1.108	1.848	-4.433
20	0.00008	0.00001	-0.00014	6.027	-2.222	-5.707
30	0.00000	0.00000	0.00000	-0.254	-7.233	11.230
\vdots		\vdots			\vdots	
100	0.00000	0.00000	0.00000	333.517	309.263	-964.170

In addition to the deterministic update orders, the random order is defined by (at each iteration) the randomly choosing $j = 1, \dots, J$ with equal probability and updating z_j given z_{-j} according to $f_j(z_j|z_{-j})$; and the random permutation order is defined by randomly choosing

each of the $J!$ deterministic orders $[z_{j_1}, z_{j_2}, \dots, z_{j_J}]$ with equal probability.

It is intuitively appealing to use a PIGS with the random orders, because of the symmetry among all J variables. One would hope that even though the deterministic orders give different stationary distributions (or no stationary distribution at all), the random or random permutation order should at least converge and give an “average” stationary distribution whose full conditionals approximately match the specified $f_j(\cdot|\cdot)$. This is, unfortunately, incorrect. Below we show two nontrivial examples of PIGS: In one, every deterministic order converges, but neither the random order nor the random permutation order does; in the other, the random and random permutation orders converge but no deterministic order does.

Example 2. Consider a PIGS defined on the state space $(z_1, z_2, z_3) \in \{0, 1\} \times \{0, 1, \dots\} \times \{0, 1, \dots\}$, with the constraint $|z_2 - z_3| \leq 1$. Let $\epsilon = 0.01$ and $r_1, r_2 > 1$ (to be determined). The full conditionals are specified below.

- $\Pr(z_1|z_2, z_3)$:

	$z_2 = z_3$	$z_2 \neq z_3$
$z_1 = 0$	$1 - \epsilon$	ϵ
$z_1 = 1$	ϵ	$1 - \epsilon$

- $\Pr(z_2|z_1 = 0, z_3)$:

	$z_3 = 1 \pmod 3$	$z_3 = 2 \pmod 3$	$z_3 = 0 \pmod 3$
$z_2 = z_3$	ϵ	ϵ	ϵ
$z_2 = z_3 + 1$	$(1 - \epsilon)/r_1$	$(1 - \epsilon)/r_2$	$(1 - \epsilon)/r_1$
$z_2 = z_3 - 1$	$(1 - \epsilon)(r_1 - 1)/r_1$	$(1 - \epsilon)(r_2 - 1)/r_2$	$(1 - \epsilon)(r_1 - 1)/r_1$

$\Pr(z_2|z_1 = 1, z_3) :$

	$z_3 = 1 \pmod 3$	$z_3 = 2 \pmod 3$	$z_3 = 0 \pmod 3$
$z_2 = z_3$	ϵ	ϵ	ϵ
$z_2 = z_3 + 1$	$(1 - \epsilon)/r_2$	$(1 - \epsilon)/r_1$	$(1 - \epsilon)/r_1$
$z_2 = z_3 - 1$	$(1 - \epsilon)(r_2 - 1)/r_2$	$(1 - \epsilon)(r_1 - 1)/r_1$	$(1 - \epsilon)(r_1 - 1)/r_1$

• $\Pr(z_3|z_1, z_2) = \Pr(z_3|z_2) :$

$z_3 = z_2$	$z_3 = z_2 + 1$	$z_3 = z_2 - 1$
$1 - 2\epsilon$	ϵ	ϵ

These specifications need to be slightly modified when z_2 and z_3 are small. Since we are only concerned about the recurrent/transient properties for different orders of update, this does not matter as long as the Markov chains are irreducible and aperiodic.

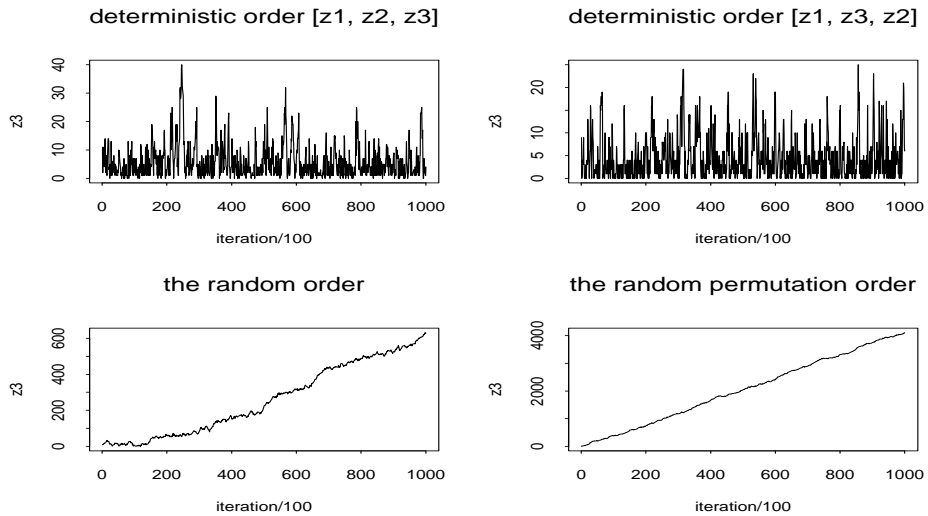
To help decipher the example, notice (a) an update for z_1 usually just sets $z_1 = |z_2 - z_3|$; (b) an update for z_2 typically results in $z_2 \neq z_3$; and (c) an update for z_3 mostly copies z_2 to z_3 . It is easy to check that for the deterministic order $[z_1, z_2, z_3]$, we have $z_1 = 0$ with high probability, whereas for $[z_1, z_3, z_2]$, we have $z_1 = 1$ with high probability. In other words, different deterministic orders give completely different probabilistic rules. Perhaps more surprising is the following result (Appendix B contains the proof):

Result 1. *If $r_1 = 1.3$, $r_2 = 31$, then every deterministic order converges but neither the random nor the random permutation order does. If $r_1 = 5$, $r_2 = 1.02$, however, the random and random permutation orders converge but no deterministic order does.*

As an illustration, for both examples we carried out a small simulation where 100,000 iterations are produced for each of the following four updating strategies: (a) deterministic order $[z_1, z_2, z_3]$, (b) deterministic order $[z_1, z_3, z_2]$, (c) the random order, and (d) the random

permutation order. When $r_1 = 1.3$, $r_2 = 31$, Figure 2 shows the time series plots of the draws for z_3 . The corresponding plots for $r_1 = 5$, $r_2 = 1.02$ in Figure 3 shows exactly the opposite. The simulation therefore supports above result.

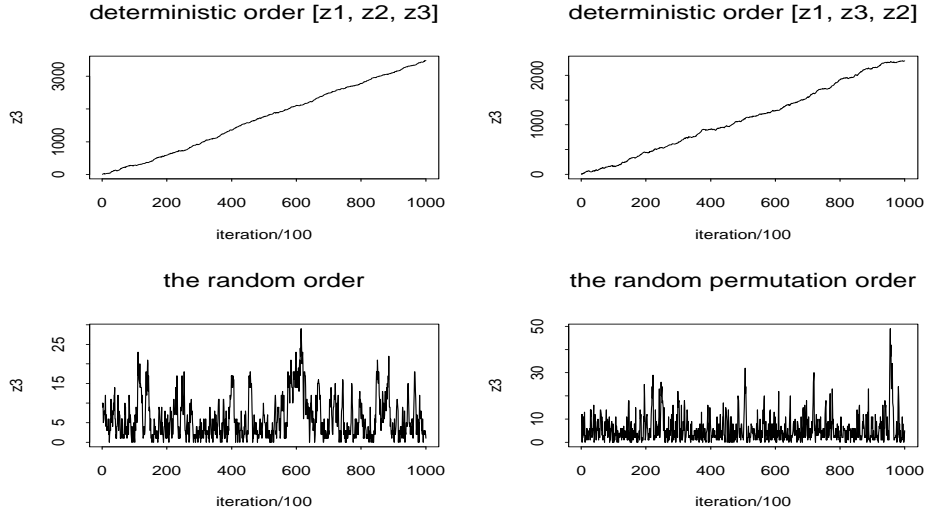
Figure 2: Trajectories of z_3 when $r_1 = 1.3$, $r_2 = 31$. The deterministic orders appear to stabilize, whereas the random and random permutation orders drift towards infinity.



2.3. Example of problems with MICE

In the above examples, we fixed the parameters of the univariate conditional distributions to reveal some of the undesirable features of PIGS. These features are mainly algorithmic and have little to do with how well the conditional models fit the data. In fact, the parameters are usually estimated and updated iteratively from fitting the specified models to the data in real implementation of MICE. This, on one hand, prevents arbitrary specification of model parameters as in the above examples; on the other hand, however, also introduces the possibility of model misspecification that may induce incompatibility and unreasonable imputations. Below we present a simple example that highlights such problems in MICE.

Figure 3: Trajectories of z_3 when $r_1 = 5$, $r_2 = 1.02$. The random and random permutation orders appear to stabilize, whereas the deterministic orders drift towards infinity.



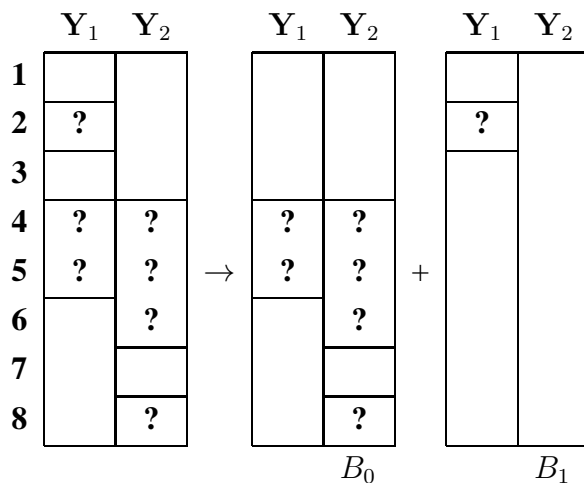
Example 3. Consider a case where the complete data of two variables Y_1, Y_2 are generated from the following normal distributions ($N = 60$):

$$Y_1 \sim N(10, 1) \quad \text{and} \quad Y_2|Y_1 \sim N(3Y_1, 2). \quad (1)$$

It is easy to show that $Y_2 \sim N(30, 11)$. The missing data pattern is shown in the left side of Figure 4, where the values of Y_2 's are arranged in increasing order from unit 1 to 8 (here each "unit" represents a set of units). Variable Y_2 is missing in the middle (units 4, 5, 6) and the high end (unit 8), and Y_1 is missing in units 2, 4, 5.

Consider two partially misspecified conditionals: $Y_1|Y_2$ and $Y_2|Y_1^2$. Upon initializing by imputing all the missing data from their marginal distributions, a standard MICE iterates between: (1) impute the missing Y_1 's by sampling from the posterior predictive distribution of $Y_1|Y_2$; and (2) impute Y_2 by the posterior predictions from fitting $Y_2|Y_1^2$. We use standard non-informative prior for the regression coefficients and variances. The standardize error in

Figure 4: An example where IMMB and IOMB outperform MICE



mean (i.e., the absolute error divided by the true standard deviation) of the imputations of Y_1, Y_2 from MICE are shown in the left two columns of Table 2: the errors rapidly evolves to infinity after only 10 iterations.

The key components of this example are misspecification and extrapolation. Here Y_1 is correctly assumed to be linearly dependent on Y_2 , but Y_2 is wrongly assumed to be linearly dependent on the square term Y_1^2 . In MICE, this model misspecification induces incompatibility between the conditional distributions $f(y_2|y_1)$ and $f(y_1|y_2)$. Since the imputation relies on extrapolation in the high end of Y_2 , the imputation error propagates rapidly as MICE iterates between conditionals $Y_1|Y_2$ and $Y_2|Y_1^2$, leading to meaningless imputations in both Y_1 and Y_2 .

In the current practice of MICE for large and complex data sets, to deal with the large number and various types of variables, the univariate conditional distributions are usually specified in an automatic fashion, e.g., assuming *a priori* linear models between the variables. Model selection, such as choice of predictors, transformations on response variables and covariates, is occasionally conducted (e.g., Li et al., 2011), but is often limited due to the

Table 2: The standardized error of mean in Y_1, Y_2 under MICE and IMMB

iteration	MICE		IMMB	
	Y_1	Y_2	Y_1	Y_2
0	0.17	0.07	0.17	0.07
1	0.06	0.02	0.12	0.15
2	0.02	0.05	0.14	0.19
3	0.04	0.56	0.14	0.17
4	0.53	5.18	0.18	0.23
\vdots	\vdots	\vdots	\vdots	\vdots
9	10e10	10e21	0.14	0.16
10	∞	∞	0.12	0.12

large p . As a result, misspecification and extrapolation can be introduced in such procedures and the scenario illustrated by the above example may not be uncommon.

The above examples, for the first time in the literature to our knowledge, explicitly reveal some undesirable theoretical and algorithmic properties of PIGS (and MICE) under certain scenarios. However, they do not undermine the usefulness of MICE as a flexible imputation tool. In fact, MICE has performed remarkably well in both simulations (e.g., van Buuren et al., 2006) and a wide range of real applications (see the references mentioned in Section 1). Nevertheless, these examples suggest that PIGS-based procedures, such as MICE, should not be used without caution as general tools to impute missing data and more intensive research on the theoretical aspects of MICE are needed.

3. A GENERAL FRAMEWORK: IMPUTATION BY MONOTONE BLOCKS

3.1. Sequential imputation for monotone missing data

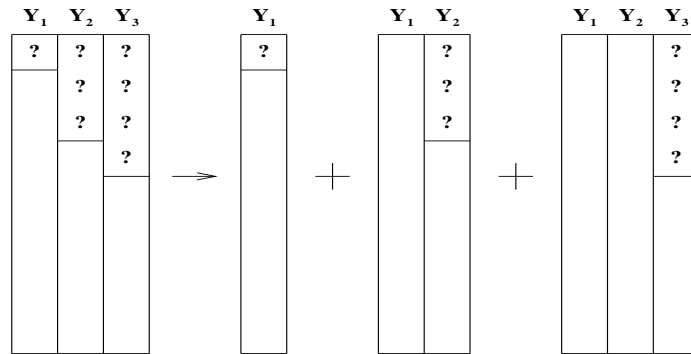
Flexible, theoretically justified and easy-to-implement methods exist when the missing data are in a monotone pattern. In fact, monotone missingness is the basis for some long-standing commercial software for MI, such as SOLAS 3.0 (Statistical Solutions Ltd., 2001). A set

of missing data are *monotone* if Y_{j+1} is missing whenever Y_j is, that is, $M_j \subset M_{j+1}$, for $j = 1, \dots, J - 1$. One can sequentially impute the monotonically missing data Y^{mis} from conditionally specified models as follows (also implemented from a Bayesian perspective):

Sequential Imputation for Monotone Missing Data For $j = 1, \dots, J$: (1) specify a distribution of Y_j conditional on the more observed variables Y_1, \dots, Y_{j-1} and a prior distribution for the parameters θ_j , (2) obtain posterior draws of θ_j using only the units with observed Y_j , and (3) then impute Y_j^{mis} by its random draws from the posterior predictive distribution given the posterior draw of θ_j and Y^{obs} .

Figure 5 illustrates the sequential imputation scheme in a data matrix with three variables. Under general conditions, the sequential imputation scheme for monotone missing

Figure 5: Sequential imputation for monotone missing data with three variables. First impute Y_1^{mis} by fitting a model on Y_1^{obs} and drawing from the posterior predictive distribution of Y_1^{mis} given Y_1^{obs} ; then impute Y_2^{mis} by fitting a (regression) model of Y_2^{obs} given Y_1 and drawing from the predictive distribution of Y_2^{mis} given both imputed and observed Y_1 ; and then impute Y_3^{mis} by fitting a model of Y_3^{obs} given Y_1, Y_2 , and drawing from the predictive distribution of Y_3^{mis} given both imputed and observed Y_1, Y_2 .



data converges in one iteration, due to the factorization of the likelihood, together with the prior independence of the parameters corresponding to the univariate conditional models (see Rubin, 1974, 2004; Little and Rubin, 2002, for details). Sequential imputation shares the advantage A.1 with MICE, but unlike MICE, it is theoretically valid when the missing data are monotone and the conditional models have distinct parameters. The disadvantage is also

obvious: It is limited to data with an exact monotone missingness pattern.

3.2. Imputation by monotone blocks: general setup

When the missing data pattern is monotone or nearly so, standard MICE takes no advantage of it and remains potentially incompatible. This motivates us to search for an imputation method that (1) at least approximately captures important relationships in the data, (2) reduces to standard joint modeling when it is appropriate, and (3) reduces to standard compatible sequential imputation when the missing data pattern is monotone. If the procedure is iterative, it is reasonable to require that, in addition, it converges under mild regularity conditions. Below we propose a general framework, “imputation by monotone blocks (IMB)”, attempting to combine the flexibility of MICE and the theoretical validity of the sequential imputation for monotone missing data. Here, a collection of (not necessarily compatible) conditional models are specified and the missing data are iteratively imputed and re-imputed based on these conditional models.

We first introduce two new concepts. A *monotone block* of missing data is a collection of missing entries in the data matrix that form a monotone missingness pattern, regarding the missing data outside the collection as observed. Formally, a monotone block B_k is represented by

- i. A putative monotone list of $J_k (\leq J)$ variables, i.e., an ordering of the variables $\{Y_{k_j}, j = 1, \dots, J_k\}$, where J_k is the total number of the variables that have missing entries in B_k , and $\{k_j, j = 1, \dots, J_k\}$ is a permutation of $\{1, \dots, J\}$; and
- ii. A specification of which missing entries belong to the block, with the requirement that if a missing entry (i, k_l) belongs to the block, then (i, k_p) , $l > p$ is missing and belongs to the block as well.

A *partition* of the missing data into monotone blocks is a collection of K mutually exclusive

monotone blocks B_1, \dots, B_K , such that,

$$M = \cup_{k=1}^K B_k, \quad \text{and} \quad B_k \cap B_l = \emptyset, \quad k \neq l.$$

One of the simplest partitions is to take missing entries in each variable as a monotone block, i.e., $K = J$ and $B_k = M_k (k = 1, \dots, J)$, a strategy used in MICE.

An IMB algorithm includes a modeling stage and an imputation stage. The modeling stage of consists of three components:

1. *Partitioning all missing entries into monotone blocks.*
2. *Specifying univariate conditional distributions sequentially within each monotone block,* regarding the most recent imputed values of the missing entries outside the block as “observed”. Suppose the list of variables in B_k is $\{Y_{k_j}, j = 1, \dots, J_k\}$ and denote the variables outside B_k by Y_{-B_k} . For each j , specify a distribution of Y_{k_j} conditional on $Y_{k_1}, \dots, Y_{k_{j-1}}$ and a subset of Y_{-B_k} . Variable selection procedures can be incorporated when the number of variables is large.
3. *Specifying the order of imputation within and between the monotone blocks.* An imputing order of the variables can be either deterministic or random. For example, an commonly adopted deterministic order is to impute the variables by the ascending order of missing proportion within a monotone block, i.e., the variable with the least missing entries is imputed first; and to impute the blocks by the ascending order of the block-wise total number of missing entries between monotone blocks.

In the imputation stage of IMB, after the initial imputation of all the missing data, e.g., by mean of or random draws from the observed marginal distributions, one iteratively cycles through the variables and the blocks according to the pre-specified imputing order as follows: (1) for each variable in each monotone block, fit its specified conditional model given both

the observed data and the most recently imputed missing data outside the block; (2) then impute its missing data in the block by sampling from their posterior predictive distributions. Iterates steps 1 and 2 until certain criterion of convergence is reached.

Because the missing pattern in each B_k is monotone, the modeling and imputing steps within each block are exactly the same as the sequential imputation for monotone missing data, which converges in one iteration. However, since there are more than one monotone block, more than one iterations is needed to ensure the final imputation to be stable.

3.3. Examples of IMB

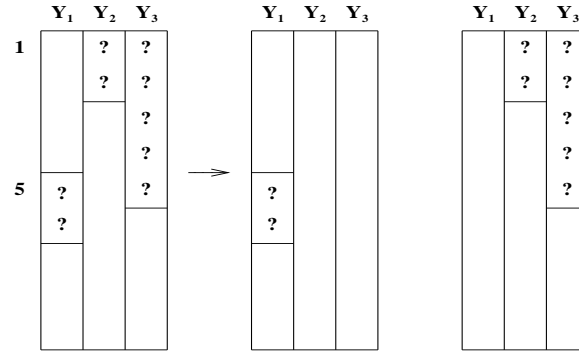
IMB defines a wide spectrum of imputation strategies, depending on the partition of the monotone blocks. As mentioned before, MICE is a special case with a simplest partition of the missing data. Another existing special case is the “Imputation by Major Monotone Block” strategy (IMMB), proposed by Rubin (2003) in the context of imputing the National Medical Expenditure Survey. In IMMB, the partition of missing data consists of: the major monotone block B_0 - a monotone block that includes as many missing entries as possible, and the blocks B_j ($j = 1, \dots, J$) that comprise the remaining missing entries in each variable j . After initial imputation, one iterate between (1) imputing the blocks B_1, \dots, B_J by the ascending order of the number of missing entries, and (2) imputing the major block B_0 by sequential imputation. Since each B_i ($i \neq 0$) contains only one variable, the first step is exactly an iteration in MICE.

Figure 6 illustrates the IMMB strategy with 3 variables. Comparing to the MICE, IMMB has the same flexibility and also takes advantage of possible monotone missingness. If the overall missing data pattern is already monotone, IMMB is reduced to the principled sequential imputation.

Furthermore, we introduce an additional example of IMB - *imputation by ordered monotone blocks (IOMB)*:

IOMB Partition the missing data by $M = \cup_{k=0}^K B_k$ as follows: First sort the data matrix

Figure 6: IMMB: The missing data are partitioned into 2 blocks. At each iteration, to update the block in the left, we fit a regression model of Y_1^{obs} given Y_2, Y_3 , and then impute Y_1^{mis} from its predictive distribution; to update the major block in the right, we first impute Y_2^{mis} conditional on Y_1 and Y_3^{obs} , and then impute Y_3^{mis} conditional on Y_1, Y_2 and Y_3^{obs} .



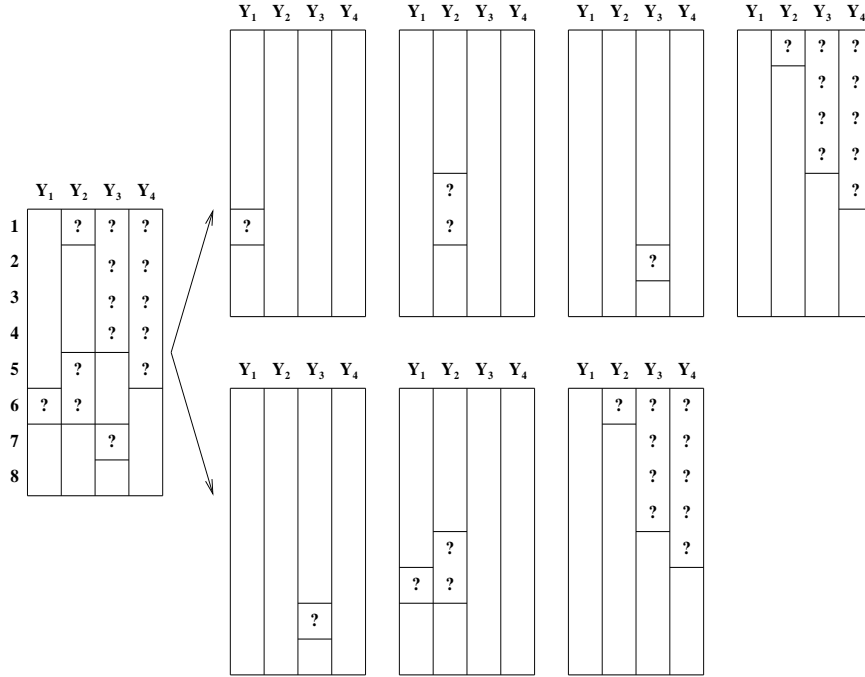
to obtain B_0 , the major monotone block; then sort the data matrix again, treating entries in B_0 as observed, and obtain the second largest monotone block B_1 ; and then recursively sort the data matrix to obtain monotone blocks B_2, \dots, B_K , until all missing entries have been accounted for.

The IOMB strategy can be viewed as an extension of the IMMB strategy. A comparison between IOMB and IMMB with three variables is illustrated in Figure 7. IOMB has been applied to impute the missing data in the Anthrax Vaccine Research Program for the Centers for Disease Control and Prevention in Li et al. (2011), where the implementation details of IOMB (e.g., how to partition an arbitrary missing data pattern into monotone blocks, the specification of conditional models for various types of variables) are extensively discussed.

We now re-examine Example 3 in Section 3.4 to show the potential benefit of IMMB and IOMB strategies over MICE when the conditional distributions are incompatible. The missing data partition of an IMMB strategy is given in the right side of Figure 4: B_0 consists of units 4,5 of Y_1 and 4,5,6,8 of Y_2 , and B_1 consists of unit 2 of Y_1 . In this case, since the B_1 only contains one variable, IMMB and the IOMB with largest B_0 strategies are equivalent.

IOMB iterates between: (1) in block B_0 , first impute the missing Y_1 's by sampling from

Figure 7: The IMMB and IOMB strategies. Top right: the IMMB strategy, which partitions the missing data into 4 monotone blocks; bottom right: the IOMB strategy, which partitions the missing data into 3 blocks.



its marginal distribution, and then impute Y_2 by the predictions from fitting $Y_2|Y_1^2$; and (2) in block B_1 , impute Y_1 from fitting $Y_1|Y_2$ to all data except for block 2, treating the updated imputed values in B_0 as observed. The standardize error in mean of the imputations from IOMB are shown in the right two columns of Table 2. Unlike MICE, the errors from IOMB stabilize over iterations. The key reason is that the model misspecification of $Y_2 \sim Y_1^2$, by construction, does not induce incompatibility in IOMB. Specifically, in B_0 , IOMB always first draws Y_1 from its marginal distribution, independent of Y_2 . Thus, the imputation error will not be propagated.

This simple example highlights the following advantages of IMMB and IOMB over the standard MICE: (1) IOMB/IMMB can avoid error propagation under model misspecification (thus resulting incompatibility); (2) it reduces the possibility of misspecification; and (3) it

reduces the possibility of overfitting when the number of variables is big. The last two are due to the fact that the conditional models within a monotone block always involve less or equal number of covariates than that in MICE.

3.4. Guidelines for choosing between IMB strategies

All IMB algorithms, like MICE, iteratively impute data based on a collection of not necessarily compatible conditional models, thus generally still have the problem of PIGS. However, within this general framework, one has the flexibility to choose the imputation strategies that reduce incompatibility as much as possible. Intuitively, all of the above combined strategies attempt to reduce the amount of incompatibility, a concept itself needed to be defined and quantified to assist meaningful comparison among IMB strategies. Here we formally introduce an *incompatibility measure*, M_{inc} , as *the minimal cardinality of any set of missing entries such that, when we fix the values of these entries, the imputation algorithm always has compatible conditional specifications, under fixed parameter imputation*.

In this definition we assume fixed parameter imputation, that is, the conditional models specified by MICE or the combined strategy have no free parameters. This focuses the algorithms at the unit level, because missing data for different units are imputed independently. Note that fixed parameter imputation is introduced to help clarify a theoretical discussion; it is by no means the recommended procedure to use in practice.

As an intuitive measure of incompatibility, M_{inc} is simple to calculate, and helps to facilitate a quantitative comparison between imputation strategies. For example, in Example 3, $M_{inc}^{mice} = 2$, $M_{inc}^{immb} = M_{inc}^{ismb} = 0$, indicating the IMMB/IOMB strategy has less combinatorial incompatibility than MICE. As another example, in the missingness pattern depicted by Figure 1 and Figure 6, MICE has $M_{inc}^{mice} = 3$, whereas the IMMB strategy has $M_{inc}^{immb} = 1$ (we only need to fill in entry y_{51} for any fixed-parameter conditional specification to be compatible). Generally, it is easy to show the following results.

Result 2. M_{inc} of MICE is the total number of missing entries minus the total number of units that have missing entries; $M_{inc}^{mice} = 0$ if and only if each unit has at most one variable missing.

Result 3. M_{inc} of IMMB $M_{inc}^{immb} \leq M_{inc}^{mice}$ for the same missing data.

Just as IMMB improves over MICE in terms of the incompatibility measure M_{inc} , we can prove by induction the IOMB strategy improves over the IMMB strategy in terms of M_{inc} . For example, in Figure 7, IMMB has $M_{inc}^{immb} = 2$, whereas its recursive generalization IOMB has $M_{inc}^{ismb} = 1$. Generally, we have:

Result 4. Suppose IOMB chooses the same major monotone block B_0 as IMMB, then $M_{inc}^{ismb} \leq M_{inc}^{immb}$.

The advantage of having a smaller M_{inc} can be viewed from the following perspective. Observe that, if a unit has more than two entries missing, then based on the examples in Section 2, imputations generated by MICE for this unit may be meaningless. The IMMB strategy has a smaller M_{inc} partly because it can group these missing entries into B_0 , and thereby eliminate such potential danger for this unit.

As an example, in Figure 1, unit 1 has this problem: imputing for this unit by MICE iterates between (a) $y_{12}|(y_{13}, y_{11})$, and (b) $y_{13}|(y_{12}, y_{11})$, and when these two conditional distributions are incompatible, imputed values of y_{12} and y_{13} cannot be relied upon. In Figure 6, using the combined strategy, unit 5 has a problem, but unit 1 does not, because both missing entries of unit 1 belong to the major block B_0 .

One drawback of M_{inc} is that it is purely combinatorial and does not reflect the actual conditional distributions specified by the algorithm. In fact, if the conditional distributions are derived from a legitimate joint model, then MICE or IMMB is automatically compatible, even though M_{inc} may not be zero. Another drawback is that $M_{inc} = 0$ does not imply a row-exchangeable complete-data model for all units, even when we assume fixed parameter imputation. Consider the following example. Let $J = 2$, and assume y_{11} and y_{22} are miss-

ing but all other entries of the data matrix are observed. Use MICE and specify the fully conditional models as follows:

$$Y_1|Y_2 \sim N(Y_2^2, 1) \quad \text{and} \quad Y_2|Y_1 \sim N(Y_1^2, 1).$$

Notice that we fix the parameters in these conditional models. Because of the special missing data pattern, the conditional specifications derived for $Y^{mis} = (y_{11}, y_{22})$ are actually compatible: one imputes y_{11} by drawing from $y_{11} \sim N(y_{12}^2, 1)$, and the other imputes y_{22} by drawing from $y_{22} \sim N(y_{21}^2, 1)$. However, clearly there is no row-exchangeable complete-data model that accommodates these two full conditionals.

Despite these drawbacks, M_{inc} has a simple interpretation and captures the combinatorial aspect of the problem. As a guideline, we propose to choose an IMB strategy with a small M_{inc} . In the extreme case when $M_{inc} = 0$, each unit is imputed under a single model rather than two or more incompatible conditional models. Although there may still be no row-exchangeable joint model for the underlying complete data for all units, IMB as an algorithm (assuming fixed parameter imputation) avoids the technical problems of PIGS.

Although we recommend using a strategy with a small M_{inc} , there are practical constraints on how small M_{inc} can be. Typically, the computing cost, for which the number of univariate regressions (URs) fitted per iteration is an approximate measure, is larger for algorithms with smaller M_{inc} . In the missing data pattern depicted in Figure 7, MICE has $M_{inc}^{mice} = 7$, and has to run four URs per iteration, one for each variable; the IMMB strategy has $M_{inc}^{immb} = 2$, and has to run 6 URs; the IOMB strategy has $M_{inc}^{ismb} = 1$, and also has to run 6 URs. $M_{inc} = 0$ does not necessarily imply a compatible MICE either, because M_{inc} is defined assuming fixed parameter imputation; in practice, when parameters for the conditional models are estimated from the data, incompatibility can still remain due to the dependence between imputed data across units.

Another measure of compatibility is a vector of ratios, $\mathbf{R} = (R_0, \dots, R_K)$, where $R_k =$

$|B_k|/|\cup_{l=0}^K B_l|$ is the ratio of the number of missing entries in the k th largest monotone block to the total number of missing entries (here B_k is ordered descendingly by its size). For an overall monotone missing pattern: $R = R_0 = 1$. Intuitively, the larger first R_k 's are, the closer the missing pattern is to an overall monotone pattern, thus has less potential incompatibility. Meanwhile, the computational burden of an IMB strategy is reflected by the total of the number of variables in each block, $L = \sum_{k=0}^K J_k$, which is the number of URs needed to be fitted per iteration. Smaller L corresponds to less computational burden. There is usually a tradeoff between R and L . For example, MICE usually has the smallest L , but also the smallest R_0 . We generally recommend to choose an IMB strategy with either large first R (s), or small L , or a combination of both. In fact, the IMMB strategy is designed to solely maximize R_0 , while the IOMB strategy is designed to both maximize R_0 and minimize K . Unlike the M_{inc} , the measure R does not assume fixed parameter imputation and it has a well defined upper bound of compatibility ($R_0 = 1$). But as M_{inc} , it is also purely combinatorial. The simple measures M_{inc} and R are certainly not adequate to represent the whole picture of imputation strategies based on fully conditional models. Nevertheless they provide an useful overall combinatorial assessment of (in)compatibility between the standard MICE strategy and the monotone blocks based strategies.

4. DISCUSSION

We examine some of the theoretical and algorithmic properties of the widely used imputation strategy MICE that is based on fully conditional models. As an algorithm, MICE is a possibly incompatible Gibbs sampler. Using simple examples, we show that MICE may not converge or generate meaningless imputations due to incompatibility between the conditional specifications. Aiming at retaining the flexibility of MICE and mitigating the problem of incompatibility, we propose a general imputation framework, IMB, utilizing the theoretically justified sequential imputation for monotone missing data pattern. Even though in general the

IMB strategies are also PIGS, it offers the possibility for users to find strategies that reduce incompatibility as much as possible within this framework. We define two combinatorial measures of (in)compatibility and provide guidelines to choose among the IMB strategies. In particular, we discussed two IMB strategies, IMMB and IOMB that show some potential in real implementation.

Our proposal by no means solves the incompatibility problem of the fully conditional models based imputation approaches. By explicitly pointing out the potential problems of the existing approaches and providing some preliminary solutions, the main purpose of this article is to attract more research in this important yet little-understood topic. For example, general measures of incompatibility of the IMB strategies that take into account model specification and computational cost deserve extensive further research. Moreover, we show that model specification is crucial for producing sensible imputations from MICE and other IMB strategies. Model selection and diagnostics with a large number of variables that are common in large surveys is a nontrivial task. The proposal of applying target analyses to posterior replicates of complete data (He and Zaslavsky, 2012) provides a promising approach for diagnosing imputation models. Flexible models, such as Bayesian nonparametric models, may be incorporated into the MICE machinery to improve model specification. But scalability of these models in large data sets can be an issue. Therefore, developing approaches that balance model flexibility and computational cost is key to improve the conditional models based imputation strategies.

Another attractive proposal to combine the advantages of MICE and sequential imputation for monotone missing data is based on the “conditional-conditional specification” introduced by Lipsitz and Ibrahim (1996), where for any missing data pattern, one specifies a series of conditional models for variables with missing data as follows: $Y_1, Y_2|Y_1, Y_3|(Y_2, Y_1), \dots, Y_J|(Y_{J-1}, \dots, Y_1)$, and then estimate the parameters and impute the missing values iteratively as in MICE. By construction, this ensures joint distribution exists for (Y_1, \dots, Y_J) . When missing data and data

analysis are considered simultaneously (as in Lipsitz and Ibrahim, 1996), this approach is theoretically valid. However, when used solely to impute missing data, this algorithm with different order of specification may lead to completely different joint distributions and resulting imputations.

Finally, IMB strategies, including MICE, all assume the missing-data mechanism is ignorable (missing at random). When the missing-data mechanism is non-ignorable, how to combine the fully conditional models based approaches with pattern-mixture models (Little, 1993) or selection models to take into account the missing-data mechanism in a principled fashion requires systematic investigation.

APPENDIX

The Gaussian and Linear PIGS

Example 1 shown in Section 2 is a Gaussian and linear PIGS. Studying PIGS in this setting is a starting point for further investigation, since the Gaussian and linear case is perhaps the simplest non-trivial PIGS whose convergence behavior is mathematically tractable.

Definition A PIGS is called linear if the full conditionals are specified by $f_j(z_j|z_{-j})$ such that

$$E_j(z_j|z_{-j}) = \alpha_j + \sum_{l \neq j} \beta_{jl} z_l, \text{ and } Var_j(z_j|z_{-j}) = \sigma^2,$$

where E_j and Var_j are the conditional mean and conditional variance operators with respect to the density $f_j(z_j|z_{-j})$.

Here we restrict our attention to univariate full conditionals and without loss of generality the conditional variances are set equal.

Definition A linear PIGS is called Gaussian and linear if the full conditionals are

$$z_j | z_{-j} \sim N(\alpha_j + \sum_{l \neq j} \beta_{jl} z_l, \sigma^2), \quad j = 1, \dots, J.$$

Let A be the $J \times J$ matrix with ones in the diagonals and $-\beta_{jl}$ in the (j, l) th entry, $j \neq l$. We shall call A/σ^2 the *concentration matrix* of the linear PIGS. Denote $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_J)^T$. We shall call α the *displacement vector*.

Proposition 4.1 For the Gaussian and linear PIGS, a necessary and sufficient condition for these full conditional distributions to be compatible is A being symmetric and positive definite (see Liu 1999).

Remark: When these full conditionals are compatible, PIGS reduces to the usual Gibbs sampler, and A/σ^2 is the concentration matrix of the invariant distribution, which is multivariate normal.

Roberts and Sahu (1997) investigate convergence properties of the Gibbs sampler for multivariate Gaussian densities (see also Amit 1991). Their results can be modified in a straightforward manner for the PIGS.

Proposition 4.2 Let (z_1, z_2, \dots, z_J) be a linear PIGS with concentration matrix A/σ^2 and displacement vector α . Let L and U be the (strictly) lower and upper triangular parts of A , respectively. Then

$$E(z^{(t+1)} | z^{(t)}) = (I + L)^{-1}(\alpha - Uz^{(t)}),$$

$$Var(z^{(t+1)} | z^{(t)}) = \sigma^2 \{(I + L)^T (I + L)\}^{-1},$$

where $z^{(t)}$ is the vector $(z_1, \dots, z_J)^T$ for iteration t and I is the identity matrix of order J .

Proof: Denote $w_j = z_j^{(t+1)} - \sum_{k=1}^{j-1} \beta_{jk} z_k^{(t+1)}$, $j = 1, \dots, J$. From the structure of the PIGS

we notice

$$E(w_j|z^{(t)}, z_k^{(t+1)}, 1 \leq k \leq j-1) = \alpha_j + \sum_{k=j+1}^J \beta_{jk} z_k^{(t)}.$$

Hence

$$E(w_j|z^{(t)}) = \alpha_j + \sum_{k=j+1}^J \beta_{jk} z_k^{(t)}.$$

Also,

$$\text{Var}(w_j|z^{(t)}, z_k^{(t+1)}, 1 \leq k \leq j-1) = \sigma^2.$$

$$\begin{aligned} \text{Var}(w_j|z^{(t)}) &= E\{\text{Var}(w_j|z^{(t)}, z_k^{(t+1)}, 1 \leq k \leq j-1)|z^{(t)}\} \\ &\quad + \text{Var}\{E(w_j|z^{(t)}, z_k^{(t+1)}, 1 \leq k \leq j-1)|z^{(t)}\} \\ &= \sigma^2 + 0 = \sigma^2. \end{aligned}$$

Furthermore, for $l < j$,

$$\begin{aligned} E(w_l w_j|z^{(t)}) &= E\{w_l E(w_j|z^{(t)}, z_k^{(t+1)}, 1 \leq k \leq j-1)|z^{(t)}\} \\ &= E\{w_l E(w_j|z^{(t)})|z^{(t)}\} \\ &= E(w_l|z^{(t)})E(w_j|z^{(t)}). \end{aligned}$$

Therefore w_l and w_j are conditionally uncorrelated given $z^{(t)}$. Notice that w_j is just the j th element of $(I + L)z^{(t+1)}$. Hence

$$\begin{aligned} E\{(I + L)z^{(t+1)}|z^{(t)}\} &= \alpha - Uz^{(t)}, \\ \text{Var}\{(I + L)z^{(t+1)}|z^{(t)}\} &= \sigma^2 I. \end{aligned}$$

The claims then follow.

Theorem 4.1 *Under the setting of Proposition 4.2, assume in addition that the PIGS is Gaus-*

sian and linear. Denote $B = -(I + L)^{-1}U$ and $C = \sigma^2\{(I + L)^T(I + L)\}^{-1}$. Then the PIGS converges iff $\rho(B) < 1$, where $\rho(B)$ is the spectral radius of B . When convergent, it converges to $N(\gamma, D)$, where $\gamma = A^{-1}\alpha$, and D satisfies $D = C + BDB^T = \sum_{k=0}^{\infty} B^k C (B^T)^k$.

Proof: The conditional distribution of $z^{(t+1)}|z^{(t)}$ is obviously multivariate normal. By Proposition 4.2, it is given by

$$z^{(t+1)}|z^{(t)} \sim N((I + L)^{-1}\alpha + Bz^{(t)}, C).$$

Therefore $z^{(t)}$ is a multivariate $AR(1)$ process, and the necessary and sufficient condition for convergence is $\rho(B) < 1$. The target distribution is clearly normal, say $N(\gamma, D)$, with $\gamma = (I - B)^{-1}(I + L)^{-1}\alpha$, and $D = C + BDB^T$. Simple calculation yields $\gamma = A^{-1}\alpha$.

Remark: 1. Since we are dealing with Gaussian processes, it does not matter very much which convergence criterion we use. To be definitive, we consider convergence in total variation norm, i.e., densities $\pi_k(\cdot)$ converging to $\pi(\cdot)$ means $\int |\pi_k(z) - \pi(z)| dz \rightarrow 0, k \rightarrow \infty$.

2. When $\sigma^2 \rightarrow 0$, PIGS reduces to the well-known Gauss-Seidel iteration (see Golub and Van Loan, 1996) for solving the linear system of equations $Az = \alpha$. Theorem 4.1 therefore slightly generalizes the corresponding result for the Gauss-Seidel iteration. Since the conditions for convergence do not depend on σ^2 , results on the Gauss-Seidel iteration can be used for the PIGS. For example, both of them converge if one of the following holds:

1. A is symmetric and positive definite.
2. A is strictly diagonally dominant, i.e., $\sum_{l \neq j} |\beta_{jl}| < 1, j = 1, \dots, J$.

As an application of Theorem 4.1, we analyze the convergence behavior of Example 1 constructed in Section 2. In this example, $J = 3$ and the concentration matrix for update

order $[z_1, z_2, z_3]$ is

$$A = \begin{pmatrix} 1 & 1.5 & 0.5 \\ 0.5 & 1 & 0.5 \\ 1.5 & 1.5 & 1 \end{pmatrix}.$$

Then $\rho(B) = 1.0607 > 1$, so this update order fails to converge. For $[z_2, z_1, z_3]$, the corresponding matrix is now

$$A_* = \begin{pmatrix} 1 & 0.5 & 0.5 \\ 1.5 & 1 & 0.5 \\ 1.5 & 1.5 & 1 \end{pmatrix}.$$

But $\rho(B_*) = 0.6124 < 1$, so this update order converges.

Proof of Result 1

We need the following result from denumerable Markov chain theory (and tedious calculations).

Theorem 4.2 *Let $z^{(t)}$ be an irreducible denumerable Markov chain whose states are grouped in regular blocks of size m , with block labels $k = 0, 1, \dots$. Denote the j th state of block k by (k, j) . Suppose transition occurs only within blocks or between adjacent blocks. Suppose for large enough k transition probabilities can be specified by $m \times m$ matrices P, Q, R , with*

$$\begin{aligned} P_{jl} &= \Pr\{z^{(t+1)} = (k+1, l) | z^{(t)} = (k, j)\}, \\ Q_{jl} &= \Pr\{z^{(t+1)} = (k, l) | z^{(t)} = (k+1, j)\}, \\ R_{jl} &= \Pr\{z^{(t+1)} = (k, l) | z^{(t)} = (k, j)\}. \end{aligned}$$

Let S be the minimal non-negative solution of the matrix equation

$$S = P + SR + S^2Q.$$

Then (a) S is finite and $\rho(S) \leq 1$; (b) the chain is recurrent iff $Q + R + SQ$ is a stochastic matrix; (c) the chain is positive recurrent iff $\rho(S) < 1$.

Group the states into blocks such that block k is $\{(z_1, z_2, z_3) : z_1 = 0, 1; z_2 = 3k, 3k + 1, 3k + 2; |z_3 - z_2| \leq 1\}$. It is easy to see that with this grouping, all deterministic and random orders for both examples satisfy conditions of Theorem 4.2, i.e., transitions only occur between states within the same block or between adjacent blocks. After some arithmetic, the matrices P, Q, R are determined, and $\rho(S)$ as well as $Q + R + SQ$ are computed numerically. The recurrent/transient status for each case is then easily determined by Theorem 4.2.

REFERENCES

- [1] Amit, Y. (1991). On rates of convergence of stochastic relaxation for gaussian and non-gaussian distributions. *Journal of Multivariate Analysis* **38**, 82–99.
- [2] Arnold, B. C. and Press, S. J. (1989). Compatible conditional distributions. *Journal of the American Statistical Association* **84**, 152–156.
- [3] Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of Royal Statistical Society, Series B* **36**, 192–236.
- [4] Brand, J. P. L. (1999). *Development, Implementation, and Evaluation of Multiple Imputation Strategies for the Statistical Analysis of Incomplete Data Sets*. Ph.D. Dissertation, Erasmus University, Rotterdam.
- [5] Faris, P. D., Ghali, W. A., Brant, R., Norris, C. M., Galbraith, P. D. and Knudtson, M. L. (2002). Multiple imputation versus data enhancement for dealing with missing data in observational health care outcome analyses. *Journal of Clinical Epidemiology* **55**, 184–191.
- [6] Gelfand, A.E. and Smith, A.F.M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410), 398–409.
- [7] Gelman, A. and Raghunathan, T. E. (2001). Discussion of Arnold et al. “Conditionally specified distributions”. *Statistical Science* **16**, 249–274.

- [8] Golub, G. and Van Loan, C. (1996). *Matrix Computations*. Johns Hopkins University Press, 3rd Edition.
- [9] He, Y., Zaslavsky, A.M., Harrington, D.P., Catalano, P. and Landrum, M.B. (2010). Multiple imputation for a large-scale complex survey: a practical guide. *Statistical Method in Medical Research* **19(6)**, 653-670.
- [10] He, Y. and Zaslavsky, A.M. (2012). Diagnosing imputation models by applying target analyses to posterior replicates of complete data. *Statistics in Medicine*. **31**, 1-18.
- [11] Heeringa, S. G., Little, R. J. A. and Raghunathan, T. E. (2002). Multivariate imputation of coarsened survey data on household wealth. In R. M. Groves et al. (Eds.), *Survey Nonresponse*. New York: Wiley, 357–371.
- [12] Hobert, J. P. and Casella, G. (1998). Functional compatibility, Markov chains and Gibbs sampling with improper posteriors. *Journal of Computational and Graphical Statistics* **7**, 42–66.
- [13] Javaras, K. N. and van Dyk, D. A. (2003). Multiple imputation for incomplete data with semi-continuous variables. *Journal of the American Statistical Association* **98**, 703–715.
- [14] Kennickell, A. B. (1991). Imputation of the 1989 survey of consumer finances: stochastic relaxation and multiple imputation. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 1–10.
- [15] Kennickell, A. B. (1999). Multiple imputation and disclosure control: the case of the 1995 survey of consumer finances. In *Record Linkage Techniques 1997*, Washington, DC: National Academy Press, 248–267.
- [16] Li, F., Baccini, M., Mealli, F., Frangakis, C. F., Rubin, D. B. and Zell, E. R. (2011). Multiple imputation by ordered monotone blocks with application to the Anthrax Vaccine trial. *Duke University Department of Statistical Science Discussion Paper 11-26*.
- [17] Lipsitz, S.R., and Ibrahim, J.G. (1996). Conditional Model for Incomplete Covariates in Parametric Regression Models. *Biometrika* **83**, 916-922.
- [18] Little, R. J. A. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the*

American Statistical Association **88**, 125–134.

- [19] Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd Ed. John Wiley & Sons, Inc. New York.
- [20] Liu, C. (1999). Compatibility conditions for a set of conditional Gaussian distributions. *Statistics & Probability Letters* **42**, 127–130.
- [21] 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.
- [22] Meng, X-L. (1994). Multiple-Imputation Inferences with Uncongenial Sources of Input. *Statistical Science* **9(4)**, 538-558.
- [23] Oudshoorn, C. G. M., van Buuren, S. and van Rijckeversel, J. L. A. (1999). *Flexible Multiple Imputation by Chained Equations of the AVO-95 Survey*. Leiden: TNO Prevention and Health. Report PG/VGZ/99.045.
- [24] Raghunathan, T. E. and Siscovick, D. S. (1996). A multiple imputation analysis of a case-control study of the risk of primary cardiac arrest among pharmacologically treated hypertensives. *Applied Statistics* **45**, 335–352.
- [25] Raghunathan, T. E., Solenberger, P. and van Hoewyk, J. (2000). *IVEware: Imputation and Variance Estimation Software: Installation Instructions and User Guide*. Survey Research Center, Institute of Social Research, University of Michigan.
- [26] Roberts, G. O. and Sahu, S. K. (1997). Updating schemes, correlation structure, blocking and parameterisation for the Gibbs sampler. *Journal of Royal Statistical Society, Series B* **59**, 291–317.
- [27] Rubin, D. B. (1974). Characterizing the estimation of parameters in incomplete data problems. *Journal of the American Statistical Association* **69**, 467–474.
- [28] Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63**, 581–592.
- [29] Rubin, D. B. (1978). Multiple imputations in sample surveys. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 20–34.

- [30] Rubin, D. B. (1996). Multiple imputation after 18+ years (with discussion). *Journal of the American Statistical Association* **91**, 473–517.
- [31] Rubin, D. B. (2003). Nested multiple imputation of NMES via partially incompatible MCMC. *Statistica Neerlandica* **57**, 3–18.
- [32] Rubin, D. B. (2004). *Multiple Imputation for Nonresponse in Surveys*, 2nd Ed. John Wiley & Sons, Inc. New York.
- [33] SAS Institute Inc. (2008). SAS 9.2 Reference. Cary, NC: SAS Institute Inc.
- [34] Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. London: Chapman and Hall.
- [35] Schenker, N., Raghunathan, T. E., Chiu, P. L., Makuc, D. M., Zhang, G., and Cohen, A. J. (2006). Multiple imputation of missing income data in the National Health Interview Survey. *Journal of the American Statistical Association* **101**, 924-933.
- [36] Statistical Solutions Ltd. (2001). SOLAS for Missing Data Analysis. Statistical Solutions, Cork, Ireland.
- [37] Stuart, E.A., Azur, M., Frangakis, C.E., and Leaf, P. (2009). Multiple imputation with large datasets: A case study of the Children’s Mental Health Initiative. *American Journal of Epidemiology* **169(9)**, 1133-1139.
- [38] Van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research* **16(3)**, 219-242.
- [39] Van Buuren, S., Boshuizen, H. C. and Knook, D. L. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine* **18**, 681–694.
- [40] Van Buuren, S., Brand, J. P. L., Oudshoorn, C. G. M. and Rubin D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation* **76**, 1049–1064.
- [41] Van Buuren, S. and Oudshoorn C. G. M. (2000). *Multivariate Imputation by Chained Equations: MICE VI.0 User’s Manual*. Report PG/VGZ/00.038. Leiden: TNO Preventie en Gezondheid.