On the Stationary Distribution of Iterative Imputations

Jingchen Liu, Andrew Gelman, Jennifer Hill, and Yu-Sung Su

Columbia University, Columbia University, New York University, Tsinghua University

February 21, 2012

Abstract

Iterative imputation, in which variables are imputed one at a time each given a model predicting from 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. More precisely, when the conditional models are compatible (defined in the text), we give a set of 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 are valid, we show that the combined imputation estimator is consistent.

1 Introduction

Iterative imputation is a widely used approach for imputation 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 (as measured, for example, by the mixing of multiple chains).

Iterative imputation can be an appealing way to express uncertainty about missing data. 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 need specify a sequence of families of conditional regression models such as linear regrsssion, logistic regression, and other standard and already programmed forms. The distribution of the resulting imputations is implicitly defined as the invariant (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` [28] and `mi` [26] in `R`, `IVEware` [16] in `SAS`, and `ice` in `Stata` [19, 20]. 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, as we would prefer, the fitting of each imputation model and the imputations themselves are performed using conditional Bayesian inference, but the stationary distribution of the algorithm (if it exists) does not in general correspond to Bayesian inference on any specified multivariate distribution.

Key questions are: (1) Under what conditions does the algorithm converge to a stationary distribution? (2) What statistical properties does the procedure admit given that a unique stationary distribution exists?

Regarding the first question, researchers have long known that the Markov chain may be nonrecurrent ("blowing up" to infinity or drifting in a nonstationary random walk), even if each of the conditional models is fitted using a proper prior distribution. analysis.

In this paper, we focus mostly on the second question—the characterization of the stationary distributions of the iterative imputation conditional its existence. Unlike usual MCMC algorithms, which are designed in such a way that the invariant distribution and target distribution are identical, the invariant distribution of iterative imputation (even if it exists) is largely unknown.

The analysis of iterative imputation is challenging for at least two reasons. First, the range of choices of conditional models is wide so that it is difficult to provide a solution applicable to all situations. Second, the literature on Markov chains focuses on known transition distributions. With iterative imputation, the distributions for the imputations are known only within specified parametric familis. 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 main contribution of this paper is to develop a mathematical framework under which the asymptotic properties of iterative imputation can be discussed. In particular, we demonstrate the following results.

1. Given the existence of a unique invariant (stationary) distribution of the iterative imputation Markov chain, 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. Among these conditions, the most important is that the conditional models are *compatible*—that there exists a joint model whose conditional distributions are identical to the conditional models specified by the iterative imputation (Definition 1). This discussion is in Section 3.

2. We consider model compatibility as a usually necessary condition for the iterative imputation distribution to converge to the posterior distribution of some Bayesian model (Section 3.4).

3. For *incompatible* models whose imputation distributions are generally different from any

Bayesian model, we show that the expectation of the imputed data maximum likelihood estimate under the imputation distribution is a consistent estimator if the set of conditional models is valid, that is, if each conditional model contains the true probability distribution (Definition 3 in Section 4.).

The analysis presented in this paper connects to the existing separate literatures on missing data imputation and Markov chain convergence. Standard textbooks on impuation inference are [22, 11], and some key papers are [14, 13, 10, 3, 21, 23, 24]. Large sample properties are studied by [25, 29, 17], small samples are by [3], and the issue of congeniality between the imputer's and analyst's models is considered by [13].

Our asymptotic results for compatible and incompatible models use results on convergence of Markov chains, a subject for which there is a vast literature on stability and rate of convergence. General results on the exponential convergence rate appear in [8]. For specific bounds on convergence rates of specific models, see [2, 1]. In addition, empirical diagnostics of Markov chains have beensuggested by many authors, for instance, [7]. In the example of this paper (cf. the illustrative example in Section 5), we prove stability and construct a bound for convergence rate using renewal theory ([15, 18, 4]), which has the advantage of not assuming the existence of an invariant distribution, which is naturally yielded by the minorization and drift conditions.

In Section 2 of this article, we lay out our notation and assumptions. We then briefly review the framework of iterative imputation and the Gibbs sampler. In Section 3, we investigate compatible conditional models. In Section 4, the discussion focuses on incompatible models. Section 5 considers one linear example in detail. An appendix is attached containing the technical developments and a brief review of the literature for Markov chain convergence via renewal theory.

2 Background

Consider a data set with n cases and p variables, where $\mathbf{x} = (\mathbf{x}_1, ..., \mathbf{x}_p)$ represents the complete data and $\mathbf{x}_i = (x_{1,i}, ..., x_{n,i})^\top$ is the i-th variable. Let $\mathbf{r}_i$ be the vector of observed data indicators for variable i, equaling 1 for observed variables and 0 for missing, and let $\mathbf{x}_i^{obs}$ and $\mathbf{x}_i^{mis}$ denote the observed and missing subsets of variable i:

$$\mathbf{x}^{obs} = \{\mathbf{x}_i^{obs} : i = 1, ..., p\}, \quad \mathbf{x}^{mis} = \{\mathbf{x}_i^{mis} : i = 1, ..., p\}, \quad \mathbf{r} = \{\mathbf{r}_i : i = 1, ..., p\}.$$

To facilitate our description of the procedures, we define

$$\mathbf{x}_{-j}^{obs} = \{\mathbf{x}_i^{obs} : i = 1, ..., j-1, j+1, ..., p\}, \quad \mathbf{x}_{-j}^{mis} = \{\mathbf{x}_i^{mis} : i = 1, ..., j-1, j+1, ..., p\}.$$

We use boldface $\mathbf{x}$ to denote the entire data set and x to denote individual observations. Therefore, x_j denotes the j-th variable of one observation and x_{-j} denotes all the variables except for the j-th one.

Throughout, we assume that the missing data process is ignorable. One set of sufficient conditions for ignorability is that the $\mathbf{r}_i$ process is missing at random and the parameter spaces for $\mathbf{r}_i$ and $\mathbf{x}$ are distinct, with independent prior distributions [11, 22].

2.1 Inference using multiple imputations

Multiple imputation is a convenient tool to handle incomplete data set by means of complete-data procedures. The framework consists of producing m copies of the imputed data and applying the users' complete data procedures to each of the multiply imputed data sets. Suppose that m copies of point estimates and variance estimates are obtained, denoted by $(\hat{\theta}^{(i)}, U^{(i)})$, $i = 1, ..., m$. The next step is to combine them into a single point estimate and a single variance estimate $(\hat{\theta}_m, \hat{T}_m)$ [11]. If the imputed data are drawn from the joint posterior distribution of the missing data under a Bayesian model, under appropriate congeniality conditions, $\hat{\theta}_m$ is asymptotically equal to the posterior mean of θ and $\hat{T}_m$ is asymptotically equal to the posterior variance of θ ([22, 13]). The large sample theory of Bayesian inference ensures that the posterior mean and variance are asymptotically equivalent to the maximum likelihood estimate and its variance based on the observed data alone (see [5]). Therefore, the combined estimator from imputed samples is efficient. Imputations can also be constructed and used under other inferential frameworks; for example, Robins and Wang ([17, 29]) propose estimates based on estimating equations and derive corresponding combining rules. For our purposes here, what is relevant is that the multiple imputations are being used to represent uncertainty about the joint distribution of missing values in a multivariate dataset.

2.2 Bayesian modeling, imputation, and Gibbs sampling

In Bayesian inference, multiply imputed data sets are treated as samples from the posterior distribution of the full (incompletely-observed) data matrix. In the parametric Bayesian approach, one specifies a family of distributions $f(\mathbf{x}|\theta)$ and a prior $\pi(\theta)$ and then performs inference using theare i.i.d. samples from the posterior predictive distribution,

$$f(\mathbf{x}^{mis}|\mathbf{x}^{obs}) = \int_{\Theta} f(\mathbf{x}^{mis}|\mathbf{x}^{obs}, \theta)p(\theta|\mathbf{x}^{obs})d\theta, \quad (1)$$

where $p(\theta|\mathbf{x}) \propto p(\theta)f(\mathbf{x}|\theta)$. 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 (MCMC) algorithm. In the scenario of missing data, one can use the "data augmentation" strategy to iteratively draw θ given $(\mathbf{x}^{obs}, \mathbf{x}^{mis})$ and $\mathbf{x}^{mis}$ given $(\mathbf{x}^{obs}, \theta)$. Under regularity conditions (positive recurrence, irreducibility, and aperiodicity; see [8]), the Markov process is ergodic with limiting distribution $p(\mathbf{x}^{mis}, \theta|\mathbf{x}^{obs})$.

In order to connect these results to the iterative imputation that is the subject of the present article, we consider a slightly different Gibbs scheme which consists of indefinite iteration of following p steps:,

Step 1. Draw $\theta \sim p(\theta|\mathbf{x}_1^{obs}, \mathbf{x}_{-1})$ and $\mathbf{x}_1^{miss} \sim f(\mathbf{x}_1^{miss}|\mathbf{x}_1^{obs}, \mathbf{x}_{-1}, \theta)$;

Step 2. Draw $\theta \sim p(\theta|\mathbf{x}_2^{obs}, \mathbf{x}_{-2})$ and $\mathbf{x}_2^{miss} \sim f(\mathbf{x}_2^{miss}|\mathbf{x}_2^{obs}, \mathbf{x}_{-2}, \theta)$;

$\vdots$

Step p. Draw $\theta \sim p(\theta|\mathbf{x}_p^{obs}, \mathbf{x}_{-p})$ and $\mathbf{x}_p^{miss} \sim f(\mathbf{x}_p^{miss}|\mathbf{x}_p^{obs}, \mathbf{x}_{-p}, \theta)$.

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 the Markov chain evolving according to steps 1 to p (under mild regularity conditions) converges to the posterior distribution of the corresponding Bayesian model.

2.3 Iterative imputation and compatibility

For iterative imputation, we need to specify p conditional models,

$$g_j(\mathbf{x}_j|\mathbf{x}_{-j}, \theta_j),$$

for $\theta_j \in \Theta_j$ with prior distributions $\pi_j(\theta_j)$ for $j = 1, ..., p$. Iterative imputation adopts the following scheme to construct a Markov chain,

Step 1. Draw θ_1 from $p_1(\theta_1|\mathbf{x}_1^{obs}, \mathbf{x}_{-1})$, which is the posterior distribution associated with g_1 and π_1; draw $\mathbf{x}_1^{miss}$ from $g_1(\mathbf{x}_1^{miss}|\mathbf{x}_1^{obs}, \mathbf{x}_{-1}, \theta_1)$;

Step 2. Draw θ_2 from $p_2(\theta_2|\mathbf{x}_2^{obs}, \mathbf{x}_{-2})$, which is the posterior distribution associated with g_2 and π_2; draw $\mathbf{x}_2^{miss}$ from $g_2(\mathbf{x}_2^{miss}|\mathbf{x}_2^{obs}, \mathbf{x}_{-2}, \theta_2)$;

$\vdots$

Step p. Draw θ_p from $p_p(\theta_p|\mathbf{x}_p^{obs}, \mathbf{x}_{-p})$, which is the posterior distribution associated with g_p and π_p; draw $\mathbf{x}_p^{miss}$ from $g_p(\mathbf{x}_p^{miss}|\mathbf{x}_p^{obs}, \mathbf{x}_{-p}, \theta_p)$.

Iterative imputation has the practical advantage that, at each step, one only needs to set up a sensible regression model of $\mathbf{x}_j$ given $\mathbf{x}_{-j}$. 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 $\mathbf{x}$. Whether it is preferable to perform p easy task 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.

But iterative imputation has conceptual problems. Except in some special cases, there will not in general exist a joint distribution of $\mathbf{x}$ such that $f(\mathbf{x}_j|\mathbf{x}_{-j},\theta) = g_j(\mathbf{x}_j|\mathbf{x}_{-j},\theta_j)$ 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.

In this paper, we discuss the properties of the stationary distribution of the iterative imputation process by first classifying the set of conditional models as compatible (defined as there existing a joint model f which is consistent with all the conditional models) or incompatible.

We refer to the Markov chain generated by the scheme in Section 2.2 as the *Gibbs chain* and that generated by the scheme in Section 2.3 as the *iterative chain*. Our central analysis works by coupling the two.

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* $\{g_j(x_j|x_{-j},\theta_j) : \theta_j \in \Theta_j, j = 1,...,p\}$ *is said to be* compatible *if there exists a joint model* $\{f(x|\theta) : \theta \in \Theta\}$ *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* $\theta \in t_j^{-1}(\theta_j) = \{\theta : t_j(\theta) = \theta_j\}$,

$$g_j(x_j|x_{-j},\theta_j) = f(x_j|x_{-j},\theta).$$

Otherwise, $\{g_j : j = 1,...,p\}$ *is said to be* incompatible.

Though imposing certain restrictions, compatible models do include quite a collection of procedures practically in use (e.g. `ice` in `Stata`). In what follows, we give a few examples of compatible and incompatible conditional models.

We begin with a simple linear model, which we shall revisit in Section 5.

Example 1 (bivariate Gaussian) *Consider a binary continuous variable* (x, y) *and conditional models*

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

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

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

with $\sigma_x, \sigma_y > 0$ *and* $\rho \in [-1, 1]$. *The reparameterization from* $(\mu_x, \mu_y, \sigma_x, \sigma_y, \rho)$ *to the parameters of the conditional models is:*

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

The following example is a natural extension.

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

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

where β_j *is a* $p \times 1$ *vector,* $\mathbf{1} = (1, ..., 1)^\top$. *Consider the joint model of* $(x_1, ..., x_p) \overset{i.i.d.}{\sim} N(\mu, \Sigma)$. *Then the conditional distribution of each* x_j *given* x_{-j} *is Gaussian. The maps* t_j*'s 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 as follows:*

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

The above conditional models are compatible with the following joint model:

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

If we let

$$\begin{aligned} t_1(p, \beta_0, \beta_1, \sigma^2) &= \left(\log\frac{p}{1-p} - \frac{\beta_1^2}{2\sigma^2}, \frac{\beta_1}{2\sigma^2}\right) = (\alpha, \beta) \\ t_2(p, \beta_0, \beta_1, \sigma^2) &= (\beta_0, \beta_1), \end{aligned}$$

the conditional models and this joint model are compatible with each other. Similarly compatible models can be defined for other natural exponential families. See [6, 12].

Example 4 (incompatible Gaussian conditionals) *There are many incompatible conditional models. For instance,*

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

are compatible only if $\beta_2 = 0$.

3.2 Total variation distance between two transition kernels

Let $\{\mathbf{x}^{mis,1}(k) : k \in \mathbb{Z}^+\}$ be the Gibbs chain and $\{\mathbf{x}^{mis,2}(k) : k \in \mathbb{Z}^+\}$ be the iterative chain. Both chains live on the space of the missing data. We write the completed data as $\mathbf{x}^i(k) = (\mathbf{x}^{mis,i}(k), \mathbf{x}^{obs})$

for the Gibbs chain ($i = 1$) and the iterative chain ($i = 2$). The transition kernels are

$$K_i(w, dw') = P(\mathbf{x}^{mis,i}(k+1) \in dw' | \mathbf{x}^{mis,i}(k) = w), \text{ for } i = 1, 2. \tag{2}$$

where w is a generic notation for the state of the processes. The transition kernels (K_1 and K_2) depend on $\mathbf{x}^{obs}$. For simplicity, we omit the index of $\mathbf{x}^{obs}$ in the notation of K_i. Also, we let

$$K_i^{(k)}(\nu, A) \triangleq P_\nu(\mathbf{x}^{mis,i}(k) \in A),$$

for $\mathbf{x}^{mis,i}(0) \sim \nu$, ν being some starting distribution. The probability measure P_ν also depends on $\mathbf{x}^{obs}$. 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

$$d_{TV}(\nu_1, \nu_2) = \sup_{A \in \mathcal{F}} |\nu_1(A) - \nu_2(A)|.$$

We further define

$$\|\nu\|_V = \sup_{|h| \le V} \int h(x)\nu(dx)$$

and $\|\nu\|_{\mathbf{1}} = \|\nu\|_V$ for $V \equiv 1$. Let $\nu_i^{\mathbf{x}^{obs}}$ be the stationary distribution of K_i. We intend to establish conditions under which

$$d_{TV}(\nu_1^{\mathbf{x}^{obs}}, \nu_2^{\mathbf{x}^{obs}}) \to 0$$

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 $\mathbf{x}^{obs}$), that is, $\|K_1(w, \cdot) - K_2(w, \cdot)\|_{\mathbf{1}} \to 0$ 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 2.2 and 2.3, it suffices to compare the following posterior predictive distributions (for each $j = 1, ..., p$),

$$f(\mathbf{x}_j^{mis} | \mathbf{x}_j^{obs}, \mathbf{x}_{-j}) = \int f(\mathbf{x}_j^{mis} | \mathbf{x}_j^{obs}, \mathbf{x}_{-j}, \theta) p(\theta | \mathbf{x}_j^{obs}, \mathbf{x}_{-j}) d\theta \tag{3}$$

$$g_j(\mathbf{x}_j^{mis} | \mathbf{x}_j^{obs}, \mathbf{x}_{-j}) = \int g_j(\mathbf{x}_j^{mis} | \mathbf{x}_j^{obs}, \mathbf{x}_{-j}, \theta_j) p_j(\theta_j | \mathbf{x}_j^{obs}, \mathbf{x}_{-j}) d\theta_j, \tag{4}$$

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:

$$f(\mathbf{x}_j^{mis}|\mathbf{x}_j^{obs},\mathbf{x}_{-j},\theta) = g_j(\mathbf{x}_j^{mis}|\mathbf{x}_j^{obs},\mathbf{x}_{-j},\theta_j),$$

if $t_j(\theta) = \theta_j$. The only difference lies in their posterior distributions. In fact, the $\|\cdot\|_1$ distance between two posterior predictive distributions is bounded by the distance between the posterior distributions of parameters. Therefore, we move to comparing $p(\theta|\mathbf{x}_j^{obs},\mathbf{x}_{-j})$ and $p_j(\theta_j|\mathbf{x}_j^{obs},\mathbf{x}_{-j})$.

Parameter augmentation. 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 a lower dimension. Consider the linear model in Example 1. The conditional models include three parameters (two regression coefficients and variance of the errors), while the joint model has five parameters $(\mu_x,\mu_y,\sigma_x,\sigma_y,\rho)$. This is because the (conditional) regression models are usually conditional on the covariates. The joint model not only parameterizes the conditional distributions of $\mathbf{x}_j$ given $\mathbf{x}_{-j}$ but also the marginal distribution of $\mathbf{x}_{-j}$. 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-degenerated 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 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,

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

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^*.

Impact of the prior distribution. Because we are assuming compatibility, we can drop the notation g_j for conditional model of the j-th variable. Instead, we unify the notation to that of the joint Bayesian model $f(\mathbf{x}_j|\mathbf{x}_{-j},\theta)$. In addition, we abuse the notation and write $f(\mathbf{x}_j|\mathbf{x}_{-j},\theta_j) = f(\mathbf{x}_j|\mathbf{x}_{-j},\theta)$ for $t_j(\theta)=\theta_j$. For instance, in Example 1, we write $f(y|x,\alpha_{y|x},\beta_{y|x},\sigma_{y|x}) = f(y|x,\mu_x,\mu_y,\sigma_x,\sigma_y,\rho)$ as long as $\alpha_{y|x} = \mu_y - \frac{\rho\sigma_y}{\sigma_x}\mu_x$, $\beta_{y|x} = \frac{\rho\sigma_y}{\sigma_x}$, and $\sigma_{y|x}^2 = (1-\rho^2)\sigma_y^2$.

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

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

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

$$p(\theta_j|\mathbf{x}_j^{obs}, \mathbf{x}_{-j}) = \int p(\theta_j, \theta_j^*|\mathbf{x}_j^{obs}, \mathbf{x}_{-j})d\theta_j^* \propto \int f(\mathbf{x}_j^{obs}, \mathbf{x}_{-j}|\theta_j, \theta_j^*)\pi_j^*(\theta_j, \theta_j^*)d\theta_j^*.$$

Because $f(\mathbf{x}_j^{obs}|\mathbf{x}_{-j}, \theta_j, \theta_j^*) = f(\mathbf{x}_j^{obs}|\mathbf{x}_{-j}, \theta_j)$, the above posterior distribution can be further reduced to

$$p(\theta_j|\mathbf{x}_j^{obs}, \mathbf{x}_{-j}) \propto f(\mathbf{x}_j^{obs}|\mathbf{x}_{-j}, \theta_j)\int f(\mathbf{x}_{-j}|\theta_j, \theta_j^*)\pi_j^*(\theta_j, \theta_j^*)d\theta_j^*.$$

If we write

$$\pi_{j,\mathbf{x}_{-j}}(\theta_j) \triangleq \int f(\mathbf{x}_{-j}|\theta_j, \theta_j^*)\pi_j^*(\theta_j, \theta_j^*)d\theta_j^*,$$

then the posterior distribution of θ_j under the joint Bayesian model is

$$p(\theta_j|\mathbf{x}_j^{obs}, \mathbf{x}_{-j}) \propto f(\mathbf{x}_j^{obs}|\mathbf{x}_{-j}, \theta_j)\pi_{j,\mathbf{x}_{-j}}(\theta_j).$$

Compared with the posterior distribution of the iterative imputation procedure, which is proportional to

$$p_j(\theta_j|\mathbf{x}_j^{obs}, \mathbf{x}_{-j}) \propto g_j(\mathbf{x}_j^{obs}|\mathbf{x}_{-j}, \theta_j)\pi_j(\theta_j) = f(\mathbf{x}_j^{obs}|\mathbf{x}_{-j}, \theta_j)\pi_j(\theta_j),$$

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

Controlling the distance between the posterior predictive distributions. We put forward tools to control the distance between the two posterior predictive distributions in (3) and (4). Let $\mathbf{x}$ be the generic notation the observed data, and let $f_{\mathbf{X}}(\theta)$ and $g_{\mathbf{X}}(\theta)$ be two posterior densities of θ. Let $h(\tilde{x}|\theta)$ be the density function for future observations given the parameter θ, and let $\tilde{f}_{\mathbf{X}}(\tilde{x})$ and $\tilde{g}_{\mathbf{X}}(\tilde{x})$ be the posterior predictive distributions:

$$\tilde{f}_{\mathbf{X}}(\tilde{x}) = \int h(\tilde{x}|\theta)f_{\mathbf{X}}(\theta)d\theta, \quad \tilde{g}_{\mathbf{X}}(\tilde{x}) = \int h(\tilde{x}|\theta)g_{\mathbf{X}}(\theta)d\theta.$$

It is straightforward to obtain that

$$\|\tilde{f}_{\mathbf{X}} - \tilde{g}_{\mathbf{X}}\|_1 \leq \|f_{\mathbf{X}} - g_{\mathbf{X}}\|_1. \tag{5}$$

The next proposition provides sufficient conditions that $\|f_{\mathbf{X}} - g_{\mathbf{X}}\|_1$ vanishes.

Proposition 1 *Let n be the sample size. Let $f_{\mathbf{X}}(\theta)$ and $g_{\mathbf{X}}(\theta)$ be two posterior density functions that share the same likelihood but have two different prior distributions π_f and π_g. Let*

$$L(\theta) = \frac{\pi_g(\theta)}{\pi_f(\theta)}, \qquad r(\theta) = \frac{g_{\mathbf{X}}(\theta)}{f_{\mathbf{X}}(\theta)} = \frac{L(\theta)}{\int L(\theta)f_{\mathbf{X}}(\theta)d\theta},$$

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

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

where " $\cdot$ " denotes inner product and $\mu_\theta = \int \theta f_{\mathbf{X}}(\theta) d\theta$. If there exists a random variable $Z(\theta)$ with finite variance under $f_{\mathbf{X}}$, such that

$$\left|\sqrt{n} \partial L(\xi) \cdot (\theta - \mu_\theta)\right| \leq |\partial L(\mu_\theta)| Z(\theta), \tag{6}$$

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

$$\|\tilde{f}_{\mathbf{X}} - \tilde{g}_{\mathbf{X}}\|_{\mathbf{1}} \leq \frac{\kappa \sqrt{|\partial \log L(\mu_\theta)|}}{n^{1/4}}. \tag{7}$$

We prove this proposition in Appendix A.

Remark 1 *We adapt Proposition 1 to the analysis of the conditional models. Expresion (6) implies that the posterior variance 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 $\hat{\theta}(\mathbf{x})$ be the complete-data MLE and $A_n = \{\mathbf{x} : |\hat{\theta}(\mathbf{x})| \leq \gamma\}$. 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 $L(\theta_j) = \pi_j(\theta_j)/\pi_{j,\mathbf{x}_{-j}}(\theta_j)$.*

The expression $\partial \log L(\mu_\theta)$ is not always bounded. For example, suppose $\pi_{j,\mathbf{x}_{-j}}(\theta_j)$ is an informative prior for which the covariates of the regression model $\mathbf{x}_{-j}$ provide strong information on the regression coefficients θ_j, as in Example 3 on page 7. The marginal distribution of the covariate x_2 provides $\sqrt{n}$ amount of information on the coefficients of the logistic regression. Under this situation, Proposition 1 still holds, but the right-hand side of (7) does not necessarily converge to zero. Consequently, Theorem 1 (the main result in this section, presented later) does not apply. For the properties of iterative imputation under such situations, we can apply the consistency result for incompatible models, as discussed in Section 4.

Remark 3 *If $L(\theta)$ is twice differentiable, the convergence rate in (7) can be improved to $O(n^{-1/2})$. However, $O(n^{-1/4})$ is sufficient for the current analysis.*

3.3 Convergence of the invariant distributions

With the results of 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 alarge region A_n. The subsequent analysis falls into several steps. First, we slightly modify the processes by conditioning them on the set A_n with stationary distributions $\tilde{\nu}_i^{\mathbf{X}^{obs}}$ (details provided below). The stationary distributions of the conditional processes and the original processes ($\tilde{\nu}_i^{\mathbf{X}^{obs}}$ and $\nu_i^{\mathbf{X}^{obs}}$) are close in total variation. Second, we show (in Lemma 2) that, with a bound on the convergence rate, $\tilde{\nu}_1^{\mathbf{X}^{obs}}$ and $\tilde{\nu}_2^{\mathbf{X}^{obs}}$ are close in total variation and so it is with $\nu_1^{\mathbf{X}^{obs}}$ and $\nu_2^{\mathbf{X}^{obs}}$. The bound of convergence rate can be verified 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 $\mathbf{x} \in A_n$. The next lemma shows 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*

$$\nu_i^{\mathbf{X}^{obs}}(A_n) \to 1 \tag{9}$$

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

$$d_{TV}(\nu_i^{\mathbf{X}^{obs}}, \tilde{\nu}_i^{\mathbf{X}^{obs}}) \to 0, \tag{10}$$

as $n \to \infty$.

The proof is elementary by the representation of $\nu_i^{\mathbf{X}^{obs}}$ through the renewal theory and therefore is omitted. Based on the above lemma, we only need to show that $d_{TV}(\tilde{\nu}_1^{\mathbf{X}^{obs}}, \tilde{\nu}_2^{\mathbf{X}^{obs}}) \to 0$. The expression $\|K_1(w, \cdot) - K_2(w, \cdot)\|_{\mathbf{1}}$ approaches 0 uniformly for $w \in A_n$. This implies that

$$\|\tilde{K}_1(w, \cdot), \tilde{K}_2(w, \cdot)\|_{\mathbf{1}} \to 0$$

as $n \to \infty$ uniformly for $w \in A_n$. With the above convergence, we use the following lemma to establish the convergence between $\tilde{\nu}_1^{\mathbf{X}^{obs}}$ and $\tilde{\nu}_2^{\mathbf{X}^{obs}}$.

Lemma 2 *Let $\tilde{\mathbf{x}}^{mis,i}(k)$ 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 $\tilde{\nu}_i^{\mathbf{X}^{obs}}$, and that the following two conditions hold:*

1. *The convergence of the two transition kernels*

$$d(A_n) \triangleq \sup_{w \in A_n} \|\tilde{K}_1(w,\cdot) - \tilde{K}_2(w,\cdot)\|_V \to 0, \tag{11}$$

in probability as $n \to \infty$. *The function* V *is either a geometric drift function for* $\tilde{K}_2$ *or a constant, i.e.,* $V = 1$.

2. *Furthermore, there exists a monotone decreasing sequence* $r_k \to 0$ *(independent of data) and a starting measure* ν *(depending on data) such that*

$$P\left[\|\tilde{K}_i^{(k)}(\nu,\cdot) - \tilde{\nu}_i^{\mathbf{X}^{obs}}(\cdot)\|_V \leq r_k, \forall k > 0\right] \to 1, \tag{12}$$

as $n \to \infty$.

Then,

$$\|\tilde{\nu}_1^{\mathbf{X}^{obs}} - \tilde{\nu}_2^{\mathbf{X}^{obs}}\|_V \to 0, \tag{13}$$

in probability as $n \to \infty$.

Remark 4 *The above lemma holds if* $V = 1$ *or* V *is a drift function. For the analysis of convergence in total variation, we only need that* $V = 1$. *The results when* V *is a drift function is prepared for the analysis of incompatible models.*

The first condition in the above lemma can be obtained by the result of Proposition 1. Condition (12) is more difficult to establish. According to the standard results in [18] (see also (32) in the appendix), one set of sufficient conditions for (12) is that the chains $\tilde{K}_1$ and $\tilde{K}_2$ admit a common small set, C; in addition, each of them admits their own drift functions associated with the small set C (c.f. Appendix C).

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

$$\tilde{K}_1(w, A) \geq q_1 \mu_1(A), \tag{14}$$

for $w \in C$, $q_1 \in (0,1)$; and for all $w \notin C$

$$\lambda_1 V(w) \geq \int V(w')\tilde{K}_1(w, dw'). \tag{15}$$

With the existence of C and V a bound of convergence r_k (with starting point $w \in C$) can be established for the Gibbs chain by standard results (see, for instance, [18]), 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 $\mathbf{x}^{obs}$. 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 "$\|\cdot\|_{\mathbf{1}}$", 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 $q_2 \in (0,1)$, all $w \in C$, and all measurable set A. The following proposition, whose proof is given in the appendix, establishes the conditions under which V is also a drift function for $\tilde{K}_2$.

Proposition 2 *Assume the following conditions hold.*

1. *The transition kernel* $\tilde{K}_1$ *admits a small set* C *and a drift function* V *satisfying* (15).

2. *Let* $L_j(\theta_j) = \pi_j(\theta_j)/\pi_{j,\mathbf{x}_{-j}}(\theta_j)$ $(j = 1, ..., p)$ *be the the ratio of prior distributions for each conditional model (possibly depending on the data) so that on the set* A_n $\sup_{|\theta_j|<\gamma} \partial L_j(\theta_j)/L_j(\theta_j) < \infty$.

3. *For each* j*, there exists a* $Z_j(\theta_j)$ *serving as the bound in* (6) *for each* L_j *such that*

$$\tilde{E}_w \left[Z_j^2(\theta_j) V^2(w') \right] = o(n) V^2(w), \tag{16}$$

where $\tilde{E}_w$ *is the expectation of one step transition under* $\tilde{K}_1$ *with starting value* w.

Then, there exits $\lambda_2 \in (0, 1)$ *such that as* n *tends to infinity with probability converging to one the following inequality holds*

$$\lambda_2 V(w) \geq \int V(w') \tilde{K}_2(w, dw'). \tag{17}$$

Remark 5 *Since the* $\tilde{K}_1$ *and* $\tilde{K}_2$ *are close to each other, the above lemmas suggest that* V *be a drift function of* $\tilde{K}_2$ *if it is a drift function of* $\tilde{K}_1$ *to the same small set* C*. Condition* (16) *is imposed for a technical purpose. In particular, we allow the expectation of* $Z_j^2(\theta_j)V^2(w')$ *to grow to infinity but at a slower rate than* n*. Therefore, it is a mild condition.*

We now summarize the results of the compatible conditional models in the following theorem.

Theorem 1 *Suppose that a set of conditional models* $\{g_j(x_j|x_{-j}, \theta_j) : \theta_j \in \Theta_j, j = 1, ..., p\}$ *is compatible with a joint model* $\{f(x|\theta) : \theta \in \Theta\}$*. The Gibbs chain and the iterative chain then admit transition kernels* K_i *and unique stationary distributions* $\nu_i^{\mathbf{X}^{obs}}$*. Suppose the following conditions are satisfied:*

- *A1 Let* $A_n = \{\mathbf{x} : |\hat{\theta}(\mathbf{x})| \leq \gamma\}$*. One can choose* γ *sufficiently large so that* $\nu_i^{\mathbf{X}^{obs}}(A_n) \to 0$*, in probability as* $n \to \infty$.

- *A2 The conditions in Proposition 2 hold.*

- *A3 Let* $L_j(\theta_j) = \pi_j(\theta_j)/\pi_{j,\mathbf{x}_{-j}}(\theta_j)$ *and* $\partial \log L_j(\theta_j)$ *is bounded by a constant for all* $|\theta_j| < \gamma$ *and* $\mathbf{x} \in A_n$.

Then,

$$d_{TV}(\nu_1^{\mathbf{X}^{obs}}, \nu_2^{\mathbf{X}^{obs}}) \to 0$$

in probability as $n \to \infty$.

Remark 6 *One sufficient condition for A1 is that the stationary distributions of* $\hat{\theta}(\mathbf{x})$ *under* $\nu_i^{\mathbf{X}^{obs}}$ *converge to a value* θ^i*, where* θ^1 *and* θ^2 *are not necessarily the same.*

Remark 7 *Condition A2 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. For illustration purpose, we construct a drift function (in the appendix) for the linear example in Section 5.*

Proof of Theorem 1. 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 "$\|\cdot\|_{\mathbf{1}}$" 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 condition A3, 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, $\|\tilde{K}_1(w,\cdot)-\tilde{K}_2(w,\cdot)\|_{\mathbf{1}} \to 0$. 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 (14) holds for both chains and a drift function V for one of the two chains such that (15) 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 (for instance, [18] and (32) in the appendix), there exists a common starting value $w \in C$ and a bound r_k such that the bound (12) in Lemma 2 is satisfied. Thus, Lemma 2 implies that

$$d_{TV}(\tilde{\nu}_1^{\mathbf{X}^{obs}}, \tilde{\nu}_2^{\mathbf{X}^{obs}}) \to 0,$$

in probability as $n \to \infty$. According to condition A1 and Lemma 1, the above convergence implies that

$$d_{TV}(\nu_1^{\mathbf{X}^{obs}}, \nu_2^{\mathbf{X}^{obs}}) \to 0.$$

Thereby, we conclude the analysis. ■

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

$$\begin{aligned}
P^f(\mathbf{x}_j^{mis} \in A|\mathbf{x}_{-j}^{mis}, \mathbf{x}^{obs}) &= \int_A f(\mathbf{x}_j^{mis}|\mathbf{x}_{-j}^{mis}, \mathbf{x}^{obs}, \theta)p(\theta|\mathbf{x}_{-j}^{mis}, \mathbf{x}^{obs})d\theta \\
P^g_j(\mathbf{x}_j^{mis} \in A|\mathbf{x}_{-j}^{mis}, \mathbf{x}^{obs}) &= \int_A g_j(\mathbf{x}_j^{mis}|\mathbf{x}_{-j}^{mis}, \mathbf{x}^{obs}, \theta)p_j(\theta|\mathbf{x}_{-j}^{mis}, \mathbf{x}^{obs})d\theta.
\end{aligned}$$

Furthermore, denote the stationary distributions of the Gibbs chain and the iterative chain by $\nu_1^{\mathbf{X}^{obs}}$ and $\nu_2^{\mathbf{X}^{obs}}$.

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

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

or

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

and $\nu_1^{\mathbf{X}^{obs}}(\mathbf{x}_{-j}^{mis} \in C) > q \in (0, 1)$.

Then there exists a set B such that

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

Proof. *Suppose that*

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

The "less than" case is completely analogous. Consider the set $B = \{\mathbf{x}^{mis} : \mathbf{x}_{-j}^{mis} \in C, \mathbf{x}_j^{mis} \in A\}$. If

$$|\nu_2^{\mathbf{X}^{obs}}(\mathbf{x}_{-j}^{mis} \in C) - \nu_1^{\mathbf{X}^{obs}}(\mathbf{x}_{-j}^{mis} \in C)| \leq q\varepsilon/2, \tag{18}$$

then, by the fact that

$$\begin{aligned}
\nu_1^{\mathbf{X}^{obs}}(\mathbf{x}^{mis} \in B) &= \nu_1^{\mathbf{X}^{obs}}(\mathbf{x}_{-j}^{mis} \in C)\int P^f(\mathbf{x}_j^{mis} \in A|\mathbf{x}_{-j}^{mis}, \mathbf{x}^{obs})\nu_1^{\mathbf{X}^{obs}}(d\mathbf{x}_{-j}^{mis}|\mathbf{x}_{-j}^{mis} \in C), \\
\nu_2^{\mathbf{X}^{obs}}(\mathbf{x}^{mis} \in B) &= \nu_2^{\mathbf{X}^{obs}}(\mathbf{x}_{-j}^{mis} \in C)\int P^g(\mathbf{x}_j^{mis} \in A|\mathbf{x}_{-j}^{mis}, \mathbf{x}^{obs})\nu_2^{\mathbf{X}^{obs}}(d\mathbf{x}_{-j}^{mis}|\mathbf{x}_{-j}^{mis} \in C),
\end{aligned}$$

we obtain

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

Otherwise, if (18) does not hold, let $B = \{\mathbf{x}^{mis} : \mathbf{x}_{-j}^{mis} \in C\}$. ■

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 $\mathbf{x}_j$ given $\mathbf{x}_{-j}$ are different for f and g_j. Their predictive distributions have to be different for $\mathbf{x}_j$. Therefore, such an iterative imputation using incompatible conditional models typically does not correspond to Bayesian imputation under any joint model.

4 Incompatible conditional models

In this section, we proceed to the discussion of incompatible conditional models. We first extend the concept of model compatibility to semi-compatibility which includes the regression models we have generally seen in practical uses of iterative imputation. We then introduce the validity of semi-compatible models. Finally, we show that if the conditional models are semi-compatible and valid (together with a few mild technical conditions) the combined imputation estimator is consistent.

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 show 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 issue is to find a natural Bayesian model associated with a set of incompatible conditional models. Naturally, we introduce the concept of semi-compatibility.

Definition 2 *A set of conditional models* $\{h_j(x_j|x_{-j},\theta_j,\varphi_j) : j = 1, ..., p\}$*, 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*

$$g_j(x_j|x_{-j},\theta_j) = h_j(x_j|x_{-j},\theta_j,\varphi_j = 0), \tag{19}$$

for $j = 1, ..., p$*. 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 (unknown) 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 [13], 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. 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 $f(x|\theta)$ 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 must assume 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 the 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:

B1 Both the Gibbs and iterative chains admit their unique invariant distributions, $\nu_1^{\mathbf{X}^{obs}}$ and $\nu_2^{\mathbf{X}^{obs}}$.

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

B3 All the score functions evaluated at the maximum likelihood estimate have finite moment generating functions under $f(\mathbf{x}^{mis}|\mathbf{x}^{obs}, \theta)$.

B4 For each variable j, there exists a subset of observations ι_j so that for each $i \in \iota_j$ have $x_{i,j}$ is missing and $x_{i,-j}$ is fully observed. In addition, the cardinality $\#(\iota_j) \to \infty$ as $n \to \infty$.

Remark 8 *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 it is not strong because it only requires that the cardinality of ι_j tend to infinity, not necessarily even of order of $O(n)$.*

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

$$\hat{\theta} = \sup_{\theta} f(\mathbf{x}^{obs}|\theta), \qquad \hat{\theta}_j = t_j(\hat{\theta}), \tag{20}$$

be the observed-data MLE and

$$\hat{\theta}^{(2)} = \arg\sup_{\theta} \int f(\mathbf{x}|\theta)\nu_2^{\mathbf{X}^{obs}}(d\mathbf{x}^{mis}), \quad (\hat{\theta}_j^{(2)}, \hat{\varphi}_j^{(2)}) = \arg\sup_{\theta_j,\varphi_j} \int h_j(\mathbf{x}_j|\mathbf{x}_{-j}, \theta_j, \varphi_j)\nu_2^{\mathbf{X}^{obs}}(d\mathbf{x}^{mis}) \tag{21}$$

where $\mathbf{x} = (\mathbf{x}^{obs}, \mathbf{x}^{mis})$.

Consider a Markov chain $x^*(k)$ corresponding to one observation—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

$$h_j(x_j|x_{-j}, \theta_j, \varphi_j),$$

where $(\theta_j, \varphi_j) = (\hat{\theta}_j, 0) + \varepsilon\xi_j$ and ξ_j is a random vector with finite MGF (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 process. Given that $\mathbf{x}^{mis,2}(k)$ admits a unique invariant distribution, $x^*(k)$ admits its stationary distribution π for ε sufficiently small. Furthermore, consider that $x(k)$ is a Gibbs sampler and it admits stationary distribution $f(x|\hat{\theta})$, that is, each component is updated according to the conditional distribution $f(x_j|x_{-j}, \hat{\theta})$ and the parameters of the updating distribution are set at the observed data maximum likelihood estimate, $\hat{\theta}$. The last condition is stated as follows.

B5 $x^*(k)$ 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

$$\partial \log f(x|\hat{\theta}) \leq \kappa V(x), \qquad \partial \log h_j(x_j|x_{-j}, \hat{\theta}_j, \varphi_j = 0) \leq \kappa V(x).$$

Remark 9 *By choosing ε small, the transition kernels of $x^*(k)$ and $x(k)$ are close. 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 $\nu_1^{\mathbf{X}^{obs}}$ and $\nu_2^{\mathbf{X}^{obs}}$, 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 notation in* (21)*, the following limits hold:*

$$\hat{\theta}^{(2)} \to \theta^0, \qquad \hat{\theta}_j^{(2)} \to t_j(\theta^0), \quad \hat{\varphi}_j^{(2)} \to 0, \tag{22}$$

in probability as sample size $n \to \infty$ for all j.

Remark 10 *The estimator $\hat{\theta}^{(2)}$ 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 of the conditional model. Therefore, Theorem 3 suggests that the combined imputation estimators are consistent under conditions B1–5.*

Remark 11 *An open problem here is how to consistently estimate the variance of the combined imputation estimator. Given that the imputation distribution of incompatible models is asymptotically different from that of any joint Bayesian imputation, there is no guarantee that Rubin's combined variance estimator is asymptotically consistent. We acknowledge that this is a challenging problem. Even for joint Bayesian imputation, estimating the variance of the combined estimator is still a nontrivial task under specific situations; see, for instance, [13, 9]. Therefore, we leave this issue to future studies.*

5 Linear example

5.1 A simple set of compatible conditional models

In this subsection, we study a linear model as an illustration. Consider n i.i.d. bivariate observations $(\mathbf{x}, \mathbf{y}) = \{(x_i, y_i) : i = 1, ..., n\}$ and a set of conditional models

$$x_i|y_i \sim N(\beta_{x|y} y_i, \tau_x^2), \quad y_i|x_i \sim N(\beta_{y|x} x_i, \tau_y^2). \tag{23}$$

To simplify the discussion, we set the intercepts to zero. As discussed previously, the joint compatible model assumes that (x, y) is a bivariate normal random variable with mean zero, variances σ_x^2 and σ_y^2, and correlation ρ. The reparameterization from the joint model to the conditional model of y given x is

$$\beta_{y|x} = \frac{\sigma_y}{\sigma_x}\rho, \quad \tau_y^2 = (1 - \rho^2)\sigma_y^2.$$

Figure 1 displays the missingness pattern we are assuming for this simple example, with a denoting the set of observations for which both x and y are observed, b denote those with missing y's, and c denoting those with missing x's; n_a, n_b, and n_c denote their respective sample sizes, and $n = n_a + n_b + n_c$. To keep the example simple, we assume that there are no cases for which both x and y are missing.

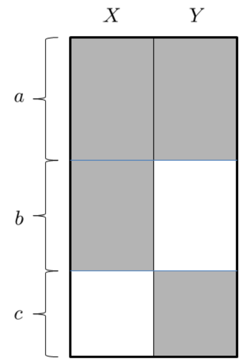

Figure 1: Missingness pattern for our simple example with two variables. Gray and white areas indicate observed and missing data, respectively. This example is constructed so that there are no cases for which both variables are missing.

Positive recurrence and limiting distributions. The Gibbs chain and the iterative chain admit a common small set. The construction of the drift functions is tedious and is not particularly relevant to the current discussion, and so we leave their detailed derivations to the supplemental materials. We proceed here by assuming that they are in force and then condition A2 in Theorem 1 has been satisfied.

Total variation distance between the kernels. The results for incompatible models apply here. Thus, condition A1 in Theorem 1 has been satisfied. We now check condition A3 in Theorem 1. The posterior distribution of the full Bayes model is

$$\begin{aligned} p(\sigma_x^2, \tau_y^2, \beta_{y|x}|\mathbf{x}, \mathbf{y}) &\propto f(\mathbf{x}, \mathbf{y}|\sigma_x^2, \tau_y^2, \beta_{y|x})\pi^*(\sigma_x^2, \tau_y^2, \beta_{y|x}) \\ &= f(\mathbf{y}|\tau_y^2, \beta_{y|x}, \mathbf{x})f(\mathbf{x}|\sigma_x^2)\pi^*(\sigma_x^2, \tau_y^2, \beta_{y|x}). \end{aligned}$$

The posterior distribution of $(\tau_y^2, \beta_{y|x})$ with σ_x^2 integrated out is

$$p(\tau_y^2, \beta_{y|x}|\mathbf{x}, \mathbf{y}) \propto f(\mathbf{y}|\tau_y^2, \beta_{y|x}, \mathbf{x})\pi_{\mathbf{X}}(\beta_{y|x}, \tau_y^2),$$

where

$$\pi_{\mathbf{X}}(\beta_{y|x}, \tau_y^2) \propto \int f(\mathbf{x}|\sigma_x^2)\pi^*(\sigma_x^2, \tau_y^2, \beta_{y|x})d\sigma_x^2.$$

The next task is to show that $\pi_{\mathbf{X}}(\beta_{y|x}, \tau_y^2)$ is a diffuse prior satisfying the conditions in Proposition 1. We impose the following independent prior distributions on σ_x^2, σ_y^2, and ρ:

$$\pi(\sigma_x^2, \sigma_y^2, \rho) \propto \sigma_x \sigma_y I_{[-1,1]}(\rho). \tag{24}$$

The distribution of $\mathbf{x}$ does not depend on (σ_y^2, ρ). Therefore, under the posterior distribution given $\mathbf{x}$, σ_x^2 and (σ_y^2, ρ) are independent. Conditional on $\mathbf{x}$, σ_x^2 is inverse-gamma. Now we proceed to develop the conditional/posterior distribution of $(\tau_y^2, \beta_{y|x})$ given $\mathbf{x}$. Consider the following change of variables

$$\sigma_y^2 = \tau_y^2 + \beta_{y|x}^2 \sigma_x^2, \quad \rho = \beta_{y|x} \sqrt{\frac{\sigma_x^2}{\tau_y^2 + \beta_{y|x}^2 \sigma_x^2}}.$$

Then,

$$\det \left(\frac{\partial(\sigma_y^2, \rho, \sigma_x^2)}{\partial(\tau_y^2, \beta_{y|x}, \sigma_x^2)} \right) = \frac{\sigma_x}{\sqrt{\tau_y^2 + \beta_{y|x}^2 \sigma_x^2}}.$$

Together with

$$\pi(\sigma_y^2, \rho^2) \propto \sigma_y,$$

we have

$$\begin{aligned} \pi_{\mathbf{x}}(\tau_y^2, \beta_{y|x}) &\propto \int \det \left(\frac{\partial(\sigma_y^2, \rho, \sigma_x^2)}{\partial(\tau_y^2, \beta_{y|x}, \sigma_x^2)} \right) \pi(\sigma_y^2, \rho) p(\sigma_x^2|\mathbf{x}) d\sigma_x^2 \\ &= \int \sigma_x p(\sigma_x^2|\mathbf{x}) d\sigma_x^2 = C(\mathbf{x}). \end{aligned}$$

Remark 12 *If one chooses $\pi_2(\tau_y^2, \beta_{y|x}) \propto 1$ for the iterative imputation and (24) for the joint Bayesian model, the iterative chain and the Gibbs chain happen to have identical transition kernels and, therefore, identical invariant distributions. This is one of the rare occasions that these two procedures yield identical imputation distributions.*

If one chooses Jeffreys' prior, $\pi_2(\tau_y^2, \beta_{y|x}) \propto \tau_y^{-2}$, then

$$L(\tau_y^2, \beta_{y|x}) = \frac{\pi_{\mathbf{x}}(\tau_y^2, \beta_{y|x})}{\pi_2(\tau_y^2, \beta_{y|x})} \propto \tau_y^2,$$

and condition A3 is satisfied.

Empirical convergence check. To numerically confirm the convergence of the two distributions, we generate the following data sets. To simplify analysis, let (x_i, y_i)'s be bivariate Gaussian random vectors with mean zero, variance one, and correlation zero. We set $n_a = 200$, $n_b = 80$, and $n_c = 80$. For the iterative imputation we use Jeffreys' prior $p(\tau_y^2, \beta_{y|x}) \propto \tau_y^{-2}$ and $p(\tau_x^2, \beta_{x|y}) \propto \tau_x^{-2}$. For the full Bayesian model, the prior distribution is chosen as in (24).

We monitor the posterior distributions of the following statistics:

$$\beta_x = \frac{\sum_{i \in b} x_i y_i}{\sum_{i \in b} y_i^2}, \quad \beta_y = \frac{\sum_{i \in c} x_i y_i}{\sum_{i \in c} x_i^2}. \tag{25}$$

Figures 2 shows the quantile-quantile plots of the distributions of β_x and β_y under $\nu_1^{\mathbf{X}^{obs}}$ and $\nu_2^{\mathbf{X}^{obs}}$

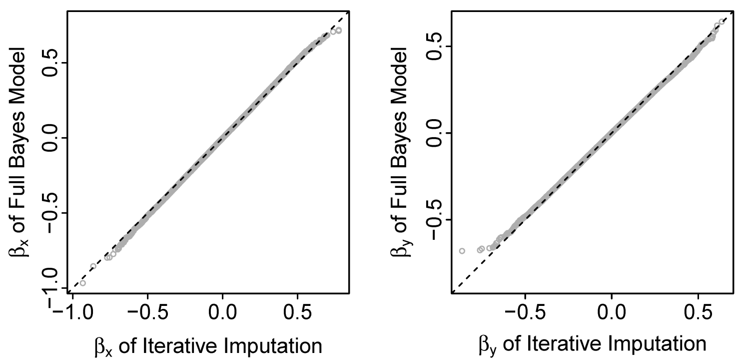

Figure 2: Quantile-quantile plots demonstrating the closeness of the posterior distribution of the Bayesian model and the compatible iterative imputation distributions for β_x and β_y with sample size $n_a = 200$.

based on 1 million MCMC iterations. The differences between these two distributions are tiny.

5.2 Higher-dimensional linear models

We next consider a more complicated and realistic situation, in which there are p continuous variables, $x_1, ..., x_p$. Each conditional model is linear in the sense that, for each j,

$$x_j | x_{-j} \sim N((1, x_{-j}^\top)\beta_j, \sigma_j^2),$$

which is the set of compatible models presented in Example 2.

In the simulation, we generate 1000 samples of $(x_1, ..., x_7)$ from a 7-dimensional multivariate normal distribution with mean 0 and covariance matrix that equals 1 on the diagonals and 0.4 on the off-diagonal elements. We then generate another variable $y \sim N(-2 + x_1 + x_2 + x_3 + x_4 - x_5 - x_6 - x_7, 1)$. Hence the dataset contains $y, x_1, x_2, \ldots, x_7$. For each variable, we randomly select 30% of the observations and set them to be missing. Thus, the missing pattern of the dataset is missing completely at random (MCAR). We impute the missing values in two ways: iterative imputation a multivariate Gaussian joint Bayesian model. After imputation, we use the imputed dataset and regress y on all x's to obtain the regression coefficients. The Q-Q plot of Figure 3 compares the imputation distribution of the least-square estimates of the regression coefficients of the iterative imputation procedure and the multivariate Gaussian joint model.

5.3 Simulation study for incompatible models

We next consider conditionals that are incompatible and valid. To study the frequency properties of the iterative imputation algorithm, we generate 1000 datasets independently each with a sample size of 2000. For each dataset, $y_1 \sim$ Bernouli(0.45), $y_2 \sim$ Bernouli(0.65), y_1 and y_2 are independent,

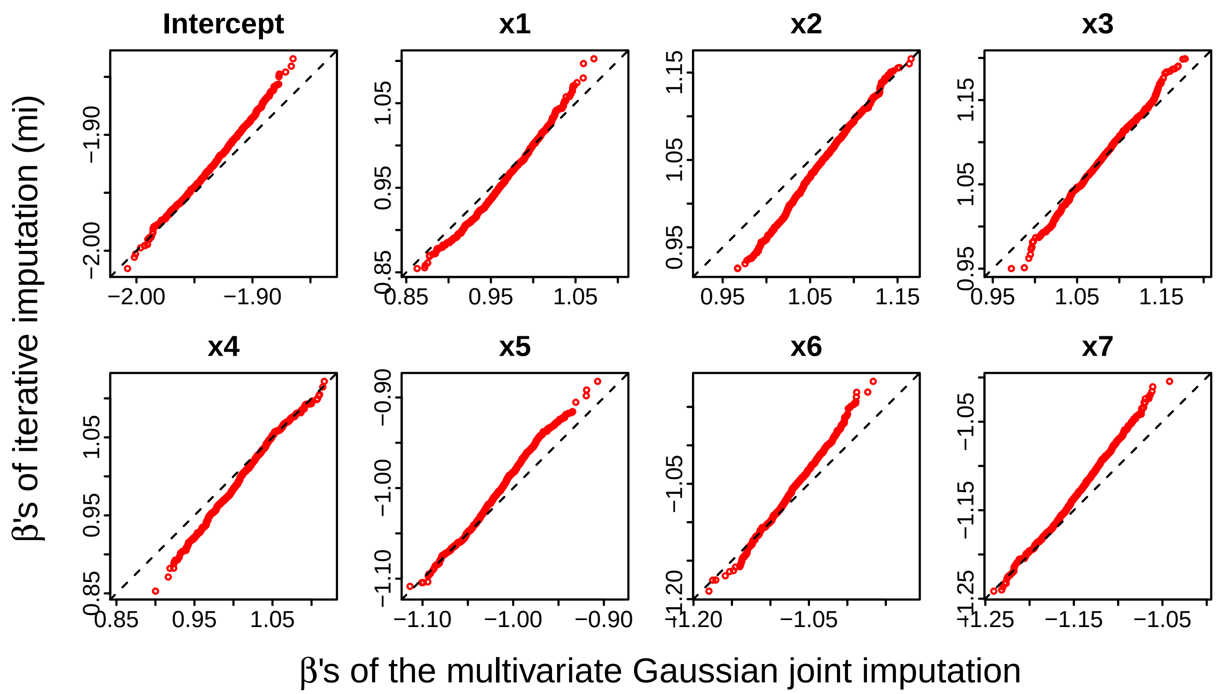

Figure 3: The Q-Q plot of the imputation distributions of the regression coefficients (y on x's) from the joint Bayesian imputation and the iterative imputation.

and the reamining variables come from this conditional distribution: $x_1, \ldots, x_5 | y_1, y_2 \sim N(\mu_1 y_1 + \mu_2 y_2, \Sigma)$, where μ_1 is a vector of 1's and μ_2 is a vector of 0.5's and Σ is a 5×5 matrix that is 1 on the diagonals and 0.2 on the off-diagonal elements.

As before, we remove 30% of the data completely at random and then impute the dataset using iterative imputation. We impute y_1 and y_2 using logistic regressions and $x_1, \ldots, x_5$ using linear regressions. In particular, y_1 is conditionally imputed given $y_2, x_1, x_2, x_3, x_4, x_5$, and the interactions $x_1 y_2$ and $x_2 y_2$; y_2 is conditionally imputed given $y_1, x_1, x_2, x_3, x_4, x_5$, and the interactions $x_1 y_1$ and $x_2 y_1$; and each x_j, $j = 1, \ldots, 5$, is conditionally imputed given y_1, y_2, and the other four x_j's. The conditional models for the x_j's are simple linear models, whereas the logistic regressions for y_i also include interactions. As a result, the set of conditional models is no longer compatible but is still valid. To check whether or not the incompatible models result in reasonable estimates, we impute the missing values using these conditional models. For each dataset, we obtain combined estimates of the regression coefficients of x_1 given the others, that is, β_{01}, β_{i1}^{y} and the vector β_1 by averaging the least-square estimates over 50 imputed datasets. That is, for each dataset, we have 50 imputations, for each of which we obtain the estimated regression coefficients of $x_1 | y_1, y_2, x_2, x_3, x_4, x_5$. Next, we average over 50 sets of coefficients to obtain a single set of coefficients. We repeat the whole procedure on 1000 datasets to get 1000 sets of estimated coefficients. Figure 4 shows the distribution of the estimated coefficients of x_1 regressing on $y_1, y_2, x_2, x_3, x_4, x_5$ based on 1000 im-

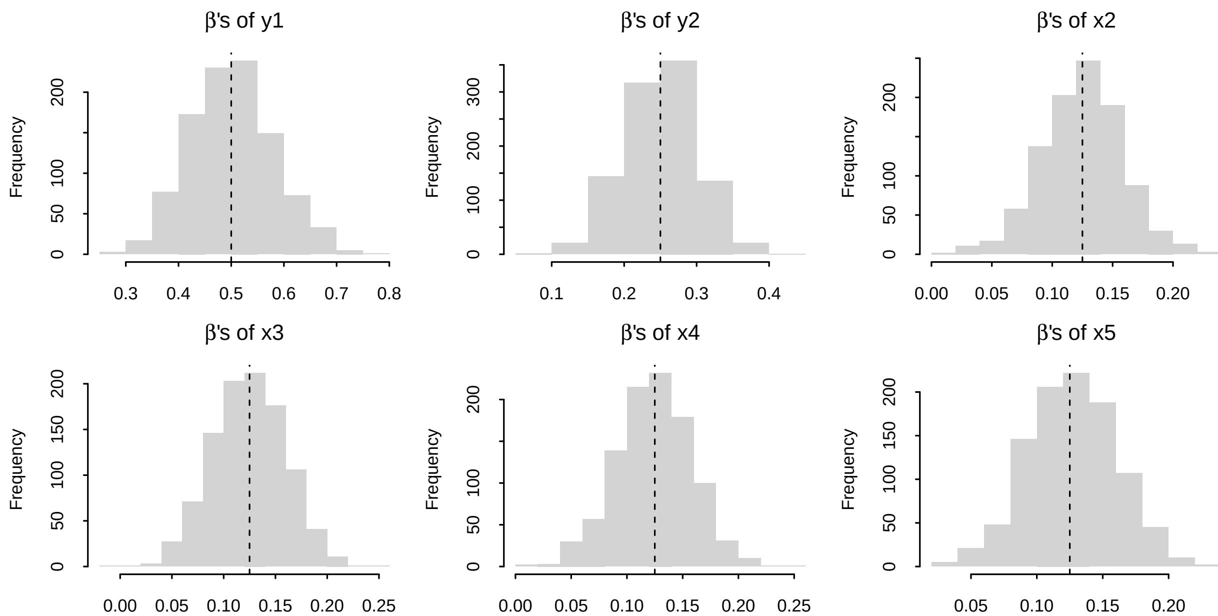

Figure 4: The histograms of coefficients of x_1 regressing on $y_1, y_2, x_2, x_3, x_4, x_5$ from 1000 imputed datasets using an iterative imputation routine [27]. The dashed vertical lines represents the true value of the regression coefficients of the simulation setting, which are $0.5, 0.25, 0.125, 0.125, 0.125, 0.125$.

puted datasets. The frequentist distributions of the combined estimate are centered around their true values, which is consistent with Theorem 3.

6 Discussion

SOMETHING IS NEEDED HERE

A Proofs in Section 3

Lemma 3 *Let Q_0 and Q_1 be probability measures defined on the same σ-field $\mathcal{F}$ and such that $dQ_1 = r^{-1}dQ_0$ for a positive r.v. $r > 0$. Suppose that for some $\varepsilon > 0$, $E^{Q_1}\left(r^2\right) = E^{Q_0}r \leq 1 + \varepsilon$. Then,*

$$\sup_{|f|\leq 1} \left|E^{Q_1}(f(X)) - E^{Q_0}(f(X))\right| \leq \varepsilon^{1/2}.$$

Proof of Lemma 3.

$$\begin{aligned}\left|E^{Q_1}(f(X)) - E^{Q_0}(f(X))\right| &= \left|E^{Q_1}\left[(1-r)f(X)\right]\right| \\ &\leq E^{Q_1}\left(|r-1|\right) \leq [E^{Q_1}(r-1)^2]^{1/2} = \left(E^{Q_1}r^2 - 1\right)^{1/2} \leq \varepsilon^{1/2}.\end{aligned}$$

■

Proof of Proposition 1. According to Lemma 3, we basically need to show that

$$\int r^2(\theta) f_{\mathbf{X}}(\theta) = 1 + \frac{\kappa^2 |\partial L(\mu_\theta)|}{\sqrt{n} L(\mu_\theta)}.$$

Let $\mu_L = E^f L(\theta)$.

$$r(\theta) = \frac{L(\theta)}{\mu_L} = \frac{L(\mu_\theta) + \partial L(\xi)(\theta - \mu_\theta)}{\mu_L}.$$

Then

$$\begin{aligned} E^f(r^2(\theta)) \frac{\mu_L^2}{L^2(\mu_\theta)} &= 1 + 2E \frac{\partial L(\xi)(\theta - \mu_\theta)}{L(\mu_\theta)} + E \frac{(\partial L(\xi))^2 (\theta - \mu_\theta)^2}{L^2(\mu_\theta)} \\ &\leq 1 + 2 \frac{|\partial L(\mu_\theta)| EZ}{L(\mu_\theta)\sqrt{n}} + \frac{(\partial L(\mu_\theta))^2}{L^2(\mu_\theta)} \frac{EZ^2}{n}. \end{aligned}$$

With a similar argument, there exists a constant κ_1 such that

$$\left| \frac{\mu_L^2}{L^2(\mu_\theta)} - 1 \right| \leq \frac{|\partial L(\mu_\theta)|}{L(\mu_\theta)} \frac{\kappa_1^2}{\sqrt{n}}.$$

Therefore, there exists some $\kappa > 0$ such that

$$\begin{aligned} E^f r^2(\theta) &\leq \left(1 + 2E \frac{|\partial L(\mu_\theta)| Z}{L(\mu_\theta)\sqrt{n}} + \frac{(\partial L(\mu_\theta))^2}{L^2(\mu_\theta)} \frac{EZ^2}{n}\right) \frac{L^2(\mu_\theta)}{\mu_L^2} \\ &\leq 1 + \frac{|\partial L(\mu_\theta)|}{L(\mu_\theta)} \frac{\kappa}{\sqrt{n}}. \end{aligned}$$

Using Lemma 3, we conclude the proof. ■

Proof of Lemma 2. For any $\varepsilon, \delta > 0$, let $k_\varepsilon = \inf\{j : \forall k > j, r_k \leq \varepsilon\}$. Then, for any $m > k_\varepsilon$

$$\begin{aligned} \|\tilde{\nu}_1^{\mathbf{X}^{obs}} - \tilde{\nu}_2^{\mathbf{X}^{obs}}\|_V &\leq \left\|\tilde{\nu}_1^{\mathbf{X}^{obs}} - \frac{1}{m}\sum_{k=1}^m \tilde{K}_1^{(k)}(\nu, \cdot)\right\|_V + \left\|\tilde{\nu}_2^{\mathbf{X}^{obs}} - \frac{1}{m}\sum_{k=1}^m \tilde{K}_2^{(k)}(\nu, \cdot)\right\|_V \\ &\quad + \left\|\frac{1}{m}\sum_{k=1}^m \tilde{K}_1^{(k)}(\nu, \cdot) - \frac{1}{m}\sum_{k=1}^m \tilde{K}_2^{(k)}(\nu, \cdot)\right\|_V. \end{aligned}$$

By the definition of k_ε, each of the first two terms is bounded by $\varepsilon + k_\varepsilon C/m$, where $C = \max\{\int V(w)\tilde{K}_i^{(k)}(\nu, dw) : i = 1, 2, \quad 1 \leq k \leq k_\varepsilon\}$. For the last term, for each $k \leq m$ and $|f| \leq V$,

$$\begin{aligned} &\left|\int f(w)[\tilde{K}_1^{(k+1)}(\nu, dw) - \tilde{K}_2^{(k+1)}(\nu, dw)]\right| \\ \leq\ &\left|\int \left(\tilde{K}_1^{(k)}(\nu, dw) - \tilde{K}_2^{(k)}(\nu, dw)\right) \int f(w')\tilde{K}_2(w, dw')\right| + \int \tilde{K}_1^{(k)}(\nu, dw)\|\tilde{K}_1(w, \cdot) - K_2(w, \cdot)\|_V \\ \leq\ &\left\|\tilde{K}_1^{(k)}(\nu, \cdot) - \tilde{K}_2^{(k)}(\nu, \cdot)\right\|_V + d(A_n) \\ =\ &\left\|\tilde{K}_1^{(k)}(\nu, \cdot) - \tilde{K}_2^{(k)}(\nu, \cdot)\right\|_V + o(1), \end{aligned}$$

as $n \to \infty$. Then, by induction, for all $k \le m$,

$$\left\| \tilde{K}_1^{(k)}(\nu, \cdot) - \tilde{K}_2^{(k)}(\nu, \cdot) \right\|_V \le o(1)$$

Therefore, the last term is

$$\left\| \frac{1}{m} \sum_{k=1}^{m} \tilde{K}_1^{(k)}(\nu, \cdot) - \frac{1}{m} \sum_{k=1}^{m} \tilde{K}_2^{(k)}(\nu, \cdot) \right\|_V = o(1).$$

Therefore,

$$\left\| \tilde{\nu}_1^{\mathbf{X}^{obs}}, \tilde{\nu}_2^{\mathbf{X}^{obs}} \right\|_V \le 2\varepsilon + 2Ck_\varepsilon/m + o(1).$$

We can choose m sufficiently large and conclude the proof. ■

Proof of Proposition 2. The proof uses a similar idea as that of Proposition 1: $\tilde{K}_1$ is equivalent to updating the missing values from the posterior predictive distribution of f condition on that $\mathbf{x} \in A_n$. Similarly, $\tilde{K}_2$ corresponds to the posterior predictive distributions of g_j's. By Proposition 1, for all $\mathbf{x} \in A_n$, $|K_1(\mathbf{x}, B) - K_2(\mathbf{x}, B)| = O(n^{-1/4})$, which implies that

$$\frac{K_1(w, A_n)}{K_2(w, A_n)} = 1 + O(n^{-1/4}).$$

The posterior distribution is a joint distribution of the parameter and the missing values. Therefore, θ_j is part of the vector w. Let

$$R = \frac{\tilde{K}_2(w, dw')}{\tilde{K}_1(w, dw')} = \prod_{j=1}^{p} r_j(\theta_j) \frac{K_1(w, A_n)}{K_2(w, A_n)},$$

where $r_j(\theta_j)$ is the normalized prior ratio corresponding to the imputation model of the j-th variable, whose definition is given in Proposition 1.

For the verification of the Lyapunov function,

$$\begin{aligned} \int \tilde{K}_2(w, dw')V(w') &= \tilde{E}_w \left[R \times V(w') \right] \\ &= (1 + O(n^{-1/4}))^p \tilde{E}_w \left(V(w') \prod_{j=1}^{p} \left(1 + 2E \frac{\partial L(\mu_{\theta_j}) Z_j}{L(\mu_{\theta_j})\sqrt{n}} \right)^2 \right). \end{aligned}$$

According to the condition in (16), the above display equals

$$(1 + o(1))\tilde{E}_w V(w') = (1 + o(1)) \int \tilde{K}_1(w, dw')V(w').$$

$V \ge 1$ and it satisfies inequality (15). Then, one can always find another $\lambda_2 \in (0, 1)$ such that the above display is bounded by $\lambda_2 V(w)$. Thus, $V(w)$ is also a drift function for $\tilde{K}_2$. ■

B Proof of Theorem 3

Let $\mathbf{x}(k)$ be the iterative chain starting from its stationary distribution $\nu_2^{\mathbf{X}^{obs}}$. Furthermore, let $\nu_{2,j}^{\mathbf{X}^{obs}}$ be the distribution of $\mathbf{x}(k)$ when the j-th variable is just updated. Due to incompatibility, $\nu_{2,j}^{\mathbf{X}^{obs}}$'s are not necessarily identical. Thanks to stationarity, $\nu_{2,j}^{\mathbf{X}^{obs}}$ does not depend on k and $\nu_2^{\mathbf{X}^{obs}} = \nu_{2,0}^{\mathbf{X}^{obs}} = \nu_{2,p}^{\mathbf{X}^{obs}}$. Let

$$(\tilde{\theta}_j, \tilde{\varphi}_j) = \arg\sup \int \log h_j(\mathbf{x}_j|\mathbf{x}_{-j}, \theta_j, \varphi_j)\nu_{2,j-1}^{\mathbf{X}^{obs}}(d\mathbf{x}^{mis}).$$

The proof consists of two steps. Step 1, we show that for all j, $\tilde{\varphi}_j \to 0, \tilde{\theta}_j \to \hat{\theta}_j$, as $n \to \infty$, where $\hat{\theta}_j$ is the observed-data maximum likelihood estimate based on the joint model f (defined as in (20)). That is, each variable is updated approximately from the conditional distribution $f(x_j|x_{-j}, \hat{\theta})$. Step 2, we establish the statement of the theorem.

Step 1. We prove this step by contradiction. Suppose that there exist ε_0 and j_0 such that $|\tilde{\varphi}_{j_0}| > \varepsilon_0$ or $|\tilde{\theta}_{j_0} - \hat{\theta}_{j_0}| > \varepsilon_0$. Let $\mathbf{x}^*(k)$ be another iterative chain with starting distribution of $\mathbf{x}^{mis}$ being

$$\pi_0(\mathbf{x}^{mis}) = f(\mathbf{x}^{mis}|\mathbf{x}^{obs}, \hat{\theta}),$$

where $\hat{\theta}$ is the observed-data maximum likelihood estimate defined as in (20). Let π_k denote the distribution of $\mathbf{x}^*(k)$ and $\pi_{k,j}$ be the distribution of $\mathbf{x}^*$ at iteration $k+1$ and the j-th variable is just updated. Define the cross-entropy

$$D(P\|Q) = E^P \left[\log \frac{dP}{dQ}\right].$$

By the second law of thermodynamics, since $\mathbf{x}(k)$ and $\mathbf{x}^*(k)$ admit the same transition kernel, we have that

$$D(\pi_{0,j}||\nu_{2,j}^{\mathbf{X}^{obs}})$$

is monotone non-increasing in j. In addition, the following lemma suggests that, after one step transition, $\mathbf{x}^*$ does not deviate much from its starting distribution.

Lemma 4 *Let $\mathbf{x}^*(k)$ be the iterative chain with starting distribution $\pi_0(\mathbf{x}^{*mis}) = f(\mathbf{x}^{*mis}|\mathbf{x}^{*obs}, \hat{\theta})$. Then,*

$$D(\pi_0||\nu_2^{\mathbf{X}^{obs}}) = D(\pi_1||\nu_2^{\mathbf{X}^{obs}}) + O(1).$$

We leave the proof of this lemma to the end of this section. Due to monotonicity, we have that

$$D(\pi_0||\nu_2^{\mathbf{X}^{obs}}) = D(\pi_{0,0}||\nu_{2,0}^{\mathbf{X}^{obs}}) \geq D(\pi_{0,1}||\nu_{2,1}^{\mathbf{X}^{obs}}) \geq ... \geq D(\pi_{0,p}||\nu_{2,p}^{\mathbf{X}^{obs}}) = D(\pi_1||\nu_2^{\mathbf{X}^{obs}}) = D(\pi_0||\nu_2^{\mathbf{X}^{obs}}) - O(1).$$

Thus for each j,

$$D(\pi_{0,j}||\nu_{2,j}^{\mathbf{X}^{obs}}) = D(\pi_{0,j+1}||\nu_{2,j+1}^{\mathbf{X}^{obs}}) + O(1). \tag{26}$$

We try to reach contradiction by showing that (26) is not true for $j_0 - 1$, that is

$$D(\pi_{0,j_0-1}||\nu_{2,j_0-1}^{\mathbf{X}^{obs}}) - D(\pi_{0,j_0}||\nu_{2,j_0}^{\mathbf{X}^{obs}}) \to \infty.$$

Let $\mathbf{x}(k, j_0)$ and $\mathbf{x}^*(k, j_0)$ be the state at iteration $k+1$ and the j_0-th variable is just updated. To simplify notation, we let

$$\begin{aligned} u &= \mathbf{x}_{j_0}^{mis}(0, j_0 - 1), v = \mathbf{x}_{-j_0}(0, j_0 - 1) = \mathbf{x}_{-j_0}(0, j_0), w = \mathbf{x}_{j_0}^{mis}(0, j_0), \\ u^* &= \mathbf{x}_{j_0}^{*mis}(0, j_0 - 1), v^* = \mathbf{x}_{-j_0}^{*}(0, j_0 - 1) = \mathbf{x}_{-j_0*}(0, j_0), w^* = \mathbf{x}_{j_0}^{*mis}(0, j_0). \end{aligned}$$

That is, u is the missing value of variable j_0 from the previous step and w is the updated missing value of variable j_0. v stands for the variables that does not change in this update. Let $p(\cdot)$ be a generic notation for density functions (u, v, w) and $p^*(\cdot)$ for (u^*, v^*, w^*). Following the derivation of the second law of thermodynamics, we have that

$$\begin{aligned} D(p^*||p) &= \int \log \frac{p^*(u,v)}{p(u,v)} p^*(u,v) du dv + \int \log \frac{p^*(w|u,v)}{p(w|u,v)} p^*(u,v,w) du dv dw \\ &= \int \log \frac{p^*(v,w)}{p(v,w)} p^*(v,w) dv dw + \int \log \frac{p^*(u|v,w)}{p(u|v,w)} p^*(u,v,w) du dv dw. \end{aligned}$$

By construction,

$$\begin{aligned} \int \log \frac{p^*(u,v)}{p(u,v)} p^*(u,v) du dv &= D(\pi_{0,j_0-1}||\nu_{2,j_0-1}^{\mathbf{X}^{obs}}) \\ \int \log \frac{p^*(v,w)}{p(v,w)} p^*(v,w) dv dw &= D(\pi_{0,j_0}||\nu_{2,j_0}^{\mathbf{X}^{obs}}). \end{aligned}$$

Since $\mathbf{x}$ and $\mathbf{x}^*$ follows the same transition rules, $p^*(w|u,v) = p(w|u,v)$ and

$$\int \log \frac{p^*(w|u,v)}{p(w|u,v)} p^*(u,v,w) du dv dw = 0.$$

Thus

$$D(p^*||p) = D(\pi_{0,j_0-1}||\nu_{2,j_0-1}^{\mathbf{X}^{obs}}) = D(\pi_{0,j_0}||\nu_{2,j_0}^{\mathbf{X}^{obs}}) + \int \log \frac{p^*(u|v,w)}{p(u|v,w)} p^*(u,v,w) du dv dw.$$

The above derivation is in fact the proof of the second law of thermodynamics. Furthermore, (26) indicates that

$$\int \log \frac{p^*(u|v,w)}{p(u|v,w)} p^*(u,v,w) du dv dw = O(1). \tag{27}$$

In what follows, we show that it is not true: u is the missing values of $\mathbf{x}_j$ from the previous step and w is the missing value for the next step. In addition, the update of $\mathbf{x}_{j_0}^{mis}$ does not depend on the previously imputed values. Therefore, u and w are independent conditional on v. Thus, (27) is reduced to

$$\begin{aligned}\int \log \frac{p^*(u|v,w)}{p(u|v,w)} p^*(u,v,w) du dv dw = O(1). &= \int \log \frac{p^*(u|v)}{p(u|v)} p^*(u,v) du dv \\ &= \int \log \frac{d\pi_{0,j_0-1}(\mathbf{x}_{j_0}^{mis}|\mathbf{x}_{-j_0})}{d\nu_{2,j_0-1}^{\mathbf{X}^{obs}}(\mathbf{x}_{j_0}^{mis}|\mathbf{x}_{-j_0})} \pi_{0,j_0-1}(d\mathbf{x}^{mis}).\end{aligned}$$

We further let ι be the set of observations where x_{j_0} is missing and x_{-j_0} are observed. Use $\mathbf{x}_{\iota,j_0}^{mis}$ to denote the missing x_{j_0}'s of the subset ι. Then

$$\int \log \frac{p^*(u|v,w)}{p(u|v,w)} p^*(u,v,w) du dv dw = O(1). \geq \int \log \frac{d\pi_{0,j_0-1}(\mathbf{x}_{\iota,j_0}^{mis}|\mathbf{x}_{-j_0})}{d\nu_{2,j_0-1}^{\mathbf{X}^{obs}}(\mathbf{x}_{\iota,j_0}^{mis}|\mathbf{x}_{-j_0})} \pi_{0,j_0-1}(d\mathbf{x}^{mis}),$$

that is the joint K-L divergence is bounded from below by the marginal of K-L divergence on the subset ι. The starting value is $\mathbf{x}_{\iota,j_0}^{*mis}(0,j_0-1)$; then $\mathbf{x}_{\iota,j_0}^{*mis}(0,j_0-1)$ was sampled from the distribution $\pi_{0,j_0-1}(\mathbf{x}_{\iota,j_0}^{mis}|\mathbf{x}_{-j_0}) = f(\mathbf{x}_{\iota,j_0}|\mathbf{x}_{\iota,-j_0},\hat{\theta})$ where $\mathbf{x}_{\iota,-j_0}$ is fully observed (by the construction of set ι).

On the other hand, $\mathbf{x}_{\iota,j_0}^{mis}(0,j_0-1)$ follows the stationary distribution and therefore can be considered as a sample from the previous step (the minus one step) according to the conditional model $h_{j_0}(x_{j_0}|x_{i,-j_0},\theta_{j_0},\varphi_{j_0})$ with $|\varphi_{j_0}| + |\theta_{j_0} - \hat{\theta}_{j_0}| > \varepsilon_0$. Together with the fact that $\int \partial^2 h_{j_0}(\mathbf{x}_{\iota,j_0}|\mathbf{x}_{\iota,-j_0},\hat{\theta}_{j_0},\varphi_{j_0}=0)\pi_{0,j_0-1}(d\mathbf{x}^{mis}) = O(\#(\iota))$, we have that

$$\begin{aligned}\int \log \frac{p^*(u|v,w)}{p(u|v,w)} p^*(u,v,w) du dv dw = O(1). &\geq \int \log \frac{f(\mathbf{x}_{\iota,j_0}|\mathbf{x}_{\iota,-j_0},\hat{\theta})}{d\nu_{2,j_0}^{\mathbf{X}^{obs}}(\mathbf{x}_{\iota,j_0}^{mis}|\mathbf{x}_{-(j_0+1)})} \pi_{0,j_0-1}(d\mathbf{x}^{mis}) \quad (28) \\ &= \int \log \frac{f(\mathbf{x}_{\iota,j_0}|\mathbf{x}_{\iota,-j_0},\hat{\theta})}{h_{j_0}(\mathbf{x}_{\iota,j_0}|\mathbf{x}_{\iota,-j_0},\theta_{j_0},\varphi_{j_0})} \pi_{0,j_0-1}(d\mathbf{x}^{mis}) \\ &\geq \delta_0 \#(\iota).\end{aligned}$$

for some $\delta_0 > 0$. Since $\#(\iota) \to \infty$, we reached a contradiction to (26). Thus, $|\tilde{\varphi}_j| + |\tilde{\theta}_j(\mathbf{x}^{obs}) - \hat{\theta}_j| = o(1)$ as $n \to \infty$. Thereby, we conclude step 1.

Step 2. We first show the consistency of $\hat{\theta}^{(2)}$. It is sufficient to show that $|\hat{\theta}^{(2)} - \hat{\theta}| \to 0$. $\hat{\theta}^{(2)}$ solves equation

$$\int \partial \log f(\mathbf{x}^{mis}|\mathbf{x}^{obs},\theta)\nu_2^{\mathbf{X}^{obs}}(d\mathbf{x}^{mis}) = 0.$$

By Taylor expansion, it is sufficient to show that

$$\int \partial \log f(\mathbf{x}^{mis}|\mathbf{x}^{obs},\hat{\theta})\nu_2^{\mathbf{X}^{obs}}(d\mathbf{x}^{mis}) = o(n).$$

Consider a single observation $x^{mis}(k)$. Without loss of generality, suppose that $x^{mis}(k) = (x_1(k), ..., x_j(k))$ and $x^{obs} = (x_{j+1}, ..., x_p)$. Then, it is sufficient to show that

$$\int \partial \log f(x^{mis}|x^{obs}, \hat{\theta})\nu_2^{\mathbf{X}^{obs}}(dx^{mis}) = o(1). \tag{29}$$

$$\int \partial \log f(x^{mis}|x^{obs}, \hat{\theta}) f(dx^{mis}|x^{obs}, \hat{\theta}) = 0.$$

The result of Step 1 suggests that each coordinate of x^{mis} is updated from

$$h_j(x_j|x_{-j}, \hat{\theta}_j + o(1), \varphi_j = o(1)).$$

Thus, $x^{mis}(k)$ follows precisely the transition kernel of x^* described in condition B5. Therefore, we apply Lemma 2 and have that

$$\begin{aligned} & \int \partial \log f(x^{mis}|x^{obs}, \hat{\theta})\nu_2^{\mathbf{X}^{obs}}(dx^{mis}) \\ = & \int \partial \log f(x^{mis}|x^{obs}, \hat{\theta}) f(dx^{mis}|x^{obs}, \hat{\theta}) + o(1). \end{aligned}$$

Then, (29) is satisfied immediately. Therefore, $\hat{\theta}^{(2)} - \hat{\theta} \to 0$. The proof for $\hat{\theta}_j^{(2)}$ and $\hat{\varphi}_j^{(2)}$ are completely analogous and therefore is omitted. Thereby, we conclude the proof.

Proof of Lemma 4. By definition

$$\begin{aligned} D(\pi_{0,1}||\nu_2^{\mathbf{X}^{obs}}) &= \int \log \frac{d\pi_{0,1}(\mathbf{x}^{mis})}{d\nu_2^{\mathbf{X}^{obs}}(\mathbf{x}^{mis})}\pi_{0,1}(d\mathbf{x}^{mis}) \\ &= \int \log \frac{d\pi_{0,1}(\mathbf{x}^{mis})}{d\pi_0(\mathbf{x}^{mis})}\pi_{0,1}(d\mathbf{x}^{mis}) + \int \log \frac{d\pi_0(\mathbf{x}^{mis})}{d\nu_2^{\mathbf{X}^{obs}}(\mathbf{x}^{mis})}\pi_{0,1}(d\mathbf{x}^{mis}). \end{aligned} \tag{30}$$

We consider each of the above two terms. We start with the first term. Since only x_1 is updated, we have that

$$\log \frac{d\pi_{0,1}(\mathbf{x}^{mis})}{d\pi_0(\mathbf{x}^{mis})} = \log \frac{d\pi_{0,1}(\mathbf{x}_1^{mis}|\mathbf{x}_{-1}^{mis})}{d\pi_0(\mathbf{x}_1^{mis}|\mathbf{x}_{-1}^{mis})}.$$

The expression $\pi_{0,1}(\mathbf{x}_1^{mis}|\mathbf{x}_{-1}^{mis})$ is the posterior predictive distribution of x_1 given x_{-1}, that is,

$$\pi_{0,1}(\mathbf{x}_1^{mis}|\mathbf{x}_{-1}^{mis}) = \int h_1(\mathbf{x}_1^{mis}|\mathbf{x}_{-1}^{mis}, \theta_1, \varphi_1)p_1(\theta_1, \varphi_1|\mathbf{x}_{-1}, \mathbf{x}_1^{obs})d\theta_1 d\varphi_1. \tag{31}$$

In addition, (θ_1, φ_1) is $O_p(n^{-1/2})$ from the MLE and $\partial f = 0$ at the MLE. Therefore,

$$\begin{aligned} \pi_{0,1}(\mathbf{x}_1^{mis}|\mathbf{x}_{-1}^{mis}) &= \int [f(\mathbf{x}_1^{mis}|\mathbf{x}_{-1}^{mis}, \hat{\theta}_1) + O(1)]p_1(\theta_1, \varphi_1|\mathbf{x}_{-1}, \mathbf{x}_1^{obs})d\theta_1 d\varphi_1 \\ &= O(1). \end{aligned}$$

Therefore, the first term in (30) is $O(1)$. We now consider the second term.

$$\int \log \frac{d\pi_0(\mathbf{x}^{mis})}{d\nu_2^{\mathbf{X}^{obs}}(\mathbf{x}^{mis})}\pi_{0,1}(d\mathbf{x}^{mis}) \int \log \frac{d\pi_0(\mathbf{x}^{mis})}{d\nu_2^{\mathbf{X}^{obs}}(\mathbf{x}^{mis})} \times \frac{d\pi_{0,1}(\mathbf{x}^{mis})}{d\pi_0(\mathbf{x}^{mis})}\pi_0(d\mathbf{x}^{mis}).$$

A similar argument and the representation in (31) show that $\frac{d\pi_{0,1}(\mathbf{x}^{mis})}{d\pi_0(\mathbf{x}^{mis})} = 1 + O(n^{-1})$. And $D(\pi_0||\nu_2^{\mathbf{X}^{obs}}) = O(n)$. Therefore,

$$\begin{aligned}\int \log \frac{d\pi_0(\mathbf{x}^{mis})}{d\nu_2^{\mathbf{X}^{obs}}(\mathbf{x}^{mis})}\pi_{0,1}(d\mathbf{x}^{mis}) &= (1+O(n^{-1}))\int \log \frac{d\pi_0(\mathbf{x}^{mis})}{d\nu_2^{\mathbf{X}^{obs}}(\mathbf{x}^{mis})}\pi_0(d\mathbf{x}^{mis}) \\ &= D(\pi_0||\nu_2^{\mathbf{X}^{obs}}) + O(1).\end{aligned}$$

Thus

$$D(\pi_{0,1}||\nu_2^{\mathbf{X}^{obs}}) = D(\pi_0||\nu_2^{\mathbf{X}^{obs}}) + O(1).$$

We apply the above derivation to each of the p variables and conclude the proof that

$$D(\pi_{0,p}||\nu_2^{\mathbf{X}^{obs}}) = D(\pi_1||\nu_2^{\mathbf{X}^{obs}}) = D(\pi_0||\nu_2^{\mathbf{X}^{obs}}) + O(1).$$

■

C Markov chain stability and rates of convergence

In this section, we discuss the pending topic of the Markov chain's convergence. A bound on the convergence rate q_k is required for both Lemma 2 and 3. In this section, we review strategies in existing literature to check the convergence. We first provide a brief summary of methods to control the rate of convergence via renewal theory.

Markov chain stability by renewal theory. We first list a few conditions (cf. [4]), which we will refer to later.

A1 Minorization condition: A homogeneous Markov process $W(n)$ with state space in $\mathcal{X}$ and transition kernel $K(w, dw') = P(W(n+1) \in dw'|W(n) = w)$ is said to satisfy a *minorization condition* if for a subset $C \subset \mathcal{X}$, there exists a probability measure ν on $\mathcal{X}$, $l \in \mathbb{Z}^+$, and $q \in (0, 1]$ such that

$$K^{(l)}(w, A) \geq q\nu(A)$$

for all $w \in C$ and measurable $A \subset \mathcal{X}$. C is called a *small set.*

A2 Strong aperiodicity condition: There exists $\delta > 0$ such that $q\nu(C) > \delta$.

A3 Geometric drift condition: there exists a non-negative and finite drift function, V and scalar

$\lambda \in (0,1)$ such that for all $w \bar{\in} C$,

$$\lambda V(w) \geq \int V(w')K(w,dw'),$$

and for all $w \in C$, $\int V(w')K(w,dw') \leq b$.

Chains satisfying A1–3 are ergodic and admit a unique stationary distribution

$$\pi(\cdot) = \lim_{n\to\infty} \frac{1}{n} \sum_{l=1}^{n} K^{(l)}(w,\cdot)$$

for all w. Moreover, there exists $\rho < 1$ depending only (and explicitly) on q, δ, λ, and b such that whenever $\rho < \gamma < 1$, there exists $M < \infty$ depending only (and explicitly) on q, δ, λ, and b such that

$$\sup_{|g|\leq V} |\int g(w')K^{(k)}(w,dw') - \int g(w')\pi(dw')| \leq MV(w)\gamma^k, \tag{32}$$

for all w and $k \geq 0$, where the supremum is taken over all measurable g satisfying $g(w) \leq V(w)$. See [18] and more recently [4] for a proof via the coupling of two Markov processes.

Therefore, once conditions A1, A2 and A3 are in force, the chain admits a unique probability invariant distribution. In addition, we can construct a bound for the convergence rates in conditions C2 (Lemma 2) and D4 (Theorem 3) according to (32).

Convergence of the iterative chain. For the Markov chain of the iterative imputation, usually conditions A1 and A2 are easy to check. For instance, sufficient conditions for A1 and A2 are that the transition kernel $K(w,\cdot)$ is continuous in w and has positive density on the space. The challenge lies in checking the positive recurrence to a small set. We let $W_1(n)$ denote the Gibbs chain and $W_2(n)$ denote the iterative chain. Assuming that conditions A1 and A2 are in force, we focus our attention on the positive recurrence of $W_2(n)$ to a small set. One sufficient condition is the existence of a drift function to a small set (A3). Though there are some principles to follow, construction of drift functions is usually done on a case by case basis. Nevertheless, if a drift function of W_1 is available, we can take advantage of the closeness of the transition kernels of W_1 and W_2 and construct a drift function to the same small set for W_2. We provide a proposition for the construction of a drift function of W_2 given that the drift function of W_1 is known. Therefore, we will need to construct a drift function for only one chain.

Proposition 3 *Suppose that $W_i(n)$, $i = 1,2$, are Markov processes. Both chains satisfy conditions A1 and A2. C is a small set for both W_i. W_1 satisfies A3 with drift function $V(w)$ to the small set C such that for all w,*

$$\lambda V(w) + b \geq \int K_1(w,dw')V(w'),$$

with $\lambda \in (0,1)$ *and* $b \in (0,\infty)$. *In addition, there exists* $q \in (0,1)$ *such that*

$$K_1(w,\cdot) = (1-q)T(w,\cdot) + qQ_1(w,\cdot), \quad K_2(w,\cdot) = (1-q)T(w,\cdot) + qQ_2(w,\cdot),$$

with T, Q_1, Q_2 *transition kernels. Furthermore, there exists a constant* κ *such that*

$$\int Q_2(w,dw')V(w') \leq V(w) + \kappa.$$

If $q < 1-\lambda$, *then there exists* $\lambda' \in (0,1)$ *and* b' *large enough such that*

$$\lambda' V(w) + b' \geq \int K_2(w,dw')V(w').$$

Proof.

$$\begin{aligned}\int K_2(w,dw')V(w') &= \int K_1(w,dw')V(w') - q\int Q_1(w,dw')V(w') + q\int Q_2(w,dw')V(w') \\ &\leq \lambda V(w) + q(V(w)+\kappa) + b \\ &\leq (\lambda+q)V(w) + q\kappa + b.\end{aligned}$$

Therefore, we choose $b' = b + q\kappa$ *and* $\lambda' = q + \lambda < 1$. *The conclusion holds.* ■

A practical alternative. In practice, one can check for convergence empirically. There are many diagnostic tools for the convergence of MCMC; see [7] and the associated discussion. Such empirical studies can show stability within the range of observed simulations. This can be important in that we would like our imputations to be coherent even if we cannot assure they are correct. In addition, most theoretical bounds are conservative in the sense that the chain usually converges much faster than what it is implied by the bounds. On the other hand, purely empirically checking supplies no theoretical guarantee that the chain converges to any distribution. Therefore, a theoretical development of the convergence is recommended when it is feasible given available resources (for instance, time constraint).

References

[1] Y. Amit. On rates of convergence of stochastic relaxation for gaussian and non-gaussian distributions. *Journal of Multivariate Analysis*, 38(1):82–99, 1991.

[2] Y. Amit and U. Grenander. Comparing sweep strategies for stochastic relaxation. *Journal of Multivariate Analysis*, 37(2):197–222, 1991.

[3] J. Barnard and D. B. Rubin. Small-sample degrees of freedom with multiple imputation. *Biometrika*, 86(4):948–955, 1999.

[4] P. H. Baxendale. Renewal theory and computable convergence rates for geometrically ergodic markov chains. *Annals of Applied Probability*, 15(1B):700–738, 2005.

[5] D. R. Cox and D. V. Hinkley. *Theoretical statistics*. Chapman and Hall, London,, 1974.

[6] B. Efron. Efficiency of logistic regression compared to normal discriminant-analysis. *Journal of the American Statistical Association*, 70(352):892–898, 1975.

[7] A. Gelman and D.B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472, 1992.

[8] S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *Ieee Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741, 1984.

[9] J. K. Kim. Finite sample properties of multiple imputation estimators. *Annals of Statistics*, 32(2):766–783, 2004.

[10] K.H. Li, X.L. Meng, T.E. Raghunathan, and D.B. Rubin. Significance levels from repeated \$p\$-values with multiply-imputed data. *Statistica Sinica*, 1:65–92, 1991.

[11] Roderick J. A. Little and Donald B. Rubin. *Statistical analysis with missing data*. Wiley series in probability and statistics. Wiley, Hoboken, N.J., 2nd edition, 2002.

[12] P. McCullagh and J.A. Nelder. *Generalized linear models*. Monographs on statistics and applied probability 37. Chapman & Hall/CRC, Boca Raton, 2nd edition, 1998.

[13] X.L. Meng. Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, 9:538–558, 1994.

[14] X.L. Meng and D.B. Rubin. Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, 80(2):267–278, 1993.

[15] S. P. Meyn and R. L. Tweedie. *Markov chains and stochastic stability*. Communications and control engineering series. Springer-Verlag, London ; New York, 1993.

[16] T.E. Raghunathan, P.W. Solenberger, and J. Van Hoewyk. *IVEware: imputation and variance estimation software*. Survey Research Center, Institute for Social Research University of Michigan, 2010.

[17] J. M. Robins and N. S. Wang. Inference for imputation estimators. *Biometrika*, 87(1):113–124, 2000.

[18] J. S. Rosenthal. Minorization conditions and convergence-rates for markov-chain monte-carlo. *Journal of the American Statistical Association*, 90(430):558–566, 1995.

[19] P. Royston. Multiple imputation of missing values. *Stata Journal*, 4:227–241, 2004.

[20] P. Royston. Multiple imputation of missing values. *Stata Journal*, 5:1–14, 2005.

[21] D.B. Rubin. Formalizing subjective notions about the effect of nonrespondents in sample surveys. *Journal of the American Statistical Association*, (72):538–543, 1977.

[22] D.B. Rubin. *Multiple Imputation for Nonresponse in Surveys.* John Wiley, NY, 1987.

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

[24] J.L. Schafer. *Analysis of Incomplete Multivariate Data.* Monographs on statistics and applied probability. Chapman & Hall/CRC, 1997.

[25] N. Schenker and A. H. Welsh. Asymptotic results for multiple imputation. *Annals of Statistics*, 16(4):1550–1566, 1988.

[26] Y.S. Su, A. Gelman, J. Hill, and M. Yajima. Multiple imputation with diagnostics (mi) in R: Opening windows into the black box. *Journal of Statistical Software*, Forthcoming.

[27] Yu-Sung Su, Andrew Gelman, Jennifer Hill, and Masanao Yajima. Multiple imputation with diagnostics (mi) in R: Opening windows into the black box. *Journal of Statistical Software*, 45(2):1–31, 2011.

[28] S. van Buuren and K. Groothuis-Oudshoorn. Mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, Forthcoming.

[29] N. Wang and J. M. Robins. Large-sample theory for parametric multiple imputation procedures. *Biometrika*, 85(4):935–948, 1998.

Supplemental materials: technical development of positive recurrence and drift function for linear model

In this section, we construct a small set and drift function for the iterative Markov chain for the linear model in Section 5. Note that it suffices to consider the following sufficient statistics for block c,

$$\sum_{i\in c} x_i y_i, \quad \sum_{i\in c} x_i^2.$$

That is, we need to identify a small set in R^2 and a drift function to that small set for the two statistics in the above display. Given that we only consider one-step transition of the Markov process, we use "$\sim$" to denote the updated values of the next iteration and (x_i, y_i)'s to denote the observed value or the imputed values from the previous iteration. Also, we adopt the following notation,

$$s^2_{\alpha,x} = \sum_{i\in\alpha} x_i^2, \quad s^2_{\alpha,y} = \sum_{i\in\alpha} y_i^2, \quad \tilde{s}^2_{\alpha,x} = \sum_{i\in\alpha} \tilde{x}_i^2, \quad \tilde{s}^2_{\alpha,y} = \sum_{i\in\alpha} \tilde{y}_i^2,$$

for $\alpha = a, b, c$. In what follows, we investigate the one step transition of $\sum_c x_i y_i$ and $\sum_c x_i^2$. To simplify the calculation and without loss of generality, we assume

$$\sum_a x_i y_i = 0.$$

Remark 13 *The construction of small set and drift function for the cases when $\sum_a x_i y_i \neq 0$ is completely analogous and more tedious. Also, we can perform a linear transformation on x or y and make the crossproduct equal to zero.*

Throughout this section, we adopt the following notations. Let n denote the sample size. We write a $a_n = O(b_n)$ if there exists $C > 0$ such that $a_n \leq C b_n$; $a_n = o(b_n)$ if $\lim a_n/b_n = 0$. We write $x_n = O_2(a_n)$ if there exists a random variable $x > 0$ such that $|x_n|$ is stochastically dominated by $a_n x$ with $Ex^2 < \infty$ and $x_n = o_2(1)$ if $Ex_n^2 \to 0$.

The general strategy of constructing a small set and a drift function is to first identify an equilibrium point and let the small set C be a compact domain around the equilibrium point. For instance, $\sum_{i\in c} x_i y_i \approx 0$ and $\sum_{i\in c} x_i^2 \approx (s^2_{a,x} + s^2_{b,x})\frac{n_c}{n_a+n_b}$. Whence a small set has been identified, we are ready to construct the drift function. The basic idea is that if the current state of the Markov chain is far away from C, the chain will in expectation move closer to C. Therefore, we need to first compute approximations of

$$g(x_i; i\in c) \triangleq E(\sum_c \tilde{x}_i y_i | x_i; i \in c), \quad \text{and } f(x_i; i\in c) \triangleq E(\sum_c \tilde{x}_i^2 | x_i; i \in c).$$

The second step is to show that both g and f are contraction mappings with one unique fixed point. The small set C is then chosen to be a domain around this fixed point. In addition, we show that the noise compared with the drift is ignorable as long as the chain is far away enough from C. In Sections C.1 and C.2, we study the one-step transition of $\sum_{i\in c} x_i y_i$ and $\sum_{i\in c} x_i^2$. In Section C.3, we give the specific form of a drift function and small set C based on the results in Section C.1 and C.2. The calculations are the same for the Gibbs chain and the iterative chain. Therefore, we do not particularly differentiate them.

C.1 One-step transition of the cross-product

The iterative imputation evolves as such that we first impute the missing y in b and then impute the missing x in c. Therefore

$$\sum_b x_i \tilde{y}_i = \sum_b x_i(\beta_{y|x} x_i + \varepsilon_i) = \beta_{y|x} s_{b,x}^2 + \sum_b x_i \varepsilon_i,$$

where $\beta_{y|x}$ is a random variable following the posterior distribution given the observations in groups a and c and is asymptotically a normal random variable

$$N\left(\frac{\sum_a x_i y_i + \sum_c x_i y_i + O(1)}{s_{a,x}^2 + s_{c,x}^2 + O(1)}, \quad \frac{\tau_y^2 + O(1)}{s_{a,x}^2 + s_{c,x}^2 + O(1)}\right).$$

The term with $O(1)$ is the impact of the prior distribution and τ_y^2 is a random variable following the corresponding posterior distribution. In addition, ε_i's are i.i.d. $N(0, \tau_y^2)$. Therefore,

$$\begin{aligned}
\sum_b x_i \tilde{y}_i &= \frac{\sum_a x_i y_i + \sum_c x_i y_i + O(1)}{s_{a,x}^2 + s_{c,x}^2 + O(1)} s_{b,x}^2 + Z s_{b,x} \sqrt{1 + \frac{s_{b,x}^2}{s_{a,x}^2 + s_{c,x}^2 + O(1)}}, \\
&= \frac{\sum_a x_i y_i + \sum_c x_i y_i}{s_{a,x}^2 + s_{c,x}^2 + O(1)} s_{b,x}^2 + Z s_{b,x} \sqrt{1 + \frac{s_{b,x}^2}{s_{a,x}^2 + s_{c,x}^2 + O(1)}} \\
&\quad + \frac{O(1) s_{b,x}^2}{s_{a,x}^2 + s_{c,x}^2 + O(1)},
\end{aligned} \tag{33}$$

where $EZ = 0$ and $Z = O_2(\tau_y)$.

Similarly, conditional on the imputed y values in block b, the imputed x values in block c (for the next iteration) satisfies,

$$\begin{aligned}
\sum_c \tilde{x}_i y_i &= \sum_b y_i(\beta_{x|y} y_i + \varepsilon_i) \\
&= \frac{\sum_a x_i y_i + \sum_b x_i \tilde{y}_i}{s_{a,y}^2 + \tilde{s}_{b,y}^2 + O(1)} s_{c,y}^2 + Z' s_{c,y} \sqrt{1 + \frac{s_{c,y}^2}{s_{a,y}^2 + \tilde{s}_{b,y}^2 + O(1)}} + \frac{O(1) s_{c,y}^2}{s_{a,y}^2 + \tilde{s}_{b,y}^2 + O(1)}.
\end{aligned}$$

Plugging in (33) and $\sum_a x_i y_i = 0$ into the above display,

$$\begin{aligned}\sum_c \tilde{x}_i y_i &= \sum_c x_i y_i \frac{s_{b,x}^2}{s_{a,x}^2 + s_{c,x}^2 + O(1)} \frac{s_{c,y}^2}{s_{a,y}^2 + \tilde{s}_{b,y}^2 + O(1)} \\ &\quad + Z' s_{c,y}\sqrt{1 + \frac{s_{c,y}^2}{s_{a,y}^2 + \tilde{s}_{b,y}^2 + O(1)}} + Z s_{b,x}\sqrt{1 + \frac{s_{b,x}^2}{s_{a,x}^2 + s_{c,x}^2 + O(1)}} \frac{s_{c,y}^2}{s_{a,y}^2 + \tilde{s}_{b,y}^2 + O(1)} \\ &\quad + \frac{O(1) s_{c,y}^2}{s_{a,y}^2 + \tilde{s}_{b,y}^2 + O(1)} \frac{s_{b,x}^2}{s_{a,x}^2 + s_{c,x}^2 + O(1)} + \frac{O(1) s_{c,y}^2}{s_{a,y}^2 + \tilde{s}_{b,y}^2 + O(1)}.\end{aligned}$$

where $E(Z) = E(Z') = 0$, $Z = O_2(\tau_y)$ and $Z' = O_2(\tau_x)$. The two terms in the last row of the above display with $O(1)$ are due to the prior. We write them as IP (impact of prior), that is

$$IP = \frac{O(1) s_{c,y}^2}{s_{a,y}^2 + \tilde{s}_{b,y}^2 + O(1)} \frac{s_{b,x}^2}{s_{a,x}^2 + s_{c,x}^2 + O(1)} + \frac{O(1) s_{c,y}^2}{s_{a,y}^2 + \tilde{s}_{b,y}^2 + O(1)}.$$

Then, the above display can be simplified to

$$\sum_c \tilde{x}_i y_i = \sum_c x_i y_i \frac{s_{b,x}^2}{s_{a,x}^2 + s_{c,x}^2 + O(1)} \frac{s_{c,y}^2}{s_{a,y}^2 + \tilde{s}_{b,y}^2 + O(1)} + O_2\left(\sqrt{s_{c,y}^2 + s_{b,x}^2}\right) + IP.$$

We assume that for some $\varepsilon > 0$,

$$s_{b,x}^2 < (1 - 2\varepsilon) s_{a,x}^2, \quad s_{c,y}^2 < (1 - 2\varepsilon) s_{a,y}^2. \tag{34}$$

Remark 14 *The above assumption is strong. It requires the fraction of missing information to be small enough and is usually not necessary. This is just to simplify our analysis.*

Let

$$\gamma = \frac{s_{b,x}^2}{s_{a,x}^2 + s_{c,x}^2 + O(1)} \frac{s_{c,y}^2}{s_{a,y}^2 + \tilde{s}_{b,y}^2 + O(1)} \in (0, 1 - \varepsilon),$$

then

$$\begin{aligned}\sum_c \tilde{x}_i y_i &= \gamma \sum_c x_i y_i + O_p\left(\sqrt{s_{c,y}^2 + s_{b,x}^2}\right) + IP, \\ &= \gamma \sum_c x_i y_i + O_p\left(\sqrt{s_{c,y}^2 + s_{b,x}^2}\right). \end{aligned} \tag{35}$$

The last step is because IP(impact of prior) is of constant order $O(1)$. An intuitive interpretation of the above result is that $\sum_c x_i y_i$ decays exponentially fast to zero with rate γ.

C.2 One-step transition of the sum of squares

Now, we proceed to the one step transition of $s_{c,x}^2 = \sum_c x_i^2$. Let

$$\bar{\sigma}_x^2 = \frac{s_{a,x}^2 + s_{b,x}^2}{n_a + n_b}, \quad \bar{\sigma}_y^2 = \frac{s_{a,y}^2 + s_{c,y}^2}{n_a + n_c}.$$

Let $\rho_{a,c}$ be the sample correlation between x and y based on samples in a and c, and $\tilde{\rho}_{a,b}$ be that based on a and b samples. The sums of squares of the x and y's satisfy the following recursion,

$$\begin{aligned}\tilde{s}^2_{b,y} &= \rho^2_{a,c}\frac{s^2_{a,y}+s^2_{c,y}}{s^2_{a,x}+s^2_{c,x}}s^2_{b,x}+(1-\rho^2_{a,c})\bar{\sigma}^2_y n_b+O_2(\sqrt{n_b}) \\ &= \bar{\sigma}^2_y n_b\left[(1-\rho^2_{a,c})+\rho^2_{a,c}\frac{(n_a+n_c)s^2_{b,x}}{n_b(s^2_{a,x}+s^2_{c,x})}\right]+O_2(\sqrt{n_b}), \end{aligned} \tag{36}$$

Similarly,

$$\tilde{s}^2_{c,x}=\bar{\sigma}^2_x n_c\left[(1-\tilde{\rho}^2_{a,b})+\tilde{\rho}^2_{a,b}\frac{(n_a+n_b)s^2_{c,y}}{n_c(s^2_{a,y}+\tilde{s}^2_{b,y})}\right]+O_2(\sqrt{n_c}). \tag{37}$$

Therefore, by plugging (36) into (37), the evolution of $s^2_{c,x}$ satisfies

$$\frac{\tilde{s}^2_{c,x}}{\bar{\sigma}^2_x n_c}=(1-\tilde{\rho}^2_{a,b})+\tilde{\rho}^2_{a,b}\frac{s^2_{c,y}/n_c}{\frac{s^2_{a,y}}{n_a+n_b}+\frac{n_b\bar{\sigma}^2_y(1-\rho^2_{a,c})}{n_a+n_b}+\frac{\rho^2_{a,c}\bar{\sigma}^2_y s^2_{b,x}/[(n_a+n_b)\bar{\sigma}^2_x]}{\frac{s^2_{a,x}}{\bar{\sigma}^2_x(n_a+n_c)}+\frac{n_c}{n_a+n_c}\frac{s^2_{c,x}}{\bar{\sigma}^2_x n_c}}+O_2(\sqrt{n_b}/(n_a+n_b))}+O_2(1/\sqrt{n_c}).$$

Define function,

$$f(\lambda,\rho,\tilde{\rho})=(1-\tilde{\rho}^2)+\tilde{\rho}^2\frac{s^2_{c,y}/n_c}{\frac{s^2_{a,y}}{n_a+n_b}+\frac{n_b\bar{\sigma}^2_y(1-\rho^2)}{n_a+n_b}+\frac{\rho^2\bar{\sigma}^2_y s^2_{b,x}/[(n_a+n_b)\bar{\sigma}^2_x]}{\frac{s^2_{a,x}}{\bar{\sigma}^2_x(n_a+n_c)}+\frac{n_c}{n_a+n_c}\lambda}}.$$

Then, the evolution of $s_{c,x}$ follows,

$$\frac{\tilde{s}^2_{c,x}}{\bar{\sigma}^2_x n_c}=f\left(\frac{s^2_{c,x}}{\bar{\sigma}^2_x n_c},\rho_{a,c},\tilde{\rho}_{a,b}\right)+O_2(n_c^{-1/2}+n_b^{-1/2}) \tag{38}$$

Let λ^* be the solution to

$$f\left(\lambda^*,\rho_{a,c},\tilde{\rho}_{a,b}\right)=\lambda^*. \tag{39}$$

The expression λ^* depends on $\rho_{a,c}$ and $\tilde{\rho}_{a,b}$. To simplify notation, we omit the indexes of $\rho_{a,c}$ and $\rho_{a,b}$ in the notation of λ^*. In what follows, we provide conditions under which f is a contraction mapping with fixed point λ^*. Consider

$$\frac{\partial f(\lambda,\rho_{a,c},\tilde{\rho}_{a,b})}{\partial\lambda} = \frac{\tilde{\rho}^2_{a,b}s^2_{c,y}/n_c}{\frac{s^2_{a,y}}{n_a+n_b}+\frac{n_b\bar{\sigma}^2_y(1-\rho^2_{a,c})}{n_a+n_b}+\frac{\rho^2_{a,c}\bar{\sigma}^2_y s^2_{b,x}/(n_a+n_b)\bar{\sigma}^2_x}{\frac{s^2_{a,x}}{\bar{\sigma}^2_x(n_a+n_c)}+\frac{n_c}{n_a+n_c}\lambda}}\frac{\frac{\rho^2_{a,c}\bar{\sigma}^2_y s^2_{b,x}/(n_a+n_b)\bar{\sigma}^2_x}{\frac{s^2_{a,x}}{\bar{\sigma}^2_x(n_a+n_c)}+\frac{n_c}{n_a+n_c}\lambda}}{\frac{s^2_{a,y}}{n_a+n_b}+\frac{n_b\bar{\sigma}^2_y(1-\rho^2_{a,c})}{n_a+n_b}+\frac{\rho^2_{a,c}\bar{\sigma}^2_y s^2_{b,x}/(n_a+n_b)\bar{\sigma}^2_x}{\frac{s^2_{a,x}}{\bar{\sigma}^2_x(n_a+n_c)}+\frac{n_c}{n_a+n_c}\lambda}}\frac{\frac{n_c}{n_a+n_c}}{\frac{s^2_{a,x}}{\bar{\sigma}^2_x(n_a+n_c)}+\frac{n_c}{n_a+n_c}\lambda}.$$

The first term on the right hand side of the above display,

$$\frac{\tilde{\rho}_{a,b}^2 s_{c,y}^2/n_c}{\frac{s_{a,y}^2}{n_a+n_b} + \frac{n_b\bar{\sigma}_y^2(1-\rho_{a,c}^2)}{n_a+n_b} + \frac{\rho_{a,c}^2\bar{\sigma}_y^2 s_{b,x}^2/(n_a+n_b)\bar{\sigma}_x^2}{\frac{s_{a,x}^2}{\bar{\sigma}_x^2(n_a+n_c)}+\frac{n_c}{n_a+n_c}\lambda}} \le \frac{s_{c,y}^2(n_a+n_b)}{s_{a,y}^2 n_c};$$

the second term is less than or equals to one; the last term,

$$\frac{\frac{n_c}{n_a+n_c}}{\frac{s_{a,x}^2}{\bar{\sigma}_x^2(n_a+n_c)}+\frac{n_c}{n_a+n_c}\lambda} \le \frac{n_c\bar{\sigma}_x^2}{s_{a,x}^2}.$$

We put all these terms together and obtain,

$$\frac{\partial f(\lambda,\rho_{a,c},\tilde{\rho}_{a,b})}{\partial\lambda} \le \frac{s_{c,y}^2(s_{a,x}^2+s_{b,x}^2)}{s_{a,y}^2 s_{a,x}^2}. \tag{40}$$

Suppose that, for some $\varepsilon > 0$,

$$\frac{s_{c,y}^2(s_{a,x}^2+s_{b,x}^2)}{s_{a,y}^2 s_{a,x}^2} < 1-\varepsilon. \tag{41}$$

Remark 15 *Similar to condition* (34)*, we assume* (41) *to simplify the complexity of the analysis. It is usually not necessary.*

We obtain that $|\partial_\lambda f(\lambda,\rho,\tilde{\rho})| < 1-\varepsilon$ for all $\lambda > 0$. One nice feature of having $|\partial_\lambda f(\lambda,\rho_{a,c},\tilde{\rho}_{a,b})| < 1-\varepsilon$ is that $f: R^+ \to R^+$ is a contraction mapping and for any $\Delta\lambda$,

$$|f(\lambda^*+\Delta\lambda,\rho_{a,c},\tilde{\rho}_{a,b}) - \lambda^*| \le (1-\varepsilon)|\Delta\lambda|,$$

where $\lambda^* = f(\lambda^*,\rho,\tilde{\rho})$, uniqueness and existence of which have been proved in standard functional analysis. Therefore, with condition (41), $\frac{s_{c,x}^2}{\bar{\sigma}_x^2 n_c}$ goes exponentially fast to λ^*.

C.3 The small set and drift function

Based on the fluid dynamics of $s_{c,x}^2$ and $\sum_c x_i y_i$, we are able to provide a drift function and a small set. Let $x_c = \{x_i : i \in c\}$ and λ^* be the solution to

$$f(\lambda^*,\rho_{a,c},E\tilde{\rho}_{a,b}) = \lambda^*.$$

Consider

$$V(x_c) = \frac{\left(\sum_c x_i y_i\right)^2}{s_{a,x}^2} + \frac{\left(\frac{s_{c,x}^2}{\bar{\sigma}_x^2 n_c} - \lambda^*\right)^2}{A^2} n_c.$$

for some A large enough. Let $\tilde{\lambda}^*$ be the solution to

$$f(\tilde{\lambda}^*,\tilde{\rho}_{a,c},E(\tilde{\rho}_{a,b}|\tilde{x}_c)) = \tilde{\lambda}^*,$$

which is the equilibrium point of the next iteration, that is,

$$V(\tilde{x}_c) = \frac{\left(\sum_c \tilde{x}_i y_i\right)^2}{s_{a,x}^2} + \frac{\left(\frac{\tilde{s}_{c,x}^2}{\bar{\sigma}_x^2 n_c} - \tilde{\lambda}^*\right)^2}{A^2} n_c.$$

Now we define a small set

$$C_A = \{x_c : V(x_c) \le A\}.$$

In C_A both $\sum_c x_i y_i / s_{c,x}^2$ and $\frac{s_{c,x}^2}{\bar{\sigma}_x^2 n_c} - \lambda^*$ are $1/\sqrt{n}$ distance away from zero. It is not hard to show that C_A is a small set. Consider the one step transition. Let $\zeta = 1 - \varepsilon$. For all $x_c \bar{\in} C_A$, thanks to (35), (38), (40), and $\gamma < \zeta$,

$$\begin{aligned}
V(\tilde{x}_c) &= \frac{\left(\sum_c \tilde{x}_i y_i\right)^2}{s_{a,x}^2} + \frac{\left(\frac{\tilde{s}_{c,x}^2}{\bar{\sigma}_x^2 n_c} - \tilde{\lambda}^*\right)^2}{A^2} n_c \\
&\le \zeta^2 \frac{\left(\sum_c x_i y_i\right)^2}{s_{a,x}^2} + O_2\left(\sqrt{s_{c,y}^2 + s_{b,x}^2}\right) \frac{\sum_c x_i y_i}{s_{a,x}^2} + O_1\left(\frac{s_{c,y}^2 + s_{b,x}^2}{s_{a,x}^2}\right) \\
&\quad + \frac{n_c}{A^2}\left[\zeta^2 \left(\frac{s_{c,x}^2}{\bar{\sigma}_x^2 n_c} - \lambda^*\right)^2 + \left|\frac{s_{c,x}^2}{\bar{\sigma}_x^2 n_c} - \lambda^*\right| \left|\tilde{\lambda}^* - \lambda^*\right| + \left(\tilde{\lambda}^* - \lambda^*\right)^2\right] \\
&\quad + \frac{1}{A^2} O_2\left(\sqrt{n_c}\right)\left(\left|\frac{s_{c,x}^2}{\bar{\sigma}_x^2 n_c} - \lambda^*\right| + \left|\tilde{\lambda}^* - \lambda^*\right| + O_2\left(\frac{1}{\sqrt{n_c}}\right)\right) \\
&\le \zeta^2 V(x_c) \\
&\quad + O_2\left(\sqrt{s_{c,y}^2 + s_{b,x}^2}\right) \frac{\sum_c x_i y_i}{s_{a,x}^2} + O_1\left(\frac{s_{c,y}^2 + s_{b,x}^2}{s_{a,x}^2}\right) \\
&\quad + \frac{n_c}{A^2}\left[\left|\frac{s_{c,x}^2}{\bar{\sigma}_x^2 n_c} - \lambda^*\right| \left|\tilde{\lambda}^* - \lambda^*\right| + \left(\tilde{\lambda}^* - \lambda^*\right)^2\right] \\
&\quad + \frac{1}{A^2} O_2\left(\sqrt{n_c}\right)\left(\left|\frac{s_{c,x}^2}{\bar{\sigma}_x^2 n_c} - \lambda^*\right| + \left|\tilde{\lambda}^* - \lambda^*\right| + O_2\left(\frac{1}{\sqrt{n_c}}\right)\right).
\end{aligned}$$

In what follows, we show that we can choose A sufficiently large, so that when $V(x_c) > A$

$$\begin{aligned}
&E\Big\{O_2\left(\sqrt{s_{c,y}^2 + s_{b,x}^2}\right) \frac{\sum_c x_i y_i}{s_{a,x}^2} + O_2\left(\frac{s_{c,y}^2 + s_{b,x}^2}{s_{a,x}^2}\right) \\
&+ \frac{n_c}{A^2}\left[\left|\frac{s_{c,x}^2}{\bar{\sigma}_x^2 n_c} - \lambda^*\right| \left|\tilde{\lambda}^* - \lambda^*\right| + \left(\tilde{\lambda}^* - \lambda^*\right)^2\right] \\
&+ \frac{1}{A^2} O_2\left(\sqrt{n_c}\right)\left(\left|\frac{s_{c,x}^2}{\bar{\sigma}_x^2 n_c} - \lambda^*\right| + \left|\tilde{\lambda}^* - \lambda^*\right| + O_2\left(\frac{1}{\sqrt{n_c}}\right)\right)\Big\} \le \varepsilon V(x_c)/2.
\end{aligned} \tag{42}$$

The first terms of the above display are all bounded by

$$\sqrt{V(x_c)}O_1\left(\sqrt{\frac{s_{c,y}^2+s_{b,x}^2}{s_{a,x}^2}}\right);$$

the second term

$$O_2\left(\frac{s_{c,y}^2+s_{b,x}^2}{s_{a,x}^2}\right)=O_2(1).$$

We focus on the second line in (42). Because λ^* is a smooth function of $\rho_{a,c}$, there exists from Taylor's expansion a κ such that

$$\left|\lambda^*-\tilde{\lambda}^*\right|\sqrt{n_c}\le\frac{\kappa\left|\sum_c x_iy_i-\sum_c\tilde{x}_iy_i\right|}{s_{a,x}^2}\sqrt{n_c}\le\frac{2\kappa\left|\sum_c x_iy_i\right|\sqrt{n_c}}{s_{a,x}^2}+O_2(1)\le 2\kappa\sqrt{\frac{n_c}{s_{a,c}^2}V(x_c)}$$

Thus,

$$\frac{n_c}{A^2}\left|\frac{s_{c,x}^2}{\bar{\sigma}_x^2n_c}-\lambda^*\right|\left|\tilde{\lambda}^*-\lambda^*\right|\le\frac{2\kappa}{A}\sqrt{\frac{n_c}{s_{a,c}^2}}V(x_c),\qquad\frac{n_c\left(\tilde{\lambda}^*-\lambda^*\right)^2}{A^2}\le\frac{4\kappa^2n_c}{A^2s_{a,c}^2}V(x_c),$$

and

$$\begin{aligned}&\frac{1}{A^2}O_2\left(\sqrt{n_c}\right)\left(\left|\frac{s_{c,x}^2}{\bar{\sigma}_x^2n_c}-\lambda^*\right|+\left|\tilde{\lambda}^*-\lambda^*\right|+O_2(\frac{1}{\sqrt{n_c}})\right)\\ \le\;&\frac{O_2(1)}{A^2}\left(A\sqrt{V(x_c)}+2\kappa\sqrt{\frac{n_c}{s_{a,c}^2}}\sqrt{V(x_c)}+O_2(1)\right).\end{aligned}$$

Therefore, for A sufficiently large and $V(x_c)>A$, (42) holds and

$$E(V(\tilde{x}_c))\le(1-\varepsilon/2)V(x_c).$$

Therefore, the Markov chain of the iterative imputation under conditions in (34) and (41) is positive recurrent and the expected recurrent time to the small set C_A is bounded by $V(x_c)+bI_{C_A}(x_c)$.