Biometrika, , , *pp*. 1–18

Printed in Great Britain

On the Stationary Distribution of Iterative Imputations

BY JINGCHEN LIU, ANDREW GELMAN
Department of Statistics, Columbia University, 1255 Amsterdam Ave, New York, 10027, U.S.A.
jcliu@stat.columbia.edu gelman@stat.columbia.edu

JENNIFER HILL
School of Education, New York University, 246 Greene St, New York 10003, U.S.A.
jennifer.hill@nyu.edu

YU-SUNG SU
Department of Political Science, Tsinghua University, Beijing 100084, China
suyusung@tsinghua.edu.cn

JONATHAN KROPKO
Department of Statistics, Columbia University, 1255 Amsterdam Ave, New York, 10027, U.S.A.
jmk2210@columbia.edu

SUMMARY

Iterative imputation, in which variables are imputed one at a time conditional on all the others, is a popular technique that can be convenient and flexible, as it replaces a potentially difficult multivariate modeling problem with relatively simple univariate regressions. In this paper, we begin to characterize the stationary distributions of iterative imputations and their statistical properties, accounting for the conditional models being iteratively estimated from data rather than being pre-specified. When the families of conditional models are compatible, we provide sufficient conditions under which the imputation distribution converges in total variation to the posterior distribution of a Bayesian model. When the conditional models are incompatible but valid, we show that the combined imputation estimator is consistent.

Some key words: chained equation; convergence; iterative imputation; Markov chain.

1. INTRODUCTION

Iterative imputation is widely used for imputing multivariate missing data. The procedure starts by randomly imputing missing values using some simple stochastic algorithm. Missing values are then imputed one variable at a time, each conditionally on all the others using a model fit to the current iteration of the completed data. The variables are looped through until approximate convergence that is measured, for example, by the mixing of multiple chains.

With iterative imputation, there is no need to explicitly construct a joint multivariate model of all types of variables: continuous, ordinal, categorical, and so forth. Instead, one only needs to specify a sequence of families of conditional models such as linear regression, logistic regression, and other standard and already programmed forms. The distribution of the resulting imputations

is implicitly defined as the stationary distribution of the Markov chain corresponding to the iterative fitting and imputation process.

Iterative, or chained, imputation is convenient and flexible and has been implemented in various ways in several statistical software packages, including mice and mi in R, IVEware in SAS, and ice in Stata; see van Buuren & Groothuis-Oudshoorn (2011), Su et al. (2011), Raghunathan et al. (2010), and Royston (2004, 2005). The popularity of these programs suggests that the resulting imputations are believed to be of practical value. However, the theoretical properties of iterative imputation algorithms are not well understood. Even if the fitting of each conditional model and the imputations themselves are performed using Bayesian inference, the stationary distribution of the algorithm, if it exists, does not in general correspond to Bayesian inference on any specified multivariate distribution.

Iterative imputation poses several key questions. Under what conditions does the algorithm converge to a stationary distribution? What statistical properties does the procedure admit given that a unique stationary distribution exists? For the first question, researchers have long known that the Markov chain may be non-recurrent, even if each of the conditional models is fitted using a proper prior distribution. For example, the chain may blow up to infinity or drift like a nonstationary random walk. In this paper, we assume that the unique stationary distribution of the iterative imputation exists and focus on the second question, the characterization of this distribution.

The analysis of iterative imputation is challenging for at least two reasons. First, the range of choices of conditional models is wide, and it would be difficult to provide a solution applicable to all situations. Second, the distributions for the imputations are known only within specified parametric families. For example, if a particular variable is to be updated conditional on all the others using logistic regression, the actual updating distribution depends on the logistic regression coefficients, which are themselves estimated given the latest update of the missing values.

The contribution of this paper is a mathematical framework under which the asymptotic properties of iterative imputation can be discussed. In particular, we demonstrate the following results. First, for a positive Harris recurrent iterative imputation Markov chain whose unique stationary distribution exists, we provide a set of conditions under which this distribution converges in total variation to the posterior distribution of a joint Bayesian model, as the sample size tends to infinity. Under these conditions, iterative imputation is asymptotically equivalent to full Bayesian imputation using some joint model. This asymptotic result does not depend on the validity of the model, that is, the asymptotic equivalence holds when the model is misspecified. Among these conditions, the most important is that the conditional models are compatible, that is, that there exists a joint model whose conditional distributions are identical to the conditional models specified by the iterative imputation. Second, we consider model compatibility as a typically necessary condition for the iterative imputation distribution to converge to the posterior distribution of some Bayesian model. This is discussed in Section 3·4. Lastly, for incompatible models, imputation distributions are generally different from any Bayesian model, and we show that the combined completed-data maximum likelihood estimator of the iterative imputation is a consistent estimator if the set of conditional models are valid, that is, if each conditional family contains the true distribution.

The theoretical results have several practical implications. If the conditional models are compatible and all other regularity conditions are satisfied, then the imputation distribution of joint Bayesian models and that of the iterative procedure are asymptotically the same. Thus, with certain moment conditions, Rubin's rules for combining imputed data sets (Little & Rubin, 2002) are applicable. For incompatible models, the combined point estimators are consistent as long as each conditional model is correctly specified. For large scale data sets and in presence of multiple

types of variables, it is generally difficult to maintain or to check model compatibility. In fact, it is precisely in the situation when a joint model is difficult to obtain that iterative imputation is preferred. With incompatible models, the most important condition is the validity of the conditional models. In addition, given that the goal is predictive, one may employ more sophisticated procedures beyond generalized linear models, such as hierarchical modeling, model selection, penalized estimation, etc. From the modeling point of view, the imputers should try as much as possible to check and to achieve model compatibility. When compatibility is difficult to obtain, one should make an effort to improve the prediction quality of each conditional regression model. Lastly, one important element in the technical development is a bound on the convergence rate, which, in practice, corresponds to the mixing of iterative chains. Slow mixing rates for typical Markov chain Monte Carlo problems mean inefficient computation and often can be solved by running longer chains to reach the stationary distribution. In the context of iterative imputation, in addition to computational inefficiency, slow mixing also indicates potentially less accurate inferences. Such inaccuracies persist even if the chain has reached stationarity and thus slow mixing should raise a warning to the imputer. Therefore, we recommend that the imputer always check the convergence of the iterative chain. If the chain mixes slowly, one might consider a smaller data set by dropping some less important variables for the imputation.

The analysis presented in this paper connects to separate existing literatures on missing data imputation and Markov chain convergence. Standard references on imputation inference include Rubin (1987), Schafer (1997), Li et al. (1991), Barnard & Rubin (1999), Meng (1994), and Rubin (1996). Large sample properties are studied by Robins & Wang (2000), Schenker & Welsh (1988), and Wang & Robins (1998); congeniality between the imputer's and analyst's models is considered by Meng (1994). A framework of fractional imputation was proposed by Kim (2011). Iterative imputation is a procedure to which these and other results in the field do not apply. The current work provides theoretical backup of this procedure by means of theories and techniques developed for Markov chains that will be discussed in the following paragraph.

Our asymptotic findings for compatible and incompatible models use results on the convergence and coupling of Markov chains, a subject on which there is a vast literature concerning stability and rate of convergence (Amit & Grenander, 1991). For the analysis of compatible models, we need to construct a bound for the convergence rate using renewal theory, which has the advantage of not assuming the existence of an invariant distribution, which is naturally yielded by minorization and drift conditions (Baxendale, 2005; Meyn & Tweedie, 1993; Rosenthal, 1995). Some other related works include incompatible Gibbs samplers studied by van Dyk & Park (2008) and functional compatible Gibbs samplers studied in Hobert & Casella (1998).

2. BACKGROUND

2·1. *Bayesian modeling, imputation, and Gibbs sampling*

Consider a data set with n cases and p variables, where $X = (X_1, ..., X_p)$ represents the complete data and $X_j = (x_{1,j}, ..., x_{n,j})^{\mathrm{T}}$ is the jth variable including both the observed and the missing data. Let r_j be the vector of observed data indicators for variable j, equaling 1 for observed variables and 0 for missing variables. Let X_j^{obs} and X_j^{mis} be the observed and missing subsets of variable i and furthermore let $X^{\text{obs}} = \{X_j^{\text{obs}} : j = 1, ..., p\}$ and $X^{\text{mis}} = \{X_j^{\text{mis}} : j = 1, ..., p\}$. To facilitate our description of the procedures, we define

$$X_{-j}^{\text{obs}} = \{X_l^{\text{obs}} : l = 1, ..., j-1, j+1, ..., p\}, \quad X_{-j}^{\text{mis}} = \{X_l^{\text{mis}} : l = 1, ..., j-1, j+1, ..., p\}.$$

Notation that does not have superscripts, such as X_j and X_{-j}, includes both the observed and the imputed data, that is, $X_j = (X_j^{\text{obs}}, X_j^{\text{mis}})$ and $X_{-j} = (X_{-j}^{\text{obs}}, X_{-j}^{\text{mis}})$. We use X to denote the entire data set and x to denote individual observations, that is, x could be a row vector of X. Furthermore, we use x_j to denote the jth variable of one observation and x_{-j} to denote all the others.

We assume that the missing data process is ignorable throughout. One set of sufficient conditions for ignorability is that the r process is missing at random and the parameter spaces for the distributions of X and r given X are distinct and have independent prior distributions (Little & Rubin, 2002).

In Bayesian inference, imputed data sets are treated as samples from the posterior distribution of the incompletely observed data matrix. In the parametric Bayesian approach, one specifies a family of distributions $f(X \mid \theta)$ and a prior $\pi(\theta)$ and then performs inference using independently and identically distributed samples from the posterior predictive distribution

$$p(X^{\text{mis}} \mid X^{\text{obs}}) = \int_\Theta f(X^{\text{mis}} \mid X^{\text{obs}}, \theta) p(\theta \mid X^{\text{obs}}) \mathrm{d}\theta, \tag{1}$$

where $p(\theta \mid X) \propto \pi(\theta) f(X \mid \theta)$. Direct simulation from (1) is generally difficult. One standard solution is to draw approximate samples using the Gibbs sampler or some more complicated Markov chain Monte Carlo algorithm. In the scenario of missing data, one can use the data augmentation strategy to iteratively draw θ given $(X^{\text{obs}}, X^{\text{mis}})$ and X^{mis} given (X^{obs}, θ). Under regularity conditions such as positive recurrence, irreducibility, and aperiodicity, the Markov process is ergodic with limiting distribution $p(X^{\text{mis}}, \theta \mid X^{\text{obs}})$ (Geman & Geman, 1984).

To connect to the iterative imputation that is the subject of the present article, we consider a slightly different Gibbs scheme. Let $X(k-1)$ be the entire data set including both the observed data and the imputed data at iteration $k-1$. To evolve to $X(k)$, we need to update one variable at a time according to Algorithm 1.

Algorithm 1 Gibbs chain

Step 0. Set $X \leftarrow X(k-1)$ and update the variables of X one at a time.
Step 1. Draw $\theta \sim p(\theta \mid X_1^{\text{obs}}, X_{-1})$ and $X_1^{\text{mis}} \sim f(X_1^{\text{mis}} \mid X_1^{\text{obs}}, X_{-1}, \theta)$.
$\vdots$
Step p. Draw $\theta \sim p(\theta \mid X_p^{\text{obs}}, X_{-p})$ and $X_p^{\text{mis}} \sim f(X_p^{\text{mis}} \mid X_p^{\text{obs}}, X_{-p}, \theta)$.
Step $p+1$. Set $X(k) \leftarrow X$.

At each step, the posterior distribution is based on the updated values of the parameters and imputed data. It is not hard to verify that, under mild regularity conditions as in Rosenthal (1995), the Markov chain evolving according to Algorithm 1 converges to the posterior distribution of the corresponding Bayesian model.

2·2. *Iterative imputation and compatibility*

For iterative imputation, we need to specify p conditional models, $g_j(X_j \mid X_{-j}, \theta_j)$, for $\theta_j \in \Theta_j$ with prior distributions $\pi_j(\theta_j)$ $(j = 1, ..., p)$. When there is no ambiguity, we shall use g_j as a generic notation for the conditional model for variable j, and its meaning may differ slightly from place to place. For instance, we shall constantly use $g_1(X_1^{\text{mis}} \mid X_1^{\text{obs}}, X_{-1}, \theta_1)$ to refer to the conditional distribution of missing data X_1^{mis} given $(X_1^{\text{obs}}, X_{-1})$ and θ_1. Iterative imputation evolves from iteration $k-1$ to iteration k according to Algorithm 2.

Algorithm 2 Iterative chain

Step 0. Set $X \leftarrow X(k-1)$ and update the variables of X one at a time.
Step 1. Draw θ_1 from $p_1(\theta_1 \mid X_1^{\text{obs}}, X_{-1})$, which is the posterior distribution associated with g_1 and π_1.
Draw X_1^{mis} from $g_1(X_1^{\text{mis}} \mid X_1^{\text{obs}}, X_{-1}, \theta_1)$;
$\vdots$
Step p. Draw θ_p from $p_p(\theta_p \mid X_p^{\text{obs}}, X_{-p})$, which is the posterior distribution associated with g_p and π_p.
Draw X_p^{mis} from $g_p(X_p^{\text{mis}} \mid X_p^{\text{obs}}, X_{-p}, \theta_p)$.
Step $p+1$. Set $X(k) \leftarrow X$.

Iterative imputation has the practical advantage that, at each step, one only needs to set up a sensible regression model for X_j given X_{-j}. This substantially reduces the modeling task, given that there are usually standard linear or generalized linear models for univariate responses of different variable types. In contrast, full Bayesian or likelihood modeling requires the more difficult task of constructing a joint model for X. Whether it is preferable to perform p easy tasks or one difficult task depends on the problem at hand. All that is needed here is the recognition that, in some settings, users prefer the p easy steps of iterative imputation.

Iterative imputation has problems. In general there is no joint distribution of X such that $f(X_j \mid X_{-j}, \theta) = g_j(X_j \mid X_{-j}, \theta_j)$ for each j. In addition, it is unclear whether the Markov process has a probability invariant distribution; if there is such a distribution, it lacks characterization.

Iterative imputation has some variations. For example, the parameter θ_j for the conditional model g_j can be sampled from the posterior distribution given the complete data sets X where the missing values are updated from the previous step, in contrast to the current scheme in which the posterior distribution is conditional on $(X_j^{\text{obs}}, X_{-j})$. For this new scheme, we can construct a Gibbs sampler similarly as in Section 2·1 for the joint Bayesian model correspondingly: at each step, θ is sampled from the posterior, $p(\theta \mid X)$. This new sampler couples with the new iterative chain. The analysis strategy and results apply to the new iterative imputation schemes. Another situation is that the parameter θ_j is only updated conditional on the fully observed cases. For this procedure, we can only apply the results in Section 4, that is, consistency can be obtained under model validity. For the detailed development, we only consider the scheme for which the updates are conditional on $(X_j^{\text{obs}}, X_{-j})$. We study the stationary distribution of the iterative imputation by first classifying the set of conditional models as compatible or incompatible.

3. Compatible conditional models

3·1. *Model compatibility*

Analysis of iterative imputation is particularly challenging partly because of the large collection of possible choices of conditional models. We begin by considering a restricted class, compatible conditional models, defined as follows:

Definition 1. *A set of conditional models $\{g_j(x_j \mid x_{-j}, \theta_j) : \theta_j \in \Theta_j, j = 1, ..., p\}$ is said to be compatible if there exists a joint model $\{f(x \mid \theta) : \theta \in \Theta\}$ and a collection of surjective maps, $\{t_j : \Theta \to \Theta_j : j = 1, ..., p\}$ such that for each j, $\theta_j \in \Theta_j$, and $\theta \in t_j^{-1}(\theta_j) = \{\theta : t_j(\theta) = \theta_j\}$, $g_j(x_j \mid x_{-j}, \theta_j) = f(x_j \mid x_{-j}, \theta)$. Otherwise, $\{g_j : j = 1, ..., p\}$ is said to be incompatible.*

Throughout this paper, we assume that all the observations are independently and identically distributed and thus one only needs to specify the distribution of a single observation. Though imposing certain restrictions, compatible models do include quite a collection of procedures practically in use, for instance, ice in Stata. In what follows, we give a few examples of compatible and incompatible conditional models.

Example 1 *(bivariate Gaussian).* Consider a binary continuous variable (x, y) with

$$x \mid y \sim N(\alpha_{x|y} + \beta_{x|y} y, \tau_x^2), \quad y \mid x \sim N(\alpha_{y|x} + \beta_{y|x} x, \tau_y^2).$$

These two conditional models are compatible if and only if $(\beta_{x|y}, \beta_{y|x}, \tau_x, \tau_y)$ lie on a subspace determined from the joint model,

$$\begin{pmatrix} x \\ y \end{pmatrix} \sim N\left\{ \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \Sigma \right\}, \quad \Sigma = \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix},$$

with $\sigma_x, \sigma_y > 0$ and $-1 \le \rho \le 1$. The reparameterization is

$$t_1(\mu_x, \sigma_x^2, \mu_y, \sigma_y^2, \rho) = (\alpha_{x|y}, \beta_{x|y}, \tau_x^2) = \left\{ \mu_x - \frac{\rho\sigma_x}{\sigma_y}\mu_y, \frac{\rho\sigma_x}{\sigma_y}, (1-\rho^2)\sigma_x^2 \right\}$$
$$t_2(\mu_x, \sigma_x^2, \mu_y, \sigma_y^2, \rho) = (\alpha_{y|x}, \beta_{y|x}, \tau_y^2) = \left\{ \mu_y - \frac{\rho\sigma_y}{\sigma_x}\mu_x, \frac{\rho\sigma_y}{\sigma_x}, (1-\rho^2)\sigma_y^2 \right\}.$$

Example 2 *(continuous data).* Consider a set of conditional linear models: for each j,

$$x_j \mid x_{-j}, \beta_j, \sigma_j^2 \sim N\left\{ (1, x_{-j})\beta_j, \sigma_j^2 \right\},$$

where β_j is a $p \times 1$ vector. Consider the joint model $(x_1, ..., x_p) \sim N(\mu, \Sigma)$. Then the conditional distribution of each x_j given x_{-j} is Gaussian. The maps t_j can be derived by conditional multivariate Gaussian calculations.

Example 3 *(continuous and binary data).* Let x_1 be a Bernoulli random variable and x_2 be a continuous random variable. The conditional models are

$$x_1 \mid x_2 \sim Bernoulli\left(\frac{e^{\alpha+\beta x_2}}{1+e^{\alpha+\beta x_2}} \right), \quad x_2 \mid x_1 \sim N(\beta_0 + \beta_1 x_1, \sigma^2).$$

The above conditional models are compatible with the joint model

$$x_1 \sim Bernoulli(p), \quad x_2 \mid x_1 \sim N(\beta_0 + \beta_1 x_1, \sigma^2).$$

If we let

$$t_1(p, \beta_0, \beta_1, \sigma^2) = \left(\log\frac{p}{1-p} - \frac{\beta_1^2}{2\sigma^2}, \frac{\beta_1}{2\sigma^2} \right) = (\alpha, \beta)$$

and $t_2(p, \beta_0, \beta_1, \sigma^2) = (\beta_0, \beta_1)$, the conditional models and this joint model are compatible with each other (Efron, 1975; McCullagh & Nelder, 1998).

Example 4 *(incompatible Gaussian conditionals).* The two conditional models,

$$x \mid y \sim N(\beta_1 y + \beta_2 y^2, 1), \quad y \mid x \sim N(\lambda_1 x, 1),$$

are compatible only if $\beta_2 = 0$. Nonetheless, this model is semi-compatible, the definition of which will be given in Section 4.

Example 5 *(ordinal and continuous variable).* Suppose x_1 is continuous, x_2 takes values in $\{0, 1, ..., m\}$, and $x_1 \mid x_2 \sim N(\alpha_0 + \alpha_1 x_2, \tau_1^2)$. The model of $x_2 \mid x_1$ assumes a latent structure,

with $z \sim N(\beta_0 + \beta_1 x_1, \tau_2^2)$ and x_2 taking on the value 0 if $z \leq \mu_1$, 1 if $\mu_1 < z \leq \mu_2, \ldots, m$ if $z > \mu_m$. This is another family of reasonable but incompatible conditional models.

3·2. *Total variation distance between two transition kernels*

Let $\{X^{\text{mis},1}(k) : k \in \mathbb{Z}^+\}$ be the Gibbs chain and $\{X^{\text{mis},2}(k) : k \in \mathbb{Z}^+\}$ be the iterative chain as described in Sections 2·1 and 2·2. Both chains live on the space of the missing data. We write the completed data as $X^i(k) = (X^{\text{mis},i}(k), X^{\text{obs}})$ for the Gibbs chain ($i = 1$) and the iterative chain ($i = 2$) at the kth iteration of the iterative chain. The transition kernels are

$$K_i(w, \mathrm{d}w') = \text{pr}\{X^{\text{mis},i}(k+1) \in \mathrm{d}w' \mid X^{\text{mis},i}(k) = w\},\ i = 1, 2, \tag{2}$$

where w is a generic notation for the state of the processes. The transition kernels, K_1 and K_2, depend on X^{obs}. For simplicity, we omit the index X^{obs} in the notation K_i. Also, we define

$$K_i^{(k)}(\nu, A) = \text{pr}_\nu\{X^{\text{mis},i}(k) \in A\},$$

for $X^{\text{mis},i}(0) \sim \nu$ and ν being some starting distribution. The probability measure pr_ν also depends on X^{obs}. Let d_{TV} denote the total variation distance between two measures; that is, for two measures, ν_1 and ν_2, defined on the same probability space $(\Omega, \mathcal{F})$ we define $d_{\text{TV}}(\nu_1, \nu_2) = \sup_{A \in \mathcal{F}} |\nu_1(A) - \nu_2(A)|$. We further define $\|\nu\|_V = \sup_{|h| \leq V} \int h(x)\nu(\mathrm{d}x)$ and $\|\nu\|_1 = \|\nu\|_V$ for $V \equiv 1$.

Throughout this paper, we assume that both chains are positive Harris recurrent and thus K_i admits its unique stationary distribution denoted by $\nu_i^{X^{\text{obs}}}$. We intend to establish conditions under which $d_{\text{TV}}(\nu_1^{X^{\text{obs}}}, \nu_2^{X^{\text{obs}}}) \to 0$ in probability as $n \to \infty$ and thus the iterative imputation and the joint Bayesian imputation are asymptotically the same.

Our basic strategy for analyzing the compatible conditional models is to first establish that the transition kernels K_1 and K_2 are close to each other in a large region A_n depending on the observed data X^{obs}, that is, $\|K_1(w, \cdot) - K_2(w, \cdot)\|_1 \to 0$ as $n \to \infty$ for $w \in A_n$; and, second, to show that the two stationary distributions are close to each other in total variation in that the stationary distributions are completely determined by the transition kernels. In this subsection, we start with the first step, that is, to show that K_1 converges to K_2.

Both the Gibbs chain and the iterative chain evolve by updating each missing variable from the corresponding posterior predictive distributions. Upon comparing the difference between the two transition kernels associated with the simulation schemes in Sections 2·1 and 2·2, it suffices to compare the following posterior predictive distributions for each $1 \leq j \leq p$,

$$f(X_j^{\text{mis}} \mid X_j^{\text{obs}}, X_{-j}) = \int f(X_j^{\text{mis}} \mid X_j^{\text{obs}}, X_{-j}, \theta) p(\theta \mid X_j^{\text{obs}}, X_{-j}) \mathrm{d}\theta, \tag{3}$$

$$g_j(X_j^{\text{mis}} \mid X_j^{\text{obs}}, X_{-j}) = \int g_j(X_j^{\text{mis}} \mid X_j^{\text{obs}}, X_{-j}, \theta_j) p_j(\theta_j \mid X_j^{\text{obs}}, X_{-j}) \mathrm{d}\theta_j, \tag{4}$$

where p and p_j denote the posterior distributions under f and g_j respectively. Due to compatibility, the distributions of the missing data given the parameters are the same for the joint Bayesian model and the iterative imputation model $f(X_j^{\text{mis}} \mid X_j^{\text{obs}}, X_{-j}, \theta) = g_j(X_j^{\text{mis}} \mid X_j^{\text{obs}}, X_{-j}, \theta_j)$ if $t_j(\theta) = \theta_j$. The only difference lies in their posterior distributions. Therefore, we move to comparing $p(\theta \mid X_j^{\text{obs}}, X_{-j})$ and $p_j(\theta_j \mid X_j^{\text{obs}}, X_{-j})$.

Upon comparing the posterior distributions of θ and θ_j, the first disparity to reconcile is that the dimensions are usually different. Typically θ_j is of lower dimension. Consider the linear model in Example 1. The conditional models include three parameters, two regression coefficients and the error variance, while the joint model has five parameters μ_x, μ_y, σ_x, σ_y, and ρ. This is

because the regression models are usually conditional on the covariates. The joint model not only parameterizes the conditional distributions of x_j given x_{-j} but also the marginal distribution of x_{-j}. Therefore, it includes extra parameters, although the distributions of the missing data is independent of these parameters. We augment the parameter space of the iterative imputation to (θ_j, θ_j^*) with the corresponding map $\theta_j^* = t_j^*(\theta)$. The augmented parameter (θ_j, θ_j^*) is a non-degenerate reparameterization of θ, that is, $T_j(\theta) = \{t_j(\theta), t_j^*(\theta)\}$ is a one-to-one invertible map.

To illustrate this parameter augmentation, we consider the linear model in Example 1 in which $\theta = (\mu_x, \sigma_x^2, \mu_y, \sigma_y^2, \rho)$, where we use μ_x and σ_x^2 to denote the mean and variance of the first variable, μ_y and σ_y^2 to denote the mean and variance of the second, and ρ to denote the correlation. The reparameterization is,

$$\theta_2 = t_2(\mu_x, \sigma_x^2, \mu_y, \sigma_y^2, \rho) = (\alpha_{y|x}, \beta_{y|x}, \tau_y^2) = \Big\{\mu_y - \frac{\rho\sigma_y}{\sigma_x}\mu_x, \frac{\rho\sigma_y}{\sigma_x}, (1-\rho^2)\sigma_y^2\Big\},$$
$$\theta_2^* = t_2^*(\mu_x, \sigma_x^2, \mu_y, \sigma_y^2, \rho) = (\mu_x, \sigma_x^2).$$

The function t_2 maps to the regression coefficients and the variance of the residuals; t_2^* maps to the marginal mean and variance of x. Similarly, we can define the map of t_1 and t_1^*.

Because we are assuming compatibility, we can drop the notation g_j for the conditional model of the jth variable. Instead, we unify the notation to that of the joint Bayesian model $f(x_j \mid x_{-j}, \theta)$. In a slight abuse of notation, we write $f(x_j \mid x_{-j}, \theta_j) = f(x_j \mid x_{-j}, \theta)$ for $t_j(\theta) = \theta_j$. For instance, in Example 1, we write $f(y \mid x, \alpha_{y|x}, \beta_{y|x}, \sigma_{y|x}) = f(y \mid x, \mu_x, \mu_y, \sigma_x, \sigma_y, \rho)$ as long as $\alpha_{y|x} = \mu_y - \rho\sigma_y\mu_x/\sigma_x$, $\beta_{y|x} = \rho\sigma_y/\sigma_x$, and $\sigma_{y|x}^2 = (1-\rho^2)\sigma_y^2$.

The prior distribution on θ for the joint Bayesian model implies a prior on (θ_j, θ_j^*) of

$$\pi_j^*(\theta_j, \theta_j^*) = \det(\partial T_j/\partial\theta)^{-1}\pi\{T_j^{-1}(\theta_j, \theta_j^*)\}.$$

For the full Bayesian model, the posterior distribution of θ_j is

$$p(\theta_j \mid x_j^{\text{obs}}, x_{-j}) = \int p(\theta_j, \theta_j^* \mid x_j^{\text{obs}}, x_{-j})\mathrm{d}\theta_j^* \propto \int f(x_j^{\text{obs}}, x_{-j} \mid \theta_j, \theta_j^*)\pi_j^*(\theta_j, \theta_j^*)\mathrm{d}\theta_j^*.$$

Because $f(x_j^{\text{obs}} \mid x_{-j}, \theta_j, \theta_j^*) = f(x_j^{\text{obs}} \mid x_{-j}, \theta_j)$, the above posterior can be further reduced to

$$p(\theta_j \mid x_j^{\text{obs}}, x_{-j}) \propto f(x_j^{\text{obs}} \mid x_{-j}, \theta_j) \int f(x_{-j} \mid \theta_j, \theta_j^*)\pi_j^*(\theta_j, \theta_j^*)\mathrm{d}\theta_j^*.$$

If we write $\pi_{j,X_{-j}}(\theta_j) = \int f(x_{-j} \mid \theta_j, \theta_j^*)\pi_j^*(\theta_j, \theta_j^*)\mathrm{d}\theta_j^*$, then the posterior distribution of θ_j under the joint Bayesian model is

$$p(\theta_j \mid x_j^{\text{obs}}, x_{-j}) \propto f(x_j^{\text{obs}} \mid x_{-j}, \theta_j)\pi_{j,X_{-j}}(\theta_j).$$

Compared with the posterior distribution under iterative imputation,

$$p_j(\theta_j \mid x_j^{\text{obs}}, x_{-j}) \propto g_j(x_j^{\text{obs}} \mid x_{-j}, \theta_j)\pi_j(\theta_j) = f(x_j^{\text{obs}} \mid x_{-j}, \theta_j)\pi_j(\theta_j),$$

the difference lies in the prior distributions, $\pi_j(\theta_j)$ and $\pi_{j,X_{-j}}(\theta_j)$.

We put forward tools to control the distance between the two posterior predictive distributions in (3) and (4). Let x be the generic notation for the observed data, and let $f_X(\theta)$ and $g_X(\theta)$ be two posterior densities of θ. Let $h(\tilde{x} \mid \theta)$ be the density function for future observations given the parameter θ, and let $\tilde{f}_X(\tilde{x}) = \int h(\tilde{x} \mid \theta)f_X(\theta)\mathrm{d}\theta$ and $\tilde{g}_X(\tilde{x}) = \int h(\tilde{x} \mid \theta)g_X(\theta)\mathrm{d}\theta$ be the posterior predictive distributions. Then it is straightforward that

$$\|\tilde{f}_X - \tilde{g}_X\|_1 \le \|f_X - g_X\|_1. \tag{5}$$

The next proposition provides sufficient conditions that $\|f_X - g_X\|_1$ vanishes.

PROPOSITION 1. *Let n be the sample size. Let $f_X(\theta)$ and $g_X(\theta)$ be two posterior density functions that share the same likelihood but have two different prior distributions π_f and π_g. Let*

$$L(\theta) = \frac{\pi_g(\theta)}{\pi_f(\theta)}, \qquad r(\theta) = \frac{g_X(\theta)}{f_X(\theta)} = \frac{L(\theta)}{\int L(\theta') f_X(\theta') d\theta'}.$$

Let $\partial L(\theta)$ be the partial derivative with respect to θ and let ξ be a random variable such that

$$L(\theta) = L(\mu_\theta) + \partial L(\xi)^T(\theta - \mu_\theta),$$

where $\mu_\theta = \int \theta f_X(\theta) d\theta$. If there exists a random variable $Z(\theta)$ with finite variance under f_X, such that

$$|n^{1/2}\partial L(\xi)^T(\theta - \mu_\theta)| \leq |\partial L(\mu_\theta)| Z(\theta), \tag{6}$$

then there exists a constant $\kappa > 0$ such that, for n sufficiently large,

$$\|\tilde{f}_X - \tilde{g}_X\|_1 \leq \frac{\kappa |\partial \log L(\mu_\theta)|^{1/2}}{n^{1/4}}. \tag{7}$$

We prove this proposition in the Supplementary Material.

Remark 1. We adapt Proposition 1 to the analysis of the conditional models. Expression (6) implies that the posterior standard deviation of θ is $O(n^{-1/2})$. For most parametric models, (6) is satisfied as long as the observed Fisher information is bounded from below by εn for some $\varepsilon > 0$. In particular, we let $\hat{\theta}(x)$ be the complete-data maximum likelihood estimator and $A_n = \{x : |\hat{\theta}(x)| \leq \gamma\}$. Then, (6) is satisfied on the set A_n for any fixed γ.

Remark 2. In order to verify that $\partial \log L(\theta)$ is bounded, one only needs to know π_f and π_g up to a normalizing constant. This is because the bound is in terms of $\partial L(\theta)/L(\theta)$. This helps to handle the situation when improper priors are used and it is not feasible to obtain a normalized prior distribution. In the current context, the prior likelihood ratio is $L(\theta_j) = \pi_j(\theta_j)/\pi_{j,X_{-j}}(\theta_j)$.

We further provide a special case where the above proposition applies. Suppose that the parameter spaces of the conditional distribution and the covariates are separable, that is, $f(x_j, x_{-j} \mid \theta_j, \theta_j^*) = f(x_j \mid x_{-j}, \theta_j) f(x_{-j} \mid \theta_j^*)$ and there exists a prior π for the joint model f such that θ_j and θ_j^* are a priori independent for all j. Then, the boundedness of $\partial \log L(\theta_j)$ becomes straightforward to obtain. The ratio $L(\theta_j) = \pi_j(\theta_j)/\pi_{j,X_{-j}}(\theta_j)$, and $\pi_{j,X_{-j}}(\theta_j) = \pi_j^*(\theta_j) \int f(x_{-j} \mid \theta_j^*) \pi_j^*(\theta_j^*) \mathrm{d}\theta_j^*$. Thus, $L(\theta_j) = \pi_j(\theta_j)/\pi_j^*(\theta_j)$ is independent of data. Further, if one chooses $\pi_j(\theta_j) = \pi_j^*(\theta_j)$, then the transition probabilities of the iterative and Gibbs chains coincide and $\nu_1^{X^{\text{obs}}} = \nu_2^{X^{\text{obs}}}$.

3·3. *Convergence of the invariant distributions*

With Proposition 1 and Remark 1, we have established that the transition kernels of the Gibbs chain and the iterative chain are close to each other in a large region A_n. The subsequent analysis falls into several steps. First, we slightly modify the processes by conditioning on the set A_n with stationary distributions $\tilde{\nu}_i^{X^{\text{obs}}}$, the details of which is provided below. The stationary distributions of the conditional processes and the original processes, $\tilde{\nu}_i^{X^{\text{obs}}}$ and $\nu_i^{X^{\text{obs}}}$, are close in total variation. Second, we show in Lemma 2 that, with a bound on the convergence rate of the Markov process, $\tilde{\nu}_1^{X^{\text{obs}}}$ and $\tilde{\nu}_2^{X^{\text{obs}}}$ are close in total variation and so it is with $\nu_1^{X^{\text{obs}}}$ and $\nu_2^{X^{\text{obs}}}$.

An upper bound for convergence rate of the Markov chains, in particular r_k in Lemma 2, can be established by Proposition 2.

To proceed, we consider the chains conditional on the set A_n where the two transition kernels are uniformly close to each other. In particular, for each set B, we let

$$\tilde{K}_i(w, B) = \frac{K_i(w, B \cap A_n)}{K_i(w, A_n)}. \tag{8}$$

That is, we create another two processes, for which we update the missing data conditional on $x \in A_n$. Next we show that we only need to consider the chains conditional on the set A_n.

LEMMA 1. *Suppose that both K_1 and K_2 are positive Harris recurrent. We can choose A_n as in the form of Remark* 1 *and γ sufficiently large so that*

$$\nu_i^{X^{obs}}(A_n) \to 1 \qquad \textit{in probability as } n \to \infty. \tag{9}$$

Let $\tilde{X}^{mis,i}(k)$ be the Markov chains following $\tilde{K}_i$, defined as in (8)*, with invariant distribution $\tilde{\nu}_i^{X^{obs}}$. Then,*

$$\lim_{n\to\infty} d_{TV}(\nu_i^{X^{obs}}, \tilde{\nu}_i^{X^{obs}}) = 0. \tag{10}$$

The proof is elementary by the representation of $\nu_i^{X^{\text{obs}}}$ through renewal theory and therefore is omitted. Based on the above lemma, we need to show that $d_{\text{TV}}(\tilde{\nu}_1^{X^{\text{obs}}}, \tilde{\nu}_2^{X^{\text{obs}}}) \to 0$. The expression $\|K_1(w, \cdot) - K_2(w, \cdot)\|_1$ approaches 0 uniformly for $w \in A_n$. This implies that

$$\lim_{n\to\infty} \|\tilde{K}_1(w, \cdot), \tilde{K}_2(w, \cdot)\|_1 = 0 \text{ uniformly for } w \in A_n. \tag{11}$$

With the above convergence, we can establish the convergence between $\tilde{\nu}_1^{X^{\text{obs}}}$ and $\tilde{\nu}_2^{X^{\text{obs}}}$.

LEMMA 2. *Let $\tilde{X}^{mis,i}(k)$ admit data-dependent transition kernels $\tilde{K}_i$ for $i = 1, 2$. We use n to denote sample size. Suppose that each $\tilde{K}_i$ admits a data-dependent unique invariant distribution, denoted by $\tilde{\nu}_i^{X^{obs}}$, and that the following two conditions hold.*

First, the convergence of the two transition kernels are in place, that is,

$$d(A_n) = \sup_{w \in A_n} \|\tilde{K}_1(w, \cdot) - \tilde{K}_2(w, \cdot)\|_V \to 0 \qquad \textit{in probability as } n \to \infty. \tag{12}$$

The function V is either a geometric drift function for $\tilde{K}_2$ or a constant, i.e., $V \equiv 1$.

Second, there exists a monotone decreasing data-independent sequence $r_k \to 0$ and a data-dependent starting measure ν such that

$$pr\left\{\|\tilde{K}_i^{(k)}(\nu, \cdot) - \tilde{\nu}_i^{X^{obs}}(\cdot)\|_V \le r_k, \textit{ for all } k > 0\right\} \to 1, \quad n \to \infty. \tag{13}$$

Then,

$$\|\tilde{\nu}_1^{X^{obs}} - \tilde{\nu}_2^{X^{obs}}\|_V \to 0, \textit{ in probability as } n \to \infty. \tag{14}$$

Remark 3. The proof is included in the Supplementary Material. Lemma 2 holds if $V = 1$ or V is a drift function, the definition of which is given in (16). For the analysis of convergence in total variation, we only need that $V = 1$. We prepare the result when V is a drift function for the analysis of incompatible models.

The first condition in the above lemma has been obtained by the result of Proposition 1 and (11). Condition (13) is more difficult to establish. According to the standard results in Rosenthal

(1995), one set of sufficient conditions for (13) is that the chains $\tilde{K}_1$ and $\tilde{K}_2$ admit a common small set, C; in addition, each of them admits its own drift functions associated with the small set C. See the Supplementary Material for more details.

Gibbs chains typically admit a small set C and a drift function V; that is,

$$\tilde{K}_1(w, A) \geq q_1 \mu_1(A), \text{ for some positive measure } \mu_1, \tag{15}$$

with $w \in C$, $0 < q_1 < 1$; for some $0 < \lambda_1 < 1$ and for all $w \notin C$,

$$\lambda_1 V(w) \geq \int V(w')\tilde{K}_1(w, \mathrm{d}w'). \tag{16}$$

With the existence of C and V, an upper bound r_k with starting point $w \in C$ can be established for the Gibbs chain by standard results, and r_k only depends on λ_1 and q_1. Therefore, it is necessary to require that λ_1 and q_1 are independent of x^{obs}. In contrast, the small set C and drift function V could be data dependent.

Given that $\tilde{K}_1$ and $\tilde{K}_2$ are close in total variation distance, the set C is also a small set for $\tilde{K}_2$, that is $\tilde{K}_2(w, A) \geq q_2\mu_2(A)$, for some $0 < q_2 < 1$, all $w \in C$, and all measurable set A. The following proposition, whose proof is given in the Supplementary Material, establishes the conditions under which V is also a drift function for $\tilde{K}_2$ so that (13) is in place.

PROPOSITION 2. *Assume the following conditions hold. First, the transition kernel $\tilde{K}_1$ admits a small set C and a drift function V satisfying* (16)*. Second, let $L_j(\theta_j) = \pi_j(\theta_j)/\pi_{j,X_{-j}}(\theta_j)$ $(j = 1, ..., p)$ be the the ratio of prior distributions for each conditional model, possibly depending on the data, so that on the set A_n $\sup_{|\theta_j|<\gamma} \partial L_j(\theta_j)/L_j(\theta_j) < \infty$. Third, for each j and $1 \leq k \leq p - j$, there exists a $Z_j(\theta_j)$ serving as the bound in* (6) *for each L_j. In addition, Z_j satisfies the following moment condition*

$$\tilde{E}_1 \left\{ Z^2_{j+1}(\theta_{j+1}) V^2(w_{j+k}) \mid w_j \right\} = o(n)V^2(w_j), \tag{17}$$

where $\tilde{E}_1$ is the expectation associated with the updating distribution of $\tilde{K}_1$, and w_j is the state of the chain when the jth variable is just updated. The convergence $o(n)/n \to 0$ is uniform in $w_j \in A_n$.

Then, there exists $0 < \lambda_2 < 1$ such that as n tends to infinity with probability converging to one the following inequality holds

$$\lambda_2 V(w) \geq \int V(w')\tilde{K}_2(w, dw'). \tag{18}$$

The proof is included in the Supplementary Material. The intuition is as follows. The function V satisfying inequality (16) is a drift function of $\tilde{K}_1$ to C. Since $\tilde{K}_1$ and $\tilde{K}_2$ are close to each other, we may expect that $\int V(w')\tilde{K}_1(w, \mathrm{d}w') \approx \int V(w')\tilde{K}_2(w, \mathrm{d}w')$. The above proposition states the conditions under which this approximation is indeed true and suggests that V be a drift function of $\tilde{K}_2$ if it is a drift function of $\tilde{K}_1$. Condition (17) is imposed for a technical purpose. In particular, we allow the expectation of $Z^2_{j+1}(\theta_{j+1})V^2(w_{j+k})$ to grow to infinity but at a slower rate than n. Therefore, it is a weak condition. We now summarize the analysis and the results of the compatible conditional models in the following theorem.

THEOREM 1. *Suppose that a set of conditional models $\{g_j(x_j \mid x_{-j}, \theta_j) : \theta_j \in \Theta_j, j = 1, ..., p\}$ is compatible with a joint model $\{f(x \mid \theta) : \theta \in \Theta\}$. Both the Gibbs chain and the iterative chain are positive Harris recurrent and thus admit their unique stationary distributions $\nu_i^{X^{obs}}$. Furthermore, the conditions in Proposition 2 hold so that $\tilde{K}_1$ and $\tilde{K}_2$ are geometrically recurrent.*

Lastly, the following conditions are satisfied. Let $A_n = \{x : |\hat{\theta}(x)| \leq \gamma\}$. *One can choose* γ *sufficiently large so that*

$$\nu_i^{X^{obs}}(A_n) \to 0, \tag{19}$$

in probability as $n \to \infty$. *Then,* $d_{TV}(\nu_1^{X^{obs}}, \nu_2^{X^{obs}}) \to 0$ *in probability as* $n \to \infty$.

One sufficient condition for (19) is that the stationary distributions of $\hat{\theta}(x)$ under $\nu_i^{X^{\text{obs}}}$ converge to a value θ^i, where θ^1 and θ^2 are not necessarily the same. Therefore, (19) is a very weak condition. In addition to the conditions of Proposition 1, Proposition 2 also requires that one constructs a drift function towards a small set for the Gibbs chain. One can usually construct q_1 and λ_1 free of data if the proportion of missing data is bounded from the above by $1 - \varepsilon$. The most difficult task usually lies in constructing a drift function.

Proof. We summarize the analysis of compatible models in this proof. If g_j's are compatible with f, then the conditional posterior predictive distributions of the Gibbs chain and the iterative chain are given in (3) and (4). Thanks to compatibility, the total variation distance between the posterior predictive distributions are bounded by the distance between the posterior distributions of their own parameters as in (5).

On the set A_n, the Fisher information of the likelihood has a lower bound of εn for some ε. Then, by Proposition 1 and the second condition in Proposition 2, the distance between the two posterior distributions is of order $O(n^{-1/4})$ uniformly on set A_n. Similar convergence result holds for the conditional transition kernels, that is, $\|\tilde{K}_1(w, \cdot) - \tilde{K}_2(w, \cdot)\|_1 \to 0$. Thus, the first condition in Lemma 2 has been satisfied.

To verify the conditions of Proposition 2, one needs to construct a small set C such that (15) holds for both chains, and a drift function V for one of the two chains such that (16) holds. Based on the results of Proposition 2, $\tilde{K}_1$ and $\tilde{K}_2$ share a common data-dependent small set C with q_i independent of data and a drift function V possibly with different rate λ_1 and λ_2.

According to the standard bound of Markov chain rate of convergence stated in the Supplementary Material, there exists a common starting value $w \in C$ and a bound r_k such that (13) holds. Then both conditions in Lemma 2 are satisfied, and $d_{\text{TV}}(\tilde{\nu}_1^{X^{\text{obs}}}, \tilde{\nu}_2^{X^{\text{obs}}}) \to 0$ in probability as $n \to \infty$. According to condition (19) and Lemma 1, this implies $d_{\text{TV}}(\nu_1^{X^{\text{obs}}}, \nu_2^{X^{\text{obs}}}) \to 0$, and so we are done. ■

3·4. *On the necessity of model compatibility*

Theorem 1 shows that for compatible models and under suitable technical conditions, iterative imputation is asymptotically equivalent to Bayesian imputation. The following theorem suggests that model compatibility is typically necessary for this convergence.

Let pr^f denote the probability measure induced by the posterior predictive distribution of the joint Bayesian model and pr_j^g denote those induced by the iterative imputation's conditional models. That is,

$$\text{pr}^f(x_j^{\text{mis}} \in A \mid x_{-j}^{\text{mis}}, x^{\text{obs}}) = \int_A f(x_j^{\text{mis}} \mid x_{-j}^{\text{mis}}, x^{\text{obs}}, \theta) p(\theta \mid x_{-j}^{\text{mis}}, x^{\text{obs}}) \mathrm{d}\theta,$$
$$\text{pr}_j^g(x_j^{\text{mis}} \in A \mid x_{-j}^{\text{mis}}, x^{\text{obs}}) = \int_A g_j(x_j^{\text{mis}} \mid x_{-j}^{\text{mis}}, x^{\text{obs}}, \theta_j) p_j(\theta_j \mid x_{-j}^{\text{mis}}, x^{\text{obs}}) \mathrm{d}\theta.$$

Denote the stationary distributions of the Gibbs and iterative chains by $\nu_1^{X^{\text{obs}}}$ and $\nu_2^{X^{\text{obs}}}$.

THEOREM 2. *Suppose that for some $j \in \mathbb{Z}^+$, sets A and C, and $0 < \varepsilon < 1/2$,*

$$\inf_{X^{mis}_{-j} \in C} pr^g_j(x^{mis}_j \in A \mid x^{mis}_{-j}, x^{obs}) > \sup_{X^{mis}_{-j} \in C} pr^f(x^{mis}_j \in A \mid x^{mis}_{-j}, x^{obs}) + \varepsilon$$

or

$$\sup_{X^{mis}_{-j} \in C} pr^g_j(x^{mis}_j \in A \mid x^{mis}_{-j}, x^{obs}) < \inf_{X^{mis}_{-j} \in C} pr^f(x^{mis}_j \in A \mid x^{mis}_{-j}, x^{obs}) - \varepsilon$$

and $\nu_1^{X^{obs}}(x^{mis}_{-j} \in C) > q \in (0,1)$. Then there exists a set B such that

$$\left| \nu_2^{X^{obs}}(x^{mis} \in B) - \nu_1^{X^{obs}}(x^{mis} \in B) \right| > q\varepsilon/4.$$

For two models with different likelihood functions, one can construct sets A and C such that the conditions in the above theorem hold. Therefore, if among the predictive distributions of all the p conditional models there is one g_j that is different from f as stated in Theorem 2, then the stationary distribution of the iterative imputation is different from the posterior distribution of the Bayesian model in total variation by a fixed amount. For a set of incompatible models and any joint model f, there exists at least one j such that the conditional likelihood functions of x_j given x_{-j} are different for f and g_j. Their predictive distributions have to be different for x_j. Therefore, such an iterative imputation using incompatible conditional models typically does not correspond to Bayesian imputation under any joint model.

4. INCOMPATIBLE CONDITIONAL MODELS

4·1. *Semi-compatibility and model validity*

As in the previous section, we assume that the invariant distribution exists. For compatible conditional models, we used the posterior distribution of the corresponding Bayesian model as the natural benchmark and showed that the two imputation distributions converge to each other. We can use this idea for the analysis of incompatible models. In this setting, the first task is to find a natural Bayesian model associated with a set of incompatible conditional models. We introduce the concept of semi-compatibility.

DEFINITION 2. *A set of conditional models $\{h_j(x_j \mid x_{-j}, \theta_j, \varphi_j) : j = 1, ..., p\}$, each of which is indexed by two sets of parameters (θ_j, φ_j), is said to be semi-compatible, if there exists a set of compatible conditional models*

$$g_j(x_j \mid x_{-j}, \theta_j) = h_j(x_j \mid x_{-j}, \theta_j, \varphi_j = 0) \quad (j = 1, ..., p). \tag{20}$$

We call $\{g_j : j = 1, ..., p\}$ a compatible element of $\{h_j : j = 1, ..., p\}$.

By definition, every set of compatible conditional models is semi-compatible. A simple and uninteresting class of semi-compatible models arises with iterative regression imputation. As typically parameterized, these models include complete independence as a special case. A trivial compatible element, then, is the one in which x_j is independent of x_{-j} under g_j for all j. Throughout the discussion of this section, we use $\{g_j : j = 1, ..., p\}$ to denote the compatible element of $\{h_j : j = 1, ..., p\}$ and f to denote the joint model compatible with $\{g_j : j = 1, ..., p\}$.

Semi-compatibility is a natural concept connecting a joint probability model to a set of conditionals. One foundation of almost all statistical theories is that data are generated according to some probability law. When setting up each conditional model, the imputer chooses a rich family

that is intended to include distributions that are close to the true conditional distribution. For instance, as recommended by Meng (1994), the imputer should try to include as many predictors as possible using regularization as necessary to keep the estimates stable. Sometimes, the degrees of flexibility among the conditional models are different. For instance, some includes quadratic terms or interactions. This richness usually results in incompatibility. Semi-compatibility includes such cases in which the conditional models are rich enough to include the true model but may not be always compatible among themselves. Example 4 in Section 3 is an incompatible but semi-compatible case. To proceed, we introduce the following definition.

DEFINITION 3. *Let $\{h_j : j = 1, ..., p\}$ be semi-compatible, $\{g_j : j = 1, ..., p\}$ be its compatible element, and f be the joint model compatible with g_j. If the joint model $f(x \mid \theta)$ includes the true probability distribution, we say $\{h_j : j = 1, ..., p\}$ is a set of valid semi-compatible models.*

In order to obtain good prediction, we need the validity of the semi-compatible models. A natural issue is the performance of valid semi-compatible models. Given that we have given up compatibility, we should not expect iterative imputation to be equivalent to any joint Bayesian imputation. Nevertheless, under mild conditions, we are able to show the consistency of the combined imputation estimator.

4·2. *Main theorem of incompatible conditional models*

Now, we list a set of conditions.

Condition B1. The Gibbs and iterative chains admit unique invariant distributions, $\nu_1^{X^{\text{obs}}}$ and $\nu_2^{X^{\text{obs}}}$.

Condition B2. The posterior distributions of θ based on f and (θ_j, φ_j) based on h_j given a complete data set x have the representation $|\theta - \tilde{\theta}| \leq \xi n^{-1/2}, |(\theta_j - \tilde{\theta}_j, \varphi_j - \tilde{\varphi}_j)| \leq \xi_j n^{-1/2}$, where $\tilde{\theta}$ is the maximum likelihood estimate of $f(x \mid \theta)$, $(\tilde{\theta}_j, \tilde{\varphi}_j)$ is the maximum likelihood estimate of $h_j(x_j \mid x_j, \theta_j, \varphi_j)$, and $Ee^{|\xi_j|} \leq \kappa$, $Ee^{|\xi|} \leq \kappa$ for some $\kappa > 0$.

Condition B3. All the score functions have finite moment generating functions under $f(x^{\text{mis}} \mid x^{\text{obs}}, \theta)$.

Condition B4. For each variable j, let ι_j be the subset of observations so that, for each $i \in \iota_j$, $x_{i,j}$ is missing and $x_{i,-j}$ is fully observed. Assume that the cardinality $\#(\iota_j) \to \infty$ as $n \to \infty$.

Remark 4. The stationary distribution of the iterative imputation $\nu_2^{X^{\text{obs}}}$ depends in general on the order of updating. Even at convergence, the joint distribution of the imputed values changes after each variable is updated. Here we define $\nu_2^{X^{\text{obs}}}$ as the imputation distribution when variable p has been updated when stationarity has been reached; thus it is well defined if the chain does not blow up or drift to infinity.

Remark 5. Conditions B2 and B3 impose moment conditions on the posterior distribution and the score functions. They are satisfied by most parametric families. Condition B4 rules out certain boundary cases of missingness patterns and is imposed for technical purposes. The condition is not very restrictive because it only requires that the cardinality of ι_j tends to infinity, not necessarily even of order $O(n)$.

We now express the fifth and final condition which requires the following construction. Assume the conditional models are valid and that x is generated from $f(x \mid \theta^0)$. We use $\theta_j^0 = t_j(\theta^0)$ and $\varphi_j^0 = 0$ to denote the true parameters under h_j. We define the observed-data maximum likelihood estimator,

$$\hat{\theta} = \sup_{\theta} f(x^{\text{obs}} \mid \theta), \qquad \hat{\theta}_j = t_j(\hat{\theta}), \tag{21}$$

and the combined estimator based on infinitely many imputations,

$$\hat{\theta}^{(2)} = \arg\sup_{\theta} \int \log f(X \mid \theta) \nu_2^{X^{\text{obs}}}(\mathrm{d}X^{\text{mis}}), \tag{22}$$
$$(\hat{\theta}_j^{(2)}, \hat{\varphi}_j^{(2)}) = \arg\sup_{\theta_j, \varphi_j} \int \log h_j(X_j \mid X_{-j}, \theta_j, \varphi_j) \nu_2^{X^{\text{obs}}}(\mathrm{d}X^{\text{mis}})$$

where $X = (X^{\text{obs}}, X^{\text{mis}})$, and $\hat{\theta}$, $\hat{\theta}^{(2)}$, $(\hat{\theta}_j^{(2)}, \hat{\varphi}_j^{(2)})$ only depend on X^{obs}.

Consider a Markov chain $x^*(k)$ corresponding to one observation, that is, one row of the data matrix, living on R^p. The chain evolves as follows. Within each iteration, each dimension j is updated conditional on the others according to the conditional distribution $h_j(x_j \mid x_{-j}, \theta_j, \varphi_j)$, where $(\theta_j, \varphi_j) = (\hat{\theta}_j, 0) + \varepsilon\xi_j$ and ξ_j is a random vector with finite moment generating function independent of everything at every step. Alternatively, one may consider (θ_j, φ_j) as a sample from the posterior distribution corresponding to the conditional model h_j. Thus, $x^*(k)$ is the marginal chain of one observation in the iterative chain. Given that $X^{\text{mis},2}(k)$ admits a unique invariant distribution, $x^*(k)$ admits its unique stationary distribution for ε sufficiently small.

Furthermore, consider another process $x(k)$ that is a Gibbs sampler and admits stationary distribution $f(x \mid \hat{\theta})$. That is, each component is updated according to the conditional distribution $f(x_j \mid x_{-j}, \hat{\theta})$ and the parameters of the updating distribution are set at the observed data maximum likelihood estimate, $\hat{\theta}$. If $\varepsilon = 0$, then $x(k)$ is equal in distribution to $x^*(k)$. The last condition is stated as follows.

Condition B5. The chains $x^*(k)$ and $x(k)$ satisfy conditions in Lemma 2 as $\varepsilon \to 0$, that is, the invariant distributions of $x^*(k)$ and $x(k)$ converges in $\|\cdot\|_V$ norm, where V is a drift function for $x^*(k)$. There exists a constant κ such that all the score functions are bounded by

$$|\partial \log f(x \mid \theta^0)| \le \kappa V(x), \qquad |\partial \log h_j(x_j \mid x_{-j}, \theta_j^0, \varphi_j = 0)| \le \kappa V(x).$$

Remark 6. By choosing ε small, the transition kernels of $x^*(k)$ and $x(k)$ converge to each other. Condition B5 requires that Lemma 2 applies in this setting, that their invariant distributions are close in the sense stated in the lemma. This condition does not suggest that Lemma 2 applies to $\nu_1^{X^{\text{obs}}}$ and $\nu_2^{X^{\text{obs}}}$, which represents the joint distribution of many such $x^*(k)$'s and $x(k)$'s.

We can now state the main theorem in this section.

THEOREM 3. *Consider a set of valid semi-compatible models $\{h_j : j = 1, ..., p\}$, and assume conditions B1–5 are in force. Then, following the notations in* (22)*, the following holds for all j:*

$$\hat{\theta}^{(2)} \to \theta^0, \quad \hat{\theta}_j^{(2)} \to \theta_j^0, \quad \hat{\varphi}_j^{(2)} \to 0, \quad \textit{in probability as sample size } n \to \infty. \tag{23}$$

Remark 7. The proof is included in the Supplementary Material. The expression $\hat{\theta}^{(2)}$ corresponds to the following estimator. Impute the missing data from distribution $\nu_2^{X^{\text{obs}}}$ m times to obtain m complete datasets. Stack the m datasets to one big dataset. Let $\hat{\theta}_m^{(2)}$ be the maximum likelihood estimator based on the big dataset. Then, $\hat{\theta}_m^{(2)}$ converges to $\hat{\theta}^{(2)}$ as $m \to \infty$. Furthermore, $\hat{\theta}^{(2)}$ is asymptotically equivalent to the combined point estimator of θ according to Rubin's combining rule with infinitely many imputations. Similarly, $(\hat{\theta}_j^{(2)}, \hat{\varphi}_j^{(2)})$ is asymptotically equivalent to the combined estimator. Therefore, Theorem 3 suggests that the combined imputation estimators are consistent under conditions B1–5.

Remark 8. The consistency results of the above theorem implicitly requires that the analysts' conditional models are consistent with the imputation models. Generally speaking if the analyst

is using a different model, then the combined imputation estimators are usually inconsistent even for joint Bayesian imputations. On the other hand, recent developments show that consistency can be maintained if the imputation model is more saturated than then that of the analyst, for instance if the missing data are imputed based on p variables and the analyst is only interested in running a model containing a subset of the p variables such as pairwise correlations.

5. SIMULATION STUDIES AND REAL-DATA ILLUSTRATION

We perform simulation studies and give a real data illustration, further details of which are in the Supplementary Material. The simulation studies confirm the results presented in the theorems and also provide empirical performance of the imputation combining rules applied to the iterative imputations.

A real data example is provided in the Supplementary Material. We consider a political science application involving imputation for the American National Election Study, a nationwide survey that asks about individuals' views on political issues, candidates, and sources of information, and records other important political and demographic data. We consider a subset of 11 variables representing different aspects of the survey: age of respondent, time to complete the survey, sex, whether the respondent sees the environment as an important issue, education, income, a seven-point scale representing attitude toward government job assistance, religion, marital status, and vote choice. We eliminate partially observed cases, leaving 1442 observations, and then simulate new missing patterns for the remaining complete data, run iterative imputation, and assess the quality of the iterative imputation algorithm by comparing imputed values against estimates for the complete data sample.

For the complete data set, we generate 1000 independent missingness patterns using a fairly complicated missingness-at-random rule. First, a matrix C is created including one variable of each type as the auxiliary variables that remain complete. These variables are age, sex, income, and martial status. The categorical variables are broken into indicators for all but one of their categories, and continuous variables are standardized. Second, for the remaining 7 variables, an $n \times 7$ matrix $M = C\beta + Z$ is created, where β is a matrix of coefficients. For each missing pattern, the elements β are drawn independently from the standard Gaussian random variables. The columns of Z are drawn from the $\mathrm{N}(0, \Sigma)$, where Σ has 1's on its diagonal but is unconstrained off-diagonal by the rdata.frame command in the mi package with option restrictions="none". The elements of M are further transformed to $\pi_{i,j} = U_{ij} - \text{logit}^{-1}(M_{i,j})$ where U_{ij} are independently and identically distributed as uniform distribution over the interval $[0, 1]$. Third, for each of the 7 variables subject to missingness, the observations corresponding to the highest 10% of π_{ij} are set to be missing.

For each missingness pattern, we use the mi package in R to impute the missing data. Linear, logistic, ordered logit, and multinomial logit regressions are used for continuous, binary, ordinal, and unordered categorical variables respectively. For such complicated conditional models, it is generally hard to verify the model compatibility and we generally believe that they are incompatible among each other. We perform empirical diagnosis of the validity for the conditional models, which is the most important condition in Theorem 3. In particular, we compare the model prediction and the observed values to ensure that they provide reasonable predictions.

We assess the quality of the imputations by calculating the differences between the combined estimators of some regression coefficients based on the imputed data and those based on the complete data and examine the differences graphically in Figure 1 for a linear model. The detailed results and more graphical illustrations are presented in Section 1.4 of the Supplementary Material. The estimates using the imputed data appear to be roughly equal in expectation to the

parameters from the true data. In this study, we have an empirical diagnosis of the conditional models, while remaining agnostic about the joint distribution of the data and without making overly restrictive assumptions about the missing data process.

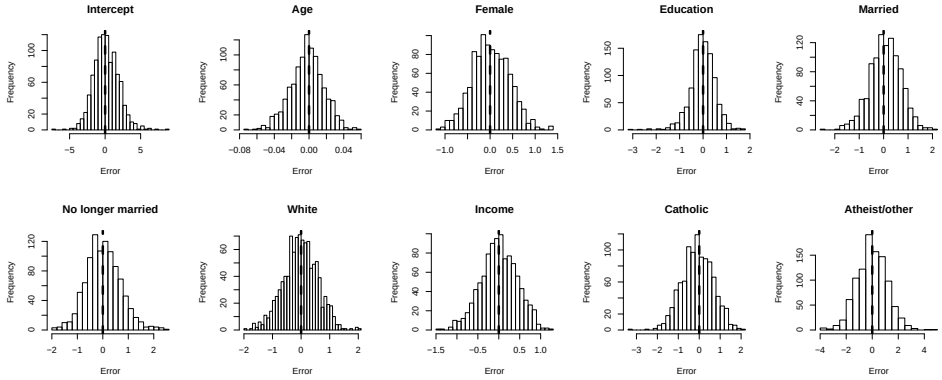

Fig. 1. Histogram of the combined estimates and the true values for 1000 missing patterns. Dependent variable: time to completion with linear regression.

6. DISCUSSION

There are several natural directions for future research. From one direction, it should be possible to obtain exact results for some particular classes of models such as linear regressions with Gaussian errors and Gaussian prior distributions, in which case convergence can be expressed in terms of simple matrix operations. In the more general case of arbitrary families of regression models, it would be desirable to develop diagnostics for stationarity along with proofs of the effectiveness of such diagnostics under some conditions and empirical measures of the magnitude of discrepancies between fitted and stationary conditional distributions.

Another open problem is how to consistently estimate the variance of the combined imputation estimator. For compatible models, iterative imputation is asymptotically equivalent to Bayesian simulation. Under model compatibility, we speculate that Rubin's variance estimator is generally applicable to the iteratively imputed datasets. For incompatible models, the imputation distribution is asymptotically different from any joint Bayesian imputation, hence there is no guarantee that the existing variance estimators are asymptotically consistent. In addition, as the combined imputation estimator cannot be represented by some estimating equations, Robins and Wang's approach does not apply either. Even for joint Bayesian imputation, estimating the variance of the combined estimator is still a nontrivial task under specific situations (Kim, 2004; Meng, 1994). We leave this issue to future studies.

ACKNOWLEDGEMENT

We thank the editors and reviewers for their valuable and instructive comments. This research is supported in part by NSF CMMI-1069064 and SES-1323977, IES grant R305D090006, and Wang Xuelian fundation.

SUPPLEMENTARY MATERIAL

Supplementary Material available at *Biometrika* online includes the technical proofs of all the theorems and numerical results from the simulation study and the real data analysis.

REFERENCES

AMIT, Y. & GRENANDER, U. (1991). Comparing sweep strategies for stochastic relaxation. *Journal of Multivariate Analysis* **37**, 197–222.

BARNARD, J. & RUBIN, D. B. (1999). Small-sample degrees of freedom with multiple imputation. *Biometrika* **86**, 948–955.

BAXENDALE, P. H. (2005). Renewal theory and computable convergence rates for geometrically ergodic Markov chains. *Annals of Applied Probability* **15**, 700–738.

EFRON, B. (1975). Efficiency of logistic regression compared to normal discriminant-analysis. *Journal of the American Statistical Association* **70**, 892–898.

GEMAN, S. & GEMAN, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *Ieee Transactions on Pattern Analysis and Machine Intelligence* **6**, 721–741.

HOBERT, J. & CASELLA, G. (1998). Functional compatibility, Markov chains and Gibbs sampling with improper posteriors. *Journal of Computational and Graphical Statistics* **7**, 42–60.

KIM, J. K. (2004). Finite sample properties of multiple imputation estimators. *Annals of Statistics* **32**, 766–783.

KIM, J. K. (2011). Parametric fractional imputation for missing data analysis. *Biometrika* **98**, 119–132.

LI, K. H., MENG, X. L., RAGHUNATHAN, T. E. & RUBIN, D. B. (1991). Significance levels from repeated p-values with multiply-imputed data. *Statistica Sinica* **1**, 65–92.

LITTLE, R. J. A. & RUBIN, D. B. (2002). *Statistical Analysis with Missing Data*. Hoboken, N.J.: Wiley, 2nd ed.

MCCULLAGH, P. & NELDER, J. A. (1998). *Generalized Linear Models*. London: Chapman & Hall/CRC, 2nd ed.

MENG, X. L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science* **9**, 538–558.

MEYN, S. P. & TWEEDIE, R. L. (1993). *Markov Chains and Stochastic Stability*. New York: Springer-Verlag.

RAGHUNATHAN, T. E., SOLENBERGER, P. W. & VAN HOEWYK, J. (2010). *IVEware*. University of Michigan.

ROBINS, J. M. & WANG, N. S. (2000). Inference for imputation estimators. *Biometrika* **87**, 113–124.

ROSENTHAL, J. S. (1995). Minorization conditions and convergence-rates for Markov-chain Monte-Carlo. *Journal of the American Statistical Association* **90**, 558–566.

ROYSTON, P. (2004). Multiple imputation of missing values. *Stata Journal* **4**, 227–241.

ROYSTON, P. (2005). Multiple imputation of missing values. *Stata Journal* **5**, 1–14.

RUBIN, D. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

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

SCHAFER, J. L. (1997). *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall/CRC.

SCHENKER, N. & WELSH, A. H. (1988). Asymptotic results for multiple imputation. *Annals of Statistics* **16**, 1550–1566.

SU, Y.-S., GELMAN, A., HILL, J. & YAJIMA, M. (2011). Multiple imputation with diagnostics (mi) in R: Opening windows into the black box. *Journal of Statistical Software* **45**, 1–31.

VAN BUUREN, S. & GROOTHUIS-OUDSHOORN, K. (2011). Mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software* **45**.

VAN DYK, D. & PARK, T. (2008). Partially collapsed Gibbs samplers: theory and methods. *Journal of the American Statistical Association* **103**, 790–796.

WANG, N. & ROBINS, J. M. (1998). Large-sample theory for parametric multiple imputation procedures. *Biometrika* **85**, 935–948.

[*Received . Revised*]