Submitted to the Annals of Statistics

ON THE STATIONARY DISTRIBUTION OF ITERATIVE IMPUTATIONS

By Jingchen Liu* Andrew Gelman† Jennifer Hill‡ and Yu-Sung Su

Columbia, Columbia, NYU, and Tsinghua

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 algorithm for multivariate missing data which proceeds as follows. First, missing values are randomly imputed using some simple stochastic algorithm. Second, the missing values for each variable are updated conditionally on all the others using a model fit to the completed data. The second step is performed on all the variables repeatedly until approximate convergence (as measured, for example, by the mixing of multiple chains). The detailed imputation scheme is given in Section 2.3.

Iterative imputation is an easy way to model uncertainty in 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 needs to specify a sequence of conditional regression models to predict each variable given the others to impute the missing data iteratively from the posterior predictive distributions of the corresponding conditional models. The imputation distribution is then the invariant (stationary) distribution of the corresponding Markov chain. Provided that regression models for univariate response have been well studied in the literature, iterative imputation is much easier to implement, especially for statistical software package development than constructing a joint Bayesian model. We call the imputation from the posterior distribution of a joint Bayesian model *joint Bayesian imputation*.

Due to its convenience and flexibility, the iterative, or chained, imputation is popular and it has been implemented in various statistical software packages, including `mice` [27] and `mi` [26] in `R`, `IVEware` [16] in `SAS`, and `ice` in `STATA` [19, 20].

*Research supported in part by Institute of Education Sciences, through Grant R305D100017.

†Research supported in part by Institute of Education Sciences, through Grant R305D090006.

‡Research supported in part by Institute of Education Sciences, through Grant R305D090006 and R305D100017.

AMS 2000 subject classifications: Primary 62D05, 62E20

Keywords and phrases: Multiple imputation, Markov chain Monte Carlo

Nonetheless, its theoretical properties have not yet been established. In this paper, we aim to fill this void. The key questions, then, 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 non-recurrent (drifting or "blowing up" to infinity), even if each of the conditional models is fitted using a proper prior distribution. There is a wealth of literature on Markov chain stability (standard textbook [15]) that one can employ for this analysis. A brief review is given in Section 5.

In this paper, we focus mostly on the second question – the characterization of the stationary distributions of the iterative imputation. Unlike usual Markov chain Monte Carlo (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 very wide so that it is difficult to provide a solution applicable to all situations. Second, there is a lack of mathematical tools to study such Markov processes. The main contribution of this paper is to develop a mathematical framework under which the asymptotic properties of iterative imputation can be discussed via the coupling of two Markov processes. 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 3.1). This discussion is in Section 3.
2. Model compatibility is usually a 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 MLE's (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 4.2 in Section 4.).

The analysis presented in this paper connects to two literatures. The first one is the literature of missing data and multiple imputation. Standard textbooks are [22, 11], and some key papers are [14, 13, 10, 3, 21, 23, 24]. Large sample properties are studied by [25, 28, 17], small samples are by [3], and the issue of congeniality between the imputer's and analyst's models is considered by [13].

The asymptotic results for both the compatible and incompatible models require bounds on the convergence rates of the Markov chains. There is a vast literature

on Markov chain 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 were also suggested by many authors, for instance, [7]. In the example of this paper (cf. the illustrative example in Section 6), we adopt the framework of renewal theory to prove stability and construct bound for convergence rate ([15, 18, 4]). The advantage of this framework is that it does not assume 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 notations and assumptions. Next, we briefly review the framework of multiple imputation, the iterative imputation procedure, and the Gibbs sampler. In Section 3, we investigate compatible conditional models. In Section 4, the discussion focuses on incompatible models. In Section 5, we review the literature for Markov chain convergence via renewal theory. Section 6 considers one linear example in detail.

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, 1 for the observed and 0 for the missing. Further, we use $\mathbf{x}_i^{obs}$ and $\mathbf{x}_i^{mis}$ to denote the observed and missing data of variable i and let

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

To facilitate our description of the procedures, we let

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

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 the discussion, 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 of 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. In literature, there are other imputation procedures based on parametric models; for instance, Robins and Wang ([17, 28]) propose an imputation procedure using estimates based on estimating equations and the corresponding combining rules. Also, there are non-parametric procedures: hot deck, last observation carried forward, and so forth.

2.2. *Bayesian modeling, imputation, and Gibbs sampling.* For Bayesian imputation, multiply imputed data sets are i.i.d. samples from the posterior distribution. In particular, we adopt a parametric family and a prior distribution

$$\mathbf{x}|\theta \sim f(\mathbf{x}|\theta), \quad \theta \sim \pi(\theta),$$

for $\theta \in \Theta$. The imputed values are i.i.d. samples from the posterior predictive distribution

$$f(\mathbf{x}^{mis}|\mathbf{x}^{obs}) = \int_{\Theta} f(\mathbf{x}^{mis}|\mathbf{x}^{obs},\theta)p(\theta|\mathbf{x}^{obs})d\theta, \tag{2.1}$$

where $p(\theta|\mathbf{x})$ is the posterior distribution associated with f and π. Direct simulation from (2.1) is generally difficult. One standard solution is to use MCMC to draw approximate samples. A popular scheme is the Gibbs sampler. In the scenario of missing data, one 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), the Markov process is ergodic with limiting distribution $p(\mathbf{x}^{mis}, \theta|\mathbf{x}^{obs})$ ([8]).

In order to connect these results to the iterative imputation that will be discussed momentarily, we consider a slightly different Gibbs scheme which consists of p steps as follows,

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

Run steps 1 to p iteratively. 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, the full Bayesian modeling approach requires a joint model for $\mathbf{x}$. There are only a handful of models of that sort and even fewer of them that accommodate both continuous and discrete variables. Moreover, these models typically impose strict distributional assumptions. For instance, the general location model ([24]) assumes that the continuous variables are multivariate Gaussian.

On the other hand, iterative imputation has conceptual problems. There may not 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 iterative imputation distribution by first classifying the conditional models into two categories: compatible and incompatible models. For compatible models, there exists a joint model f which is consistent with each conditional model. For incompatible models, there does not exist such an f.

We refer to the Markov chain generated by the scheme in Section 2.2 as the *Gibbs chain* and the one generated by the scheme in Section 2.3 as the *iterative chain*. The central analysis lies in coupling the Gibbs chain and the iterative chain.

3. Compatible conditional models.

3.1. *Model compatibility.* Analysis of the properties of iterative imputation is particularly challenging. This is partly because of the large collection of possible choices of conditional models. Therefore, we first focus on a smaller class, *compatible conditional models*, defined as follows:

DEFINITION 3.1. *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. 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 6.

EXAMPLE 3.2 (Illustrating example). *Consider a binary continuous variable* (x, y) *and conditional models*

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

These two conditional models are compatible. The corresponding joint model is

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

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 3.3 (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 3.4 (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 3.5 (an incompatible example). *There are many incompatible conditional models. For instance,*

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

These are compatible only if $\beta_2 = 0$.

3.2. *The main theorem for compatible conditional models.* 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. They admit transition kernels

$$K_i(w, dw') = P(\mathbf{x}^{mis,i}(k+1) \in dw' | \mathbf{x}^{mis,i}(k) = w), \tag{3.1}$$

for $i = 1, 2$. 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

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

In order to accommodate different variations of the imputation scheme, we write the theorem in a generic way. Then, we discuss how one can adapt the theorem to the analysis of iterative imputation.

THEOREM 3.6. *Let* $\mathbf{x}^{mis,i}(k)$ *admit transition kernels* K_i *for* $i = 1, 2$. *The transition kernels can be data dependent. We use* n *to denote sample size. Suppose the following conditions hold:*

C1 The transition kernel K_i *admits a unique invariant distribution, denoted by* $\nu_i^{\mathbf{x}^{obs}}$.

C2 There exists a monotone decreasing sequence $q_k \to 0$ *(independent of* $\mathbf{x}^{obs}$*) and a starting measure* ν *(depending on* $\mathbf{x}^{obs}$*) such that*

$$P\left[d_{TV}(K_i^{(k)}(\nu, \cdot), \nu_i^{\mathbf{x}^{obs}}(\cdot)) \le q_k, \forall k > 0\right] \to 1, \tag{3.2}$$

as $n \to \infty$.

C3 There exists a sequence of sets A_n *(depending on sample size), such that for each* $m \in \mathbb{Z}^+$,

$$P_\nu\left(\mathbf{x}^{mis,i}(k) \in A_n \text{: for all } k = 1, ..., m\right) \to 1, \tag{3.3}$$

in probability as $n \to \infty$[1]. *In addition,*

$$d(A_n) \triangleq \sup_{w \in A_n} d_{TV}(K_1(w, \cdot), K_2(w, \cdot)) \to 0, \tag{3.4}$$

in probability as $n \to \infty$.

[1]The probability measure P_ν depends on $\mathbf{x}^{obs}$. That is why the convergence is in probability.

Then,

$$d_{TV}(\nu_1^{\mathbf{X}^{obs}}, \nu_2^{\mathbf{X}^{obs}}) \to 0, \tag{3.5}$$

in probability as $n \to \infty$.

REMARK 3.7. *By construction of the Gibbs chain,* $\nu_1^{\mathbf{X}^{obs}}$ *is the posterior distribution of the corresponding Bayesian model. Iterative imputations satisfying the conditions in Theorem 3.6 are asymptotically equivalent to Bayesian imputation. Then the asymptotic statistical properties developed for multiple imputations using joint Bayesian models are applicable to such iterative imputations too.*

REMARK 3.8. *In practice, the specific imputation scheme might be different from the one presented in Section 2.3. We list two such variations. First, at each of the* p *steps within one iteration, one may sample* θ *from the posterior distribution given the entire data set* $\mathbf{x}$ *from the previous iteration, instead of* $\mathbf{x}_j^{obs}, \mathbf{x}_{-j}$. *Second, instead of updating the* p *variables in a fixed order, one may randomly select one variable to update. Given that Theorem 3.6 is presented in a general form, the results are applicable to these schemes as well. We will then discuss how to check conditions in specific contexts.*

REMARK 3.9. *C1 and C2 can be obtained simultaneously using existing results in the literature of MCMC convergence. Given that this is an independent topic, we discuss tools for the existence of an invariant distribution, its uniqueness, and the rate of convergence in Section 5.*

REMARK 3.10. *The statement of the theorem does not explicitly require compatibility. However, as we will show in later this section, compatibility is usually a necessary condition for C3 (Theorem 3.19). In addition, C3 is generally difficult to check directly. We will provide a set of sufficient and checkable conditions in Section 3.3. Compatibility is one of them.*

REMARK 3.11. *In order to have the two stationary distributions converging to each other, it is necessary to have a bound on the convergence rate (C2) in addition to the condition (C3) that the two transition kernels converge to each other. One illustrative example is given as follows. Consider two order-1 autoregressive processes:*

$$W_i(n+1) = \rho_i W_i(n) + \varepsilon_i(n+1), \quad \text{for } i = 1, 2$$

with $\varepsilon_i(n) \sim N(0,1)$ *and* W_i *having probability invariant distribution* $N(0, (1-\rho_i^2)^{-1})$. *Suppose* $\rho_1^2 = 1-\delta$ *and* $\rho_2^2 = 1-2\delta$. *With* δ *small, the two transition kernels are close to each other. However, the variances of their invariant distributions are approximately different by a factor of* 2. *This is because the mixing rate is low for both chains.*

REMARK 3.12. *For iterative chains, the invariant distribution typically depends on the order of variables within each iteration. For procedures satisfying conditions*

in Theorem 3.6, the impact of this order vanishes asymptotically. This is because the theorem does not require a particular order to have the convergence result.

REMARK 3.13. *The result of Theorem 3.6 does not rely on the validity of the imputation model. Even if the model is misspecified, the convergence in* (3.5) *still applies.*

REMARK 3.14. *The set A_n is introduced to ensure* (3.4)*. It is not always feasible to have $K_1(w,\cdot)$ and $K_2(w,\cdot)$ close in total variation for all w. Very often, it is easy to identify a compact set A_n such that $K_1(w,\cdot)$ and $K_2(w,\cdot)$ are close on A_n and the chains stay in A_n with very high probability in a finite number of steps.*

Before the proof of Theorem 3.6, we first present a lemma.

LEMMA 3.15. *Consider two positive measures ν_1, ν_2 and a non-negative bounded function h. Then,*

$$\left|\int h(x)\nu_1(dx) - \int h(x)\nu_2(dx)\right| \leq d_{TV}(\nu_1,\nu_2)\sup_x h(x).$$

PROOF. The proof is immediate by considering measure $\phi = \nu_1 - \nu_2$ and $\int h(x)\nu_1(dx) - \int h(x)\nu_2(dx) = \int h(x)\phi(dx)$. □

PROOF OF THEOREM 3.6. For any $\varepsilon, \delta > 0$, let $k_\varepsilon = \inf\{K : \forall k > K, q_k \leq \varepsilon\}$. Then, for any $m > k_\varepsilon$

$$\begin{aligned} d_{TV}(\nu_1^{\mathbf{X}^{obs}}, \nu_2^{\mathbf{X}^{obs}}) \quad \leq \quad & d_{TV}\left(\nu_1^{\mathbf{X}^{obs}}, \frac{1}{m}\sum_{k=1}^{m} K_1^{(k)}(\nu,\cdot)\right) + d_{TV}\left(\nu_2^{\mathbf{X}^{obs}}, \frac{1}{m}\sum_{k=1}^{m} K_2^{(k)}(\nu,\cdot)\right) \\ & + d_{TV}\left(\frac{1}{m}\sum_{k=1}^{m} K_1^{(k)}(\nu,\cdot), \frac{1}{m}\sum_{k=1}^{m} K_2^{(k)}(\nu,\cdot)\right). \end{aligned}$$

By the definition of k_ε, each of the first two terms is bounded by $\varepsilon + k_\varepsilon/m$. For the last term, using Lemma 3.15, for each $k \leq m$ and A,

$$\begin{aligned} & \left|K_1^{(k+1)}(\nu,A) - K_2^{(k+1)}(\nu,A)\right| \\ \leq \quad & \left|\int \left(K_1^{(k)}(\nu,dx) - K_2^{(k)}(\nu,dx)\right) K_2(x,A)\right| + \int K_1^{(k)}(\nu,dx)\left|K_1(x,A) - K_2(x,A)\right| \\ \leq \quad & d_{TV}(K_1^{(k)}(\nu,\cdot), K_2^{(k)}(\nu,\cdot)) + d(A_n) + 1 - P_\nu\left(\mathbf{x}^{mis,1}(k) \in A_n\text{: for all } k = 1,...,m\right). \end{aligned}$$

Then, by induction, for all $k \leq m$,

$$d_{TV}(K_1^{(k)}(\nu,\cdot), K_2^{(k)}(\nu,\cdot)) \leq m\left[d(A_n) + 1 - P_\nu\left(\mathbf{x}^{mis,1}(k) \in A_n\text{: for all } k = 1,...,m\right)\right]$$

Therefore, the last term is bounded by

$$d_{TV}\left(\frac{1}{m}\sum_{k=1}^{m}K_1^{(k)}(\nu,\cdot),\frac{1}{m}\sum_{k=1}^{m}K_2^{(k)}(\nu,\cdot)\right)\leq m\left[d(A_n)+1-P_\nu\left(\mathbf{x}^{mis,1}(k)\in A_n\text{: for all } k=1,...,m\right)\right].$$

Therefore,

$$d_{TV}(\nu_1^{\mathbf{x}^{obs}},\nu_2^{\mathbf{x}^{obs}})\leq 2\varepsilon+2k_\varepsilon/m+m\left[d(A_n)+1-P_\nu\left(\mathbf{x}^{mis,1}(k)\in A_n\text{: for all } k=1,...,m\right)\right].$$

We first choose m sufficiently large such that $2k_\varepsilon/m<\varepsilon$. Given the choice of m, thanks to C3,

$$d(A_n)\to 0,$$

in probability as $n\to\infty$, one can choose n sufficiently large such that

$$P\left(d(A_n)>m^{-1}\varepsilon\right)<\delta,$$

$$P\left(1-P_\nu\left(\mathbf{x}^{mis,1}(k)\in A_n\text{: for all } k=1,...,m\right)>m^{-1}\varepsilon\right)<\delta,$$

and the probability in (3.2) is greater than $1-\delta$. Therefore, with this choice of n, we have

$$P\left(d_{TV}(\nu_1^{\mathbf{x}^{obs}},\nu_2^{\mathbf{x}^{obs}})<5\varepsilon\right)>1-3\delta.$$

We conclude the proof. □

3.3. *Total variation distance between the two transition kernels.* Theorem 3.6 is written in a general format. One needs to check conditions C1-3 in the specific context of the Gibbs chain and the iterative chain. As mentioned in Remark 3.9, conditions C1 and C2 will be discussed in Section 5. The rest of this section focuses on the discussion of condition C3. It is difficult to provide checkable, sufficient, and necessary conditions for condition C3 in Theorem 3.6. Instead, we provide a set of sufficient and checkable conditions for C3 and argue that they cover a wide range of practical situations. Note that 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 bound the total variation distance between the following posterior predictive distributions (for each $j=1,...,p$),

$$f(\mathbf{x}_j^{mis}|\mathbf{x}_j^{obs},\mathbf{x}_{-j}) = \int f(\mathbf{x}_j^{mis}|\mathbf{x}_j^{obs},\mathbf{x}_{-j},\theta)p(\theta|\mathbf{x}_j^{obs},\mathbf{x}_{-j})d\theta \tag{3.6}$$

$$g_j(\mathbf{x}_j^{mis}|\mathbf{x}_j^{obs},\mathbf{x}_{-j}) = \int g_j(\mathbf{x}_j^{mis}|\mathbf{x}_j^{obs},\mathbf{x}_{-j},\theta_j)p_j(\theta_j|\mathbf{x}_j^{obs},\mathbf{x}_{-j})d\theta_j, \tag{3.7}$$

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 both the joint Bayesian model and the iterative imputation model:

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

with $t_j(\theta) = \theta_j$. The only difference lies in their posterior distributions. In fact, thanks to Lemma 3.15, the total variation distance between two posterior predictive distributions is bounded by the distance between their posterior distributions of parameters. Therefore, we move on to compare $p(\theta|\mathbf{x}_j^{obs}, \mathbf{x}_{-j})$ 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 lower dimension. Consider the linear model in Example 3.2. 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 3.2 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. Thanks to compatibility, we can drop the notation g_j which we employed to denote the 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 3.2, 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 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

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

Comparing 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 total variation distance of the posterior predictive distributions. We put forward tools to control the total variation distance between the two posterior predictive distributions in (3.6) and (3.7). Let $\mathbf{x}$ be the generic notation for the observed data and $f_{\mathbf{X}}(\theta)$ and $g_{\mathbf{X}}(\theta)$ be two 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:

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

Then, by Lemma 3.15,

$$d_{TV}(\tilde{f}_{\mathbf{X}},\tilde{g}_{\mathbf{X}}) \le d_{TV}(f_{\mathbf{X}},g_{\mathbf{X}}). \tag{3.8}$$

The next proposition provides sufficient conditions that $d_{TV}(f_{\mathbf{X}},g_{\mathbf{X}})$ vanishes.

PROPOSITION 3.16. *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 $L(\theta) = \pi_g(\theta)/\pi_f(\theta)$ and n denote sample size. Let*

$$\mu_\theta = \int \theta f_{\mathbf{X}}(\theta)d\theta.$$

Suppose that on the set $\mathbf{x} \in A_n$ (c.f. Theorem 3.6) there exists $\kappa > 0$ such that

$$|SD^f(\theta)| \le \frac{\kappa}{\sqrt{n}}, \tag{3.9}$$

where $SD^f(\theta)$ is the posterior standard deviation under $f_{\mathbf{X}}$. Let $\partial L(\theta)$ be the partial derivative with respect to θ and ξ be a random variable such that

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

where "$\cdot$" denotes inner product. If there exists a random variable y with finite expectation under $f_{\mathbf{X}}$, such that for $\mathbf{x} \in A_n$

$$\left|\frac{\sqrt{n}\partial L(\xi)\cdot(\theta-\mu_\theta)}{L(\mu_\theta)}\right|^2 \le y, \tag{3.10}$$

then for $\mathbf{x} \in A_n$

$$d_{TV}(f_{\mathbf{x}}, g_{\mathbf{x}}) = O(1)\frac{|\partial L(\mu_\theta) \cdot SD^f(\theta)|}{L(\mu_\theta)}.$$

By (3.8)*, we further obtain*

$$d_{TV}(\tilde{f}_{\mathbf{x}}, \tilde{g}_{\mathbf{x}}) \leq O(1)\frac{|\partial L(\mu_\theta) \cdot SD^f(\theta)|}{L(\mu_\theta)} = o(1)\frac{|L'(\mu_\theta)|}{L(\mu_\theta)}. \tag{3.11}$$

We delay the proof of this proposition to Appendix A.

REMARK 3.17. *We adapt Proposition 3.16 to the analysis of conditional models. For most parametric models,* (3.9) *is satisfied. Under mild moment conditions for the posterior distribution of* θ_j*,* (3.10) *is satisfied. Therefore, we only need to verify that* $d \log L(\theta_j)/d\theta_j$ *is bounded. One sufficient condition is that* $L(\theta_j) = \pi_j(\theta_j)/\pi_{j,\mathbf{x}_{-j}}(\theta_j)$ *grows polynomially in* θ_j*.*

REMARK 3.18. *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.*

Summary. As a result of (3.11) and Proposition 3.16, the total variation distances between the posterior predictive distributions of $\mathbf{x}_j^{mis}$ given $(\mathbf{x}^{obs}, \mathbf{x}_{-j}^{mis})$ associated with the two models (under mild technical conditions) can be controlled by the posterior variance of the parameters and $d \log L(\theta_j)/d\theta_j$ where

$$L(\theta_j) = \frac{\pi_j(\theta_j)}{\pi_{j,\mathbf{x}_{-j}}(\theta_j)}.$$

This forms a set of checkable sufficient conditions for C3 in Theorem 3.6.

3.4. *On the necessity of model compatibility.* Theorem 3.6 and Proposition 3.16 show that for compatible models and under suitable technical conditions, iterative imputation is asymptotically equivalent to Bayesian imputation. The following proposition 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_j^g denote those induced by the iterative imputation's conditional models. That is,

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

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

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

or

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

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

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

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

$$|\nu_2^{\mathbf{x}^{obs}}(\mathbf{x}_{-j}^{mis} \in C) - \nu_1^{\mathbf{x}^{obs}}(\mathbf{x}_{-j}^{mis} \in C)| \leq q\varepsilon/2, \tag{3.12}$$

then, by the fact that

$$\nu_1^{\mathbf{x}^{obs}}(\mathbf{x}^{mis} \in B) = \nu_1^{\mathbf{x}^{obs}}(\mathbf{x}_{-j}^{mis} \in C) \int P^f(\mathbf{x}_j^{mis} \in A | \mathbf{x}_{-j}^{mis}, \mathbf{x}^{obs}) \nu_1^{\mathbf{x}^{obs}}(d\mathbf{x}_{-j}^{mis} | \mathbf{x}_{-j}^{mis} \in C),$$

$$\nu_2^{\mathbf{x}^{obs}}(\mathbf{x}^{mis} \in B) = \nu_2^{\mathbf{x}^{obs}}(\mathbf{x}_{-j}^{mis} \in C) \int P^g(\mathbf{x}_j^{mis} \in A | \mathbf{x}_{-j}^{mis}, \mathbf{x}^{obs}) \nu_2^{\mathbf{x}^{obs}}(d\mathbf{x}_{-j}^{mis} | \mathbf{x}_{-j}^{mis} \in C),$$

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

For two models with different likelihood functions, it is not hard to 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 3.19, 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. The imputation distribution of incompatible conditional models. In this section, we proceed to the discussion of incompatible conditional models. In particular, we first extend the concept of model compatibility to semi-compatibility which includes essentially all the regression models practically in use. Second, we introduce the validity of semi-compatible models. Lastly, 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 always assume that the invariant distribution and a bound on the convergence rate exist. 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 transfer this idea to the analyis 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 4.1. *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* semi-compatible, *if there exists a set of compatible conditional models*

$$g_j(x_j|x_{-j},\theta_j) = h_j(x_j|x_{-j},\theta_j,\varphi_j = 0), \tag{4.1}$$

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. We call a set of semi-compatible models to be *trivial* if it only contains a trivial compatible element and *nontrivial* otherwise. 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 conditional models. 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 such that it includes the true conditional model. For instance, as recommended by [13], the imputer should always try to include as many predictors as possible. 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 4.2. *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. *Notations and technical conditions.* Let S^f denote the score function of the joint model $f(x|\theta)$ and I^f be the observed Fisher information:

$$S^f(\theta; x) = \frac{\partial \log f(x|\theta)}{\partial \theta}, \quad I^f(\theta; x) = -\frac{\partial^2 \log f(x|\theta)}{\partial \theta^2}.$$

We use the second argument (after ";") for the generic notation of data. For instance, $S^f(\theta; x^{obs})$ is the score function of the observed data and $S^f(\theta; x^{mis}|x^{obs})$ is the conditional score function of x^{mis} conditional on x^{obs}. In addition, we use $S^f(\theta; \mathbf{x})$ to denote the score function of the entire data set; this is calculated as the sum of individual scores. Similarly, we use such a notation for the observed Fisher information I^f. Recall that $S^f(\theta^0; x) = O_p(1)$ and $S^f(\theta^0; \mathbf{x}) = O_p(\sqrt{n})$, where θ^0 is the true parameter of f. Furthermore, we let S_j be the score function of h_j and I_j be its Fisher information, that is,

$$\begin{aligned} S_j(\theta_j, \varphi_j; x_j|x_{-j}) &= \frac{\partial \log h_j(x_j|x_{-j}, \theta_j, \varphi_j)}{\partial(\theta_j, \varphi_j)}, \\ I_j(\theta_j, \varphi_j; x_j|x_{-j}) &= -\frac{\partial^2 \log h_j(x_j|x_{-j}, \theta_j, \varphi_j)}{\partial(\theta_j, \varphi_j)^2}. \end{aligned}$$

We use $\theta_j^0 = t_j(\theta^0)$ and $\varphi_j^0 = 0$ to denote the true parameters under h_j.

For each j, we let $F_j(x|x_{-j}, \theta_j, \varphi_j)$ be the conditional cumulative distribution function associated with model $h(x_j|x_{-j}, \theta_j, \varphi_j)$ and $F_j^{-1}(x|x_{-j}, \theta_j, \varphi_j)$ be its generalized inverse function.

We also slightly change the definition of "iteration." In Section 2, one iteration consists of p steps updating each of the p variables. For convenience, we now let $\mathbf{x}^{mis,i}(k+1)$ be the state of the chain after updating just one variable from $\mathbf{x}^{mis,i}(k)$. Therefore, the original chain is a p-skeleton of the chain under the current definition.

Throughout this discussion, we use κ to denote a generic constant for upper bounds whose specific values may vary from case to case. Now, we list a set of technical conditions.

D1 Let $U_j \sim U(0, 1)$, and define

$$x_j = F_j^{-1}(U_j|\tilde{x}_{-j}, \theta_j, \varphi_j) \quad \text{and } x'_j = F_j^{-1}(U_j|\tilde{x}'_{-j}, \theta'_j, \varphi'_j). \tag{4.2}$$

Then, there exists $\kappa > 0$ such that

$$E|x_j - x'_j|^2 \leq \kappa \left(|\tilde{x}'_{-j} - \tilde{x}_{-j}|^2 + |\theta'_j - \theta_j|^2 + |\varphi'_j - \varphi_j|^2 \right),$$

for all (θ_j, φ_j) and (θ'_j, φ'_j) in a neighborhood of $(\theta^0_j, 0)$. The expectation is taken with respect to U_j conditional on $\tilde{x}_{-j}$ and $\tilde{x}'_{-j}$.

D2 Let x_j and x'_j be defined as in (4.2) for $j = 1, ..., p$. Then

$$E|S^f(\theta^0; x_1, ..., x_n) - S^f(\theta^0; x'_1, ..., x'_n)|^2 \leq \kappa \sum_{j=1}^{p} |\tilde{x}'_{-j} - \tilde{x}_{-j}|^2 + |\theta'_j - \theta_j|^2 + |\varphi'_j - \varphi_j|^2$$

and

$$E|S_j(\theta^0_j, \varphi^0_j; x_j|x_{-j}) - S_j(\theta^0_j, \varphi^0_j; x'_1|x'_{-j})|^2 \leq \kappa \sum_{j=1}^{p} |\tilde{x}'_{-j} - \tilde{x}_{-j}|^2 + |\theta'_j - \theta_j|^2 + |\varphi'_j - \varphi_j|^2,$$

where $\varphi^0_j = 0$ and the expectation is taken with respect to $U_1, ..., U_n$.

D3 Let $\hat{\theta}$, $(\hat{\theta}_j, \hat{\varphi}_j)$ be the MLE's based on f and h_j. They satisfy regularity conditions,

$$\begin{aligned} \hat{\theta} - \theta^0 &= I^f(\theta^0; \mathbf{x})^{-1} S^f(\theta^0; \mathbf{x}) + o(n^{-1/2})\xi, \\ (\hat{\theta}_j - \theta^0_j, \hat{\varphi}_j) &= I_j(\theta^0_j, 0; \mathbf{x}_j|\mathbf{x}_{-j})^{-1} S_j(\theta^0_j, 0; \mathbf{x}_j|\mathbf{x}_{-j}) + o(n^{-1/2})\xi_j, \end{aligned}$$

where ξ and ξ_j are random vectors with bounded second moments.

D4 Let $\mathbf{x}^{mis,i}(k)$ admit a unique stationary distribution denoted by $\nu_i^{\mathbf{x}^{obs}}$. Let $\mathbf{x} = (\mathbf{x}^{obs}, \mathbf{x}^{mis})$ and

$$\mu^{(i)}(\mathbf{x}^{obs}) = \int \hat{\theta}(\mathbf{x}) \nu_i^{\mathbf{x}^{obs}}(d\mathbf{x}^{mis}), \tag{4.3}$$

$$\mu^{(i)}_{j,\theta}(\mathbf{x}^{obs}) = \int \hat{\theta}_j(\mathbf{x}) \nu_i^{\mathbf{x}^{obs}}(d\mathbf{x}^{mis}), \tag{4.4}$$

$$\mu^{(i)}_{j,\varphi}(\mathbf{x}^{obs}) = \int \hat{\varphi}_j(\mathbf{x}) \nu_i^{\mathbf{x}^{obs}}(d\mathbf{x}^{mis}). \tag{4.5}$$

There exists $q_k \to 0$ such that if $\mathbf{x}^{mis,2}(0) \sim \nu_1^{\mathbf{x}^{obs}}$, then for any $k > 0$

$$P\left(\left| E[\hat{\theta}(\mathbf{x}^{obs}, \mathbf{x}^{mis,2}(k))|\mathbf{x}^{obs}] - \mu^{(2)}(\mathbf{x}^{obs}) \right| \leq q_k \right) \to 1, \tag{4.6}$$

$$P\left(\left| E[\hat{\theta}_j(\mathbf{x}^{obs}, \mathbf{x}^{mis,2}(k))|\mathbf{x}^{obs}] - \mu^{(2)}_{j,\theta}(\mathbf{x}^{obs}) \right| \leq q_k \right) \to 1, \tag{4.7}$$

$$P\left(\left| E[\hat{\varphi}_j(\mathbf{x}^{obs}, \mathbf{x}^{mis,2}(k))|\mathbf{x}^{obs}] - \mu^{(2)}_{j,\varphi}(\mathbf{x}^{obs}) \right| \leq q_k \right) \to 1, \tag{4.8}$$

as $n \to \infty$, for all $j = 1, ..., p$. The expectation is taken with respect to $\mathbf{x}^{mis,2}(k)$ under the distribution associated with the iterative Markov chain. The probability outside is taken with respect to $\mathbf{x}^{obs}$ under the sampling distribution. In addition, for some $\kappa > 0$

$$E|\mu^{(1)}(\mathbf{x}^{obs}) - \theta^0|^2 \leq \frac{\kappa}{n}, \tag{4.9}$$

where the expectation is taken with respect to $\mathbf{x}^{obs}$ over the sampling distribution.

D5 The posterior distribution of θ given complete data set $\mathbf{x}$ has the representation

$$\theta - \hat{\theta} = Z + o(n^{-1/2})\xi,$$

where Z is a random vector with mean $[I^f(\theta^0;\mathbf{x})]^{-1}S^f(\theta^0;\mathbf{x})$ and covariance matrix $[I^f(\theta^0;\mathbf{x})]^{-1}$ and $E\xi^2 \le \kappa$. Similarly, the posterior of (θ_j, φ_j) is

$$(\theta_j - \hat{\theta}_j, \varphi_j - \hat{\varphi}_j) = Z_j + o(n^{-1/2})\xi_j,$$

where Z_j is a random vector with mean $[I_j(\theta_j^0, 0; \mathbf{x}_j|\mathbf{x}_{-j})]^{-1}S_j(\theta_j^0, 0; \mathbf{x}_j|\mathbf{x}_{-j})$ and covariance matrix $[I_j(\theta_j, 0; \mathbf{x}_j|\mathbf{x}_{-j})]^{-1}$ and $E\xi_j^2 \le \kappa$.

REMARK 4.3. *Conditions D1 and D2 are satisfied by typical parametric families. For instance, $X \sim N(\mu, 1)$ can be represented by $X = \mu + \Phi^{-1}(U)$ and the score function is $S(\mu'; X) = \mu - \mu' + \Phi^{-1}(U)$, where Φ is the c.d.f. of standard normal distribution.*

Condition D4 requires a bound on the convergence rate of the Markov chain. (4.9) *suggests that the posterior mean of the complete data MLE is $O_p(1/\sqrt{n})$ from the true parameter.*

Conditions D3 and D5 require that the maximum likelihood estimate satisfies the normal equation and the posterior distributions of θ and θ_j are centered around the MLE and have $O(n^{-1/2})$ standard deviation.

4.3. *Main theorem of incompatible conditional models.*

THEOREM 4.4. *Consider a set of valid semi-compatible models $\{h_j : j = 1, ..., p\}$. Suppose that conditions D1-5 are in force. We use n to denote sample size. Then, the combined maximum likelihood estimator with infinitely many imputations is a consistent estimator, that is,*

$$\mu^{(2)}(\mathbf{x}^{obs}) \to \theta^0 \tag{4.10}$$

and

$$\mu^{(2)}_{j,\theta}(\mathbf{x}^{obs}) \to t_j(\theta^0), \quad \mu^{(2)}_{j,\varphi}(\mathbf{x}^{obs}) \to 0, \tag{4.11}$$

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

REMARK 4.5. *$\mu^{(2)}(\mathbf{x}^{obs})$ is the expectation of the complete data MLE under the iterative imputation distribution and is also the combined point estimator of θ according to Rubin's combining rule (with infinitely many imputations). Similarly,* (4.11) *contains the combined estimators of the conditional models. Therefore, the result of Theorem 4.4 suggests that the combined imputation estimator is consistent under conditions D1-5.*

REMARK 4.6. *There is one problem that remains open, that is, how one can 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 very 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.*

PROOF OF THEOREM 4.4. We construct $\mathbf{x}^{mis,1}(k)$ and $\mathbf{x}^{mis,2}(k)$ such that they are coupled as follows. Both chains start from $\nu_1^{\mathbf{X}^{obs}}$. Then $\mathbf{x}^{mis,1}(k) \sim \nu_1^{\mathbf{X}^{obs}}$ for all k. Let $\mathbf{x}^{mis,1}(0) = \mathbf{x}^{mis,2}(0)$. Suppose that variable j is updated at step $k+1$. Then, $\theta(k)$ is first sampled from

$$p(\theta|\mathbf{x}^{obs}, \mathbf{x}_{-j}^{mis,1}(k))$$

and $\mathbf{x}_j^{mis,1}(k+1)$ is sampled from $f(\mathbf{x}_j^{mis}|\mathbf{x}_{-j}^{mis,1}(k), \mathbf{x}^{obs}, \theta(k))$ and $\mathbf{x}_{-j}^{mis,1}(k+1) = \mathbf{x}_{-j}^{mis,1}(k)$. Using the representation in (4.2), we write

$$x_{i,j}^{mis,1}(k+1) = F_j^{-1}(U_{i,j}|x_{i,-j}^{mis,1}(k), x_{i,-j}^{obs}, t_j(\theta(k)), 0), \tag{4.12}$$

where $x_{i,j}^{mis,1}(k+1)$ is the j-th missing variable of the i-th observation (given it is missing). Similarly, for $\mathbf{x}^{mis,2}(k)$, at each step $(\theta_j(k), \varphi_j(k))$ is sampled from

$$p_j(\theta_j, \varphi_j|\mathbf{x}^{obs}, \mathbf{x}_{-j}^{mis,1}(k)).$$

Then $\mathbf{x}_j^{mis,2}(k+1)$ is sampled as follows

$$x_{i,j}^{mis,2}(k+1) = F_j^{-1}(U_{i,j}|x_{i,-j}^{mis,2}(k), x_{i,-j}^{obs}, \theta_j(k), \varphi_j(k)). \tag{4.13}$$

We let the $U_{i,j}$ in (4.12) and (4.13) be identical. Therefore, $\mathbf{x}^{mis,1}(k)$ and $\mathbf{x}^{mis,2}(k)$ are coupled through $U_{i,j}$. In what follows, we prove by induction that $\mathbf{x}^{mis,1}(k)$ and $\mathbf{x}^{mis,2}(k)$ are sufficiently close.

We define notation

$$E^*(\cdot) = E(\cdot|\mathbf{x}^{obs})$$

that is the expectation associated with the probability measure induced by $\mathbf{x}^{mis,i}(k)$ for $i = 1, 2$. Let κ^* be a generic (data-dependent) constant whose value may vary from case to case. In addition, κ^* depends on $\mathbf{x}^{obs}$ and its expectation $E\kappa^* < \infty$ does not increase with sample size. To simplify notation, we write κ^* instead of $\kappa^*(\mathbf{x}^{obs})$.

First, by construction, we have $\mathbf{x}^{mis,1}(0) = \mathbf{x}^{mis,2}(0)$. Suppose that for some k and each $1 \le l \le k$ there exist A_l and A_l' (possibly increasing with l, depending on $\mathbf{x}^{obs}$, and their expectations $EA_l, EA_l' < \infty$ do not increase with sample size) such that for all i and j,

$$E^*\left(x_{i,j}^{mis,1}(l) - x_{i,j}^{mis,2}(l)\right)^2 \le \frac{A_l}{n}, \quad E^*\left|t_j(\theta(l-1)) - \theta_j(l-1)\right|^2 + E^*|\varphi_j(l-1)|^2 \le \frac{A_{l-1}'}{n}. \tag{4.14}$$

Now we consider the case of $k+1$. Without loss of generality, we assume that the j-th variable is updated at step $k+1$. According to (4.12), (4.13), and condition D1,

$$\begin{aligned} & E^*\left[\left|x_{i,j}^{mis,1}(k+1)-x_{i,j}^{mis,2}(k+1)\right|^2\middle|\theta_j(k),\varphi_j(k),\theta(k),\mathbf{x}^{mis,1}(k),\mathbf{x}^{mis,2}(k)\right] \\ \leq\ & \kappa\left[\left|\theta_j(k)-t_j(\theta(k))\right|^2+|\varphi_j(k)|^2+\sum_{l=1}^{p}\left|x_{i,l}^{mis,1}(k)-x_{i,l}^{mis,2}(k)\right|^2\right]. \end{aligned}$$

By the induction assumptions, the last term is controlled by

$$E^*\left|x_{i,l}^{mis,1}(k)-x_{i,l}^{mis,2}(k)\right|^2\leq\frac{A_k}{n}.$$

For the first two terms, let $\mu_l(k)$ be the posterior mean of $(\theta_l(k),\varphi_l(k))$ given $\mathbf{x}_{-l}^{mis,2}(k)$ and $\mathbf{x}^{obs}$ (under the conditional model h_l) for each $1\leq l\leq p$. Then,

$$\begin{aligned} & E^*\left(\left|\theta_l(k)-t_l(\theta(k))\right|^2+|\varphi_l(k)|^2\right) \\ \leq\ & E^*\left|t_l(\theta(k))-\theta_l^0\right|^2+E^*\left|(\theta_l(k),\varphi_l(k))-\mu_l(k)\right|^2+E^*\left|\mu_l(k)-(\theta_l^0,0)\right|^2. \end{aligned}$$

To control each of the three terms on the right hand side, we collect the following facts. For each $1\leq l\leq p$, we derive the following bounds.

1. Since $\mathbf{x}^{mis,1}(k)\sim\nu_1^{\mathbf{x}^{obs}}$, D5, and (4.9), there exists $\kappa>0$ such that

$$E^*|t_l(\theta(k))-\theta_l^0|^2\leq\frac{\kappa^*}{n}. \tag{4.15}$$

2. By condition D5, we obtain that

$$E^*|(\theta_l(k),\varphi_l(k))-\mu_l(k)|^2\leq\frac{\kappa^*}{n}. \tag{4.16}$$

3. Lastly, we control $E^*\left|\mu_l(k)-(\theta_l^0,0)\right|^2$. Note that

$$\begin{aligned} \mu_l(k)-(\theta_l^0,0) &= (\hat{\theta}_l(\mathbf{x}_{-j}^{mis,2}(k),\mathbf{x}^{obs})-\theta_l^0,\hat{\varphi}_l(\mathbf{x}_{-j}^{mis,2}(k),\mathbf{x}^{obs}))+o(n^{-1/2})\xi_l \\ &= I_l(\theta_l^0,0;\mathbf{x}_l^{obs}|\mathbf{x}_{-l}^{obs},\mathbf{x}_{-l}^{mis,2}(k))^{-1}S_l(\theta_l^0,0;\mathbf{x}_l^{obs}|\mathbf{x}_{-l}^{obs},\mathbf{x}_{-l}^{mis,2}(k))+o(n^{-1/2})\xi_l. \end{aligned}$$

In addition, for some j^*, suppose $\mathbf{x}_{j^*}^{mis,1}(k)$ is updated according to h_{j^*} given $\mathbf{x}^{mis,1}(k-1)$, $t_{j^*}(\theta(k-1))$ and $\varphi_{j^*}=0$; $\mathbf{x}_{j^*}^{mis,2}(k)$ is updated according to h_{j^*} given $\mathbf{x}^{mis,2}(k-1)$, $\theta_{j^*}(k-1)$ and $\varphi_{j^*}(k-1)$. Therefore, by D2

$$\begin{aligned} & E^*\left|S_l(\theta_l^0,0;x_{i,l}^{obs}|x_{i,-l}^{obs},x_{i,-l}^{mis,1}(k))-S_l(\theta_l^0,0;x_{i,l}^{obs}|x_{i,-l}^{obs},x_{i,-l}^{mis,2}(k))\right|^2 \\ \leq\ & \kappa\left[E^*|\varphi_{j^*}(k-1)|^2+E^*\left|t_{j^*}(\theta(k-1))-\theta_{j^*}(k-1)\right|^2+\sum_{l'=1}^{p}E^*\left(x_{i,l'}^{mis,1}(k-1)-x_{i,l'}^{mis,2}(k-1)\right)^2\right] \\ \leq\ & \kappa\frac{A'_{k-1}+pA_{k-1}}{n}. \end{aligned}$$

Therefore, we add up the scores functions of the individual observations and obtain that

(4.17)
$$E^*\left|S_l(\theta_l^0,0;\mathbf{x}_l^{obs}|\mathbf{x}_{-l}^{obs},\mathbf{x}_{-l}^{mis,1}(k))-S_l(\theta_l^0,0;\mathbf{x}_l^{obs}|\mathbf{x}_{-l}^{obs},\mathbf{x}_{-l}^{mis,2}(k))\right|^2\leq n\kappa\left(A'_{k-1}+pA_{k-1}\right),$$

for some $\kappa>0$. Since $\mathbf{x}^{mis,1}(k)\sim\nu_1^{\mathbf{x}^{obs}}$, (4.9), and D5, we have that (possibly by enlarging κ^*)

$$E^*|S_l(\theta_l^0,0;\mathbf{x}_l^{obs}|\mathbf{x}_{-l}^{obs},\mathbf{x}_{-l}^{mis,1}(k))|^2\leq\kappa^* n. \tag{4.18}$$

Combining (4.17) and (4.18), we obtain

$$E^*|S_l(\theta_l^0,0;\mathbf{x}_l^{obs}|\mathbf{x}_{-l}^{obs},\mathbf{x}_{-l}^{mis,2}(k))|^2\leq 2\kappa n(A'_{k-1}+pA_{k-1})+2n\kappa^*.$$

Let $n\lambda_l$ be the smallest eigenvalue of I_l. Then,

$$\begin{aligned}
&E^*\left|\mu_l(k)-(\theta_l^0,0)\right|^2 \\
=\;&E^*\left|I_l(\theta_l^0,0;\mathbf{x}_l^{obs}|\mathbf{x}_{-l}^{obs},\mathbf{x}_{-l}^{mis,2}(k))^{-1}S_l(\theta_l^0,0;\mathbf{x}_l^{obs}|\mathbf{x}_{-l}^{obs},\mathbf{x}_{-l}^{mis,2}(k))\right|^2 \\
&+o(n^{-1}) \\
\leq\;&\frac{2\kappa(A'_{k-1}+pA_{k-1})+2\kappa^*+o(1)}{n\lambda_l^2}.
\end{aligned} \tag{4.19}$$

According to (4.15), (4.16), and (4.19), we obtain that for each l

$$\begin{aligned}
&E^*\left(|\theta_l(k)-t_l(\theta(k))|^2+|\varphi_l(k)|^2\right) \\
\leq\;&E^*\left|t_l(\theta(k))-\theta_l^0\right|^2+E^*\left|(\theta_l(k),\varphi_l(k))-\mu_l(k)\right|^2+E^*\left|\mu_l(k)-(\theta_l^0,0)\right|^2 \\
\leq\;&\max_l\frac{2\kappa^*(1+\lambda_l^{-2})+2\kappa\lambda_l^{-2}(A'_{k-1}+pA_{k-1})+o(1)}{n} \\
\triangleq\;&\frac{A'_k}{n},
\end{aligned}$$

and,

$$E^*\left[\left|x_{i,j}^{mis,1}(k+1)-x_{i,j}^{mis,2}(k+1)\right|^2\middle|\theta_j,\varphi_j,\theta,\mathbf{x}^{mis,1}(k),\mathbf{x}^{mis,2}(k)\right]\leq\frac{\kappa(A'_k+pA_k)}{n}\triangleq\frac{A_{k+1}}{n}.$$

Therefore we conclude by induction that

(4.20)
$$E^*\left(x_{i,j}^{mis,1}(k)-x_{i,j}^{mis,2}(k)\right)^2\leq\frac{A_k}{n},\quad E^*\left|t_j(\theta(k-1))-\theta_j(k-1)\right|^2+E^*|\varphi(k-1)|^2\leq\frac{A'_{k-1}}{n}.$$

for all i,j,k. In addition, $EA_k, EA'_k<\infty$ do not increase with sample size (but do increase with k).

With the result from induction and exactly the same argument as in (4.17) and (4.18), we can find C_k sufficiently large (depending on $\mathbf{x}^{obs}$ and its expectation $EC_k < \infty$ does not increase with sample size) such that

$$\begin{aligned} E^* \left|S(\theta^0; \mathbf{x}^{mis,1}(k), \mathbf{x}^{obs}) - S(\theta^0; \mathbf{x}^{mis,2}(k), \mathbf{x}^{obs})\right|^2 &\leq C_k n, \\ E^* \left|S(\theta^0; \mathbf{x}^{mis,1}(k), \mathbf{x}^{obs})\right|^2 &\leq \kappa^* n. \end{aligned}$$

Then,

$$E^* \left|S(\theta^0; \mathbf{x}^{mis,2}(k), \mathbf{x}^{obs})\right|^2 \leq 2(C_k + \kappa^*)n.$$

Let λn be the smallest eigenvalue of I^f. Then, by D3

$$E^* \left(\hat{\theta}(\mathbf{x}^{mis,2}(k), \mathbf{x}^{obs}) - \theta^0\right)^2 \leq \frac{2(C_k + \kappa^*)}{\lambda^2 n}. \tag{4.21}$$

In addition, (4.15) and the second inequality in (4.20) suggest

$$E^*(\theta_j(k) - \theta_j^0)^2 + E^*(\varphi_j(k))^2 \leq \frac{2A'_k + 2\kappa^*}{n},$$

and therefore

$$E^* \left(\hat{\theta}_j(\mathbf{x}^{obs}, \mathbf{x}^{mis,2}(k)) - \theta_j^0\right)^2 + E^* \left(\hat{\varphi}_j(\mathbf{x}^{obs}, \mathbf{x}^{mis,2}(k))\right)^2 \leq \frac{2A'_k + 2\kappa^*}{n}. \tag{4.22}$$

Recall that $E^*(\cdot) = E(\cdot|\mathbf{x}^{obs})$. Putting together (4.6), (4.7), and (4.8) in D4, (4.21) and (4.22), we conclude the proof. □

5. 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 Theorems 3.6 and 4.4. 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

$$\sup_{|g|\leq V} \left| \int g(w')K^{(k)}(w, dw') - \int g(w')\pi(dw') \right| \leq MV(w)\gamma^k, \tag{5.1}$$

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 (Theorem 3.6) and D4 (Theorem 4.4) according to (5.1).

On the 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 5.1. *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

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

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

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

PROOF.

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

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).

6. Linear example.

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

$$x_i|y_i \sim N(\beta_{x|y}y_i, \tau_x^2), \quad y_i|x_i \sim N(\beta_{y|x}x_i, \tau_y^2). \tag{6.1}$$

To simplify the discussion, we let the intercept be 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

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

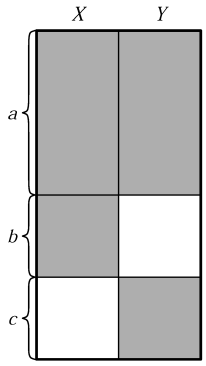

FIG 6.1. *Missingness pattern for our simple example. The gray area indicates observed data. This example thus avoids the potential instability that can arise when multible variables are missing simultaneously.*

As shown in Figure 6.1, let a denote the set of observations for which both x and y are observed, b denote those with missing y's and c denote 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 simplify discussion, we assume that there is no case in which both x and y are missing.

Positive recurrence and the limiting distribution. Both the Gibbs chain and the iterative chain satisfy conditions A1, A2, and A3. The verification of conditions A1-3 is tedious and 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 therefore C1 and C2 in Theorem 3.6 have been satisfied.

Total variation distance between the kernels. We now check condition C3 in Theorem 3.6. 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 3.16. We impose independent prior distributions on σ_x^2, σ_y^2, and ρ

$$\pi(\sigma_x^2, \sigma_y^2, \rho) \propto \sigma_x\sigma_y I_{[-1,1]}(\rho). \tag{6.2}$$

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 6.1. *If one chooses $\pi_2(\tau_y^2, \beta_{y|x}) \propto 1$ for the iterative imputation and (6.2) for the joint Bayesian model, the iterative chain and the Gibbs chain happen to have identical transition kernels and, therefore, identical invariant distributions. Note that 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

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

Let $\mu_{\tau_y^2}$ be the posterior mean of τ_y^2. Then, $SD^f(\tau_y^2) = O(\mu_{\tau_y^2} n^{-1/2})$ and $SD^f(\beta_{y|x}) = O(\mu_{\tau_y^2}^{-1/2} n^{-1/2})$. Therefore, according to Proposition 3.16, condition C3 in Theorem 3.6 is satisfied by choosing A_n be a set such that the sample variances of the x and y variables are bounded by some constant. According to Theorem 3.6, the iterative imputation distribution converges to the posterior distribution of the bivariate Gaussian Bayesian model as the sample size tend to infinity.

Empirical check of the convergence in total variation. To 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 (6.2).

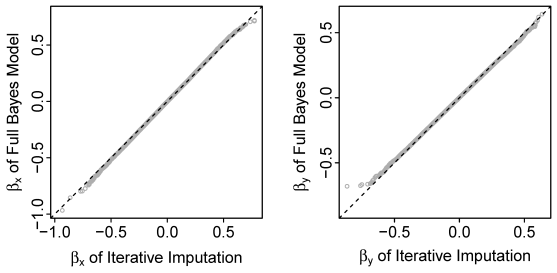

FIG 6.2. *Quantile-quantile plots of the posterior distributions of the Bayesian model versus the compatible iterative imputation distribution for β_x and β_y with sample size $n_a = 200$.*

We monitor the posterior distributions of the following statistics:

$$\beta_x = \frac{\sum_{i\in b} x_i y_i}{\sum_{i\in b} y_i^2}, \quad \beta_y = \frac{\sum_{i\in c} x_i y_i}{\sum_{i\in c} x_i^2}. \tag{6.3}$$

Figures 6.2 shows the quantile-quantile plots of the distributions of β_x and β_y under $\nu_1^{\mathbf{X}^{obs}}$ and $\nu_2^{\mathbf{X}^{obs}}$ based on 1 million MCMC iterations. The differences between these two distributions are tiny.

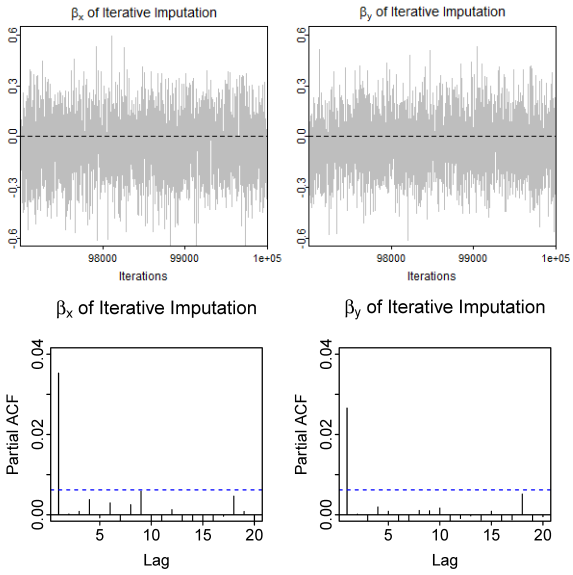

FIG 6.3. *Trace plots and partial autocorrelation functions of β_x and β_y.*

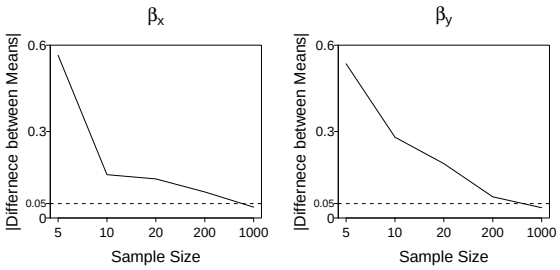

FIG 6.4. *Difference of the expectations of β_x (and β_y) under the two stationary distributions versus sample size n_a*

6.2. *Empirical check of incompatible models with quadratic terms.* Now, we consider an incompatible model with quadratic terms,

$$x_i|y_i \sim N(\alpha_0 + \alpha_1 y_i + \alpha_2 y_i^2, \tau_x^2), \quad y_i|x_i \sim N(\beta_0 + \beta_1 x_i + \beta_2 x_i^2, \tau_y^2). \tag{6.4}$$

This set of conditional models is semi-compatible with a compatible element as in (6.1). The corresponding joint model of this compatible element is

$$(x_i, y_i)^\top \sim N(0, \Sigma). \tag{6.5}$$

It is not hard to verify that the combined estimates associated with models (6.4) and (6.5) satisfy the regularity conditions in D1, D2, D3, and D5. We also empirically check the convergence of the chains (condition D4). We generate bivariate Gaussian random vectors with mean zero, variance one, and correlation zero and set $n_b = n_c = 0.4n_a$. The conditional models are valid. We implemented iterative imputation using the above model with flat prior distributions $\pi_1(\alpha_0, \alpha_1, \alpha_2, \tau_x^2) \propto 1$ and $\pi_2(\beta_0, \beta_1, \beta_2, \tau_y^2) \propto 1$. We check the convergence of the iterative chain empirically via the partial autocorrelation function of β_x and β_y shown in Figure 6.3.

We compare the invariant distributions of the iterative chain and the posterior distribution of the joint Bayesian model

$$(x_i, y_i)^\top \sim N(0, \Sigma)$$

using the statistics monitored in (6.3). As shown in Figure 6.4, the differences of the expectations of β_x (and β_y) under $\nu^{1,\mathbf{X}^{obs}}$ and $\nu^{2,\mathbf{X}^{obs}}$ vanishes as sample size (n_a) becomes large. On the other hand, Figure 6.5 shows the Q-Q plots of the two distributions based on 100,000 iterations. As expected, the total variation distance between the two distributions does not vanish even for $n_a = 1000$. Nevertheless, their expectations are close (difference less than 0.05).

Appendix A: Proof of Proposition 3.16

Let $r(\theta) = g_{\mathbf{X}}(\theta)/f_{\mathbf{X}}(\theta)$ and

$$m(\theta) = f_{\mathbf{X}}(\theta)\min(r(\theta), 1), \quad p = \int_\Theta m(\theta)d\theta \leq 1.$$

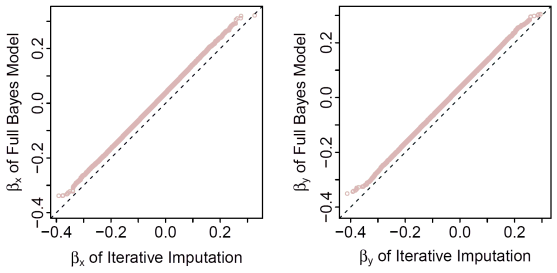

FIG 6.5. *Q-Q plots of the posterior distributions of the joint Bayesian model versus the iterative imputation distribution with squared terms for β_x and β_y with sample size $n = 1000$.*

We write $f_{\mathbf{X}}(\theta) = m(\theta) + d_f(\theta)$ and $g_{\mathbf{X}}(\theta) = m(\theta) + d_g(\theta)$, where $d_f(\theta)$ and $d_g(\theta)$ are nonnegative functions. By defining $q = 1 - p$, we have

$$\int_\Theta d_f(\theta)dx = \int_\Theta d_g(\theta)d\theta = q.$$

Therefore,

$$f_{\mathbf{X}}(\theta) = pc(\theta) + qe_f(\theta), \quad g_{\mathbf{X}}(\theta) = pc(\theta) + qe_g(\theta),$$

where $c(\theta) = m(\theta)/p$, $e_f(\theta) = d_f(\theta)/q$, and $e_g(\theta) = d_g(\theta)/q$ are normalized density functions. Then

$$d_{TV}(f_{\mathbf{X}}, g_{\mathbf{X}}) \leq q.$$

The next step is to provide bounds on q. By the fact that $d_f(\theta)d_g(\theta) = 0$, we have

$$\begin{aligned} q &= \int_\Theta \frac{1}{2}(d_f(\theta) + d_g(\theta))d\theta = \int_\Theta \frac{1}{2}|d_f(\theta) - d_g(\theta)|d\theta \\ &= \int_\Theta \frac{1}{2}|f_{\mathbf{X}}(\theta) - g_{\mathbf{X}}(\theta)|d\theta = \int_\Theta \frac{1}{2}|r(\theta) - 1|f_{\mathbf{X}}(\theta)d\theta \end{aligned}$$

Let $\mu_L = E^f L(\theta)$ and $s_L = SD^f(L(\theta))$. By the Cauchy-Schwarz inequality,

$$q = \frac{E^f|L(\theta) - \mu_L|}{2\mu_L} \leq \frac{s_L}{2\mu_L}.$$

Condition (3.10) and the dominated convergence theorem imply that

$$\mu_L \to L(\mu_\theta), \qquad SD^f(L(\theta)) = (1 + o(1))|\partial L(\mu_\theta) \cdot SD^f(\theta)|,$$

where μ_θ and $SD^f(\theta)$ are the mean and standard deviation under distribution f. Therefore,

$$q \leq \frac{s_L}{2\mu_L} = (1 + o(1))\frac{|\partial L(\mu_\theta) \cdot SD^f(\theta)|}{2L(\mu_\theta)}. \tag{A.1}$$

Hereby, we conclude the proof.

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

[28] 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 6. 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 "$\tilde{}$" 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 A.1. *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 $a_n \leq Cb_n$; $a_n = o(b_n)$ 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 A.1 and A.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 A.3, we give the specific form of a drift function and small set C based on the results in Section A.1 and A.2. The calculations are the same for the Gibbs chain and the iterative chain. Therefore, we do not particularly differentiate them.

A.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, we have

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

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,

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

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 (A.2) and $\sum_a x_i y_i = 0$ into the above display, we have

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

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 $\varepsilon > 0$,

$$s_{b,x}^2 < (1 - 2\varepsilon) s_{a,x}^2, \quad s_{c,y}^2 < (1 - 2\varepsilon) s_{a,y}^2. \tag{A.3}$$

REMARK A.2. *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

$$\begin{aligned}\sum_c \tilde{x}_i y_i &= \gamma \sum_c x_i y_i + O_p\left(\sqrt{s_{c,y}^2 + s_{b,x}^2}\right) + IP, \\ &= \gamma \sum_c x_i y_i + O_p\left(\sqrt{s_{c,y}^2 + s_{b,x}^2}\right). \qquad \text{(A.4)}\end{aligned}$$

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.

A.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,

$$\begin{aligned} \tilde{s}_{b,y}^2 &= \rho_{a,c}^2 \frac{s_{a,y}^2 + s_{c,y}^2}{s_{a,x}^2 + s_{c,x}^2} s_{b,x}^2 + (1 - \rho_{a,c}^2)\bar{\sigma}_y^2 n_b + O_2(\sqrt{n_b}) \\ &= \bar{\sigma}_y^2 n_b \left[(1 - \rho_{a,c}^2) + \rho_{a,c}^2 \frac{(n_a + n_c) s_{b,x}^2}{n_b (s_{a,x}^2 + s_{c,x}^2)} \right] + O_2(\sqrt{n_b}), \end{aligned} \tag{A.5}$$

Similarly,

$$\tilde{s}_{c,x}^2 = \bar{\sigma}_x^2 n_c \left[(1 - \tilde{\rho}_{a,b}^2) + \tilde{\rho}_{a,b}^2 \frac{(n_a + n_b) s_{c,y}^2}{n_c (s_{a,y}^2 + \tilde{s}_{b,y}^2)} \right] + O_2(\sqrt{n_c}). \tag{A.6}$$

Therefore, by plugging (A.5) into (A.6), the evolution of $s_{c,x}^2$ satisfies

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

Define function,

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

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

$$\frac{\tilde{s}_{c,x}^2}{\bar{\sigma}_x^2 n_c} = f\left(\frac{s_{c,x}^2}{\bar{\sigma}_x^2 n_c}, \rho_{a,c}, \tilde{\rho}_{a,b} \right) + O_2(n_c^{-1/2} + n_b^{-1/2}) \tag{A.7}$$

Let λ^* be the solution to

$$f(\lambda^*, \rho_{a,c}, \tilde{\rho}_{a,b}) = \lambda^*. \tag{A.8}$$

Note that 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

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

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

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

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

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

We put all these terms together and obtain,

$$\frac{\partial f(\lambda,\rho_{a,c},\tilde{\rho}_{a,b})}{\partial\lambda} \leq \frac{s_{c,y}^2(s_{a,x}^2+s_{b,x}^2)}{s_{a,y}^2 s_{a,x}^2}. \tag{A.9}$$

Suppose for some $\varepsilon > 0$, we have

$$\frac{s_{c,y}^2(s_{a,x}^2+s_{b,x}^2)}{s_{a,y}^2 s_{a,x}^2} < 1-\varepsilon. \tag{A.10}$$

REMARK A.3. *Similar to condition* (A.3)*, we assume* (A.10) *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 $\Delta\lambda$,

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

where $\lambda^* = f(\lambda^*,\rho,\tilde{\rho})$, uniqueness and existence of which have been proved in standard functional analysis. Therefore, with condition (A.10), $\frac{s_{c,x}^2}{\bar{\sigma}_x^2 n_c}$ goes exponentially fast to λ^*.

A.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

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

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 (A.4), (A.7), (A.9),and $\gamma < \zeta$, we have

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

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

$$
\begin{aligned}
(\text{A.11}) \quad & E\Big\{ O_2\left(\sqrt{s_{c,y}^2 + s_{b,x}^2}\right) \frac{\sum_c x_i y_i}{s_{a,x}^2} + O_2\left(\frac{s_{c,y}^2 + s_{b,x}^2}{s_{a,x}^2}\right) \\
& + \frac{n_c}{A^2}\left[\left|\frac{s_{c,x}^2}{\bar{\sigma}_x^2 n_c} - \lambda^*\right| \left|\tilde{\lambda}^* - \lambda^*\right| + \left(\tilde{\lambda}^* - \lambda^*\right)^2\right] \\
& + \frac{1}{A^2} O_2\left(\sqrt{n_c}\right)\left(\left|\frac{s_{c,x}^2}{\bar{\sigma}_x^2 n_c} - \lambda^*\right| + \left|\tilde{\lambda}^* - \lambda^*\right| + O_2(\frac{1}{\sqrt{n_c}})\right)\Big\} \leq \varepsilon V(x_c)/2.
\end{aligned}
$$

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 (A.11). Note that λ^* is a smooth function of $\rho_{a,c}$. Then, by Taylor's expansion, there exists κ such that

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

Thus,

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

and

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

Therefore, for A sufficiently large and $V(x_c) > A$, we have that (A.11) holds and

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

Therefore, the Markov chain of the iterative imputation under conditions in (A.3) and (A.10) is positive recurrent and the expected recurrent time to the small set C_A is bounded by $V(x_c) + bI_{C_A}(x_c)$.