Convergence Properties of a Sequential Regression Multiple Imputation Algorithm

Jian Zhu[1], Trivellore E. Raghunathan[2]
Department of Biostatistics, University of Michigan, Ann Arbor

Abstract

A sequential regression or chained equations imputation approach is a Gibbs sampling type iterative algorithm that imputes the missing values using a sequence of conditional regression models. It is a flexible approach for handling different types of variables and complex data structures. Many simulation studies have shown that the multiple imputation inferences based on this procedure have desirable repeated sampling properties. However, a theoretical weakness of this approach arises from the fact that the specification of a set of conditional regression models may not be compatible with a joint distribution of the variables being imputed. Hence, the convergence properties of the iterative algorithm are not well understood. In this paper, we develop conditions for convergence and assess the properties of inferences from both compatible and incompatible sequence of regression models. The results are established for the missing data pattern where each subject may be missing a value on at most one variable. We assume that the sequence of regression models are empirically good fit for the data and the imputer has performed appropriate model diagnostics in developing these models. Also, we develop criteria for model choice when specifying the sequence of regression models.

Key Words: Bayesian analysis; Chained equations, Compatible conditionals; Conditional specifications; Exponential family; Gibbs sampling; Missing data; Multiple imputation

1 Introduction

1.1 Background

Consider a data set with p variables, $Y_1, \ldots, Y_p$, with some missing values. The sequential regression (or chained equations, flexible conditional specifications) imputation approach is a Gibbs sampling style iterative algorithm where, at iteration $t = 1, \ldots, T$, the imputations for missing values in variable Y_i are drawn from the posterior predictive distribution, $p(Y_i \mid Y_1^{(t)}, \ldots, Y_{i-1}^{(t)}, Y_{i+1}^{(t-1)}, \ldots, Y_p^{(t-1)})$, where $Y_j^{(t)}$ equals the observed value if available, or an imputed value at iteration t if missing. Denoting $Y_{[-i]}^{(t)} = \{Y_1^{(t)}, \ldots, Y_{i-1}^{(t)}, Y_{i+1}^{(t-1)}, \ldots, Y_p^{(t-1)}\}$,

[1] jianzhu@umich.edu
[2] teraghu@umich.edu

the posterior predictive distribution corresponds to a parametric regression model, $p(Y_i \mid \theta_i, Y_{[-i]}^{(t)})$ and a prior distribution $\pi(\theta_i)$. Denoting $Y_{i,obs}$ and $Y_{i,mis}$ as the observed and missing values of Y_i, the following two step procedure is used to draw the missing values:

Step 1: Draw a value of θ_i, say, θ_i^*, from its posterior density $\pi(\theta_i \mid Y_{i,obs}, Y_{[-i]}^{(t)})$.

Step 2: Draw the set of missing values $Y_{i,mis}$ from the model $p(Y_i \mid \theta_i^*, Y_{i,obs}, Y_{[-i]}^{(t)})$.

For large samples, one may skip the first step and substitute the maximum likelihood or any other consistent estimate of θ_i in the second step. This approach is not Bayesianly proper and may result in understating the variability among the imputed values but may be negligible for large samples. Since our interest is in establishing the asymptotic convergence properties, we skip the draw in Step 1 and use a consistent estimate of θ_i obtained from the data $\{Y_{i,obs}, Y_{[-i]}^{(t)}\}$ (typically the maximum likelihood estimate $\widehat{\theta}_i^{(t)}$) in Step 2.

This approach was first used by Kennickel (1991) for imputing the missing values in continuous variables in the Survey of Consumer Finances using a sequence of linear regression model and used the maximum likelihood estimates of the regression coefficients and the residual variance in Step 2. Brand (1999), Van Buuren and Oudshoorn (1999) and Raghunathan et al (2001) generalized this approach by considering linear regression for continuous, logistic for binary, multinomial logit for more than two categories, Poisson for count and a two-stage model (logistic and then conditional normal) for semi-continuous variables which are generally continuous but have a spike at 0 (For example, real estate income, it is zero for a sizable fraction of the population and a continuous value for the rest).

The sequential regression approach has two major practical advantages over other model-based imputation methods. It enables handling of complex data structures by focusing on individual unidimensional models. The flexible selection of regression models enables better prediction of the missing values based on other variables, and the regression models are more intuitive to analysts than a joint model. Also, individual regression models can easily account for study designs such as skipping patterns, logical constraints, bounds for imputed values and consistency requirements. The software IVEWARE (Raghunathan et al, 2001) implements this approach using a fully Bayesian approach described above for each model. Several other additional features such as placing bounds on the imputed values, restricting the sample to accommodate skip patterns, model tuning and diagnostics are built into the software. A similar approach has been implemented in program MICE in the R programing environment based on the work of Van Buuren and Oudshoorn (1999) and in STATA by Royston (2005). This approach is also available as a part of PROC MI in SAS (2011). The recent issue of the Journal of Statistical Software (2012) has published several articles on this approach.

A theoretical weakness of this approach is due to the fact that the specifications of conditional distributions do not guarantee the existence of a joint distribution, and hence, it is not clear whether the iterative algorithm will achieve any stability. The convergence results established for the standard Gibbs sampling algorithms or its variations may not be applicable. Though, many simulation studies have shown that the multiple imputation inferences using this approach have desirable repeated sampling properties under a variety of conditions, the convergence properties of these algorithms are not known and is often arguable due to incompatibility. This problem was also discussed in the context of spatial analysis (Besag 1974), and necessary and sufficient conditions for the existence of a joint model were given by Arnold and Press (1989) for bivariate conditional densities. Gelman and Speed (1993) also discussed the existence of a unique joint distribution given a set of conditional and marginal distributions. Arnold et. al (2001) gave a thorough introduction to the problem in general, and Gelman and Raghunathan (2001) joined the discussion regarding the effect of incompatible conditionally specified models in missing data analysis. Van Buuren et al (2006) showed through simulations that incompatibility caused minimal effects in some cases.

The fact that incompatibility does not necessarily lead to divergence can be illustrated using the following simple bivariate example. Suppose that the data set of two variables (X, Y) can be divided into three groups: the n_{XY} individuals with both (X, Y) observed, the n_X individuals with missing Y and the n_Y individuals with missing X. Assume that the missing data mechanism is ignorable as defined in Rubin (1976). After an empirical investigation, suppose that an imputer decides to use $m_1(y \mid x, \theta_1) \sim \mathit{exponential}(\theta_1 x)$ and $m_2(x \mid y, \theta_2) \sim \mathit{exponential}(\theta_2 x)$ as conditional regression models. There is no joint distribution with these two conditional distributions. At iteration t, the imputation of missing Y is drawn from $\mathit{exponential}(\widehat{\theta}_1^{(t)} x)$ where

$$\widehat{\theta}_1^{(t)} = (n_{XY} + n_Y)/(\sum_{i \in R_{XY}} x_i y_i + \sum_{i \in R_Y} x_i^{(t-1)} y_i),$$

and the imputed values for the missing X is drawn from $\mathit{exponential}(\widehat{\theta}_2^{(t)} y)$ where

$$\widehat{\theta}_2^{(t)} = (n_{XY} + n_X)/(\sum_{i \in R_{XY}} x_i y_i + \sum_{i \in R_X} x_i y_i^{(t)}).$$

Let ρ_{XY}, ρ_X and ρ_Y be the limiting values of n_{XY}/n, n_X/n and n_Y/n, respectively, as $n \to \infty$. The above two equations, in the limit, are

$$\theta_1^{(t)} = (\rho_{XY} + \rho_Y)/(\rho_Y/\theta_2^{(t-1)} + \rho_{XY} E_o),$$

and

$$\theta_2^{(t)} = (\rho_{XY} + \rho_X)/(\rho_X/\theta_1^{(t)} + \rho_{XY} E_o).$$

where E_o is the expected value of the product XY for the complete cases. It is easy to show that the limiting case of the iterative algorithm given above converges to $\theta_1^* = \theta_2^* = 1/E_o$.

Thus asymptotically, as the sample size, n, the number of iterations, t, and the number of imputations, m, all tend to infinity, the completed-data joint density function (X, Y) averaged over infinite number of imputations, $f_{MI}(x, y)$ tends to $\rho_{XY} f_o(x, y) + (\rho_Y y f_1(y) + \rho_X x f_2(x)) \exp(-xy/E_o)/E_o$ where $f_o(x, y)$ is the joint density of (X, Y) for complete cases, $f_1(y)$ is the marginal density of Y for subjects with missing X and $f_2(x)$ is the marginal density of X for subjects with missing Y. Thus, the practical validity of the multiple imputation inferences depends on the closeness of $m_1(y \mid x, \theta_1)$ and $m_2(x \mid y, \theta_2)$ to the corresponding true conditional distributions $f_1(y \mid x)$ and $f_2(x \mid y)$. Under the missing at random assumption, if the model diagnostics based on the observed data indicate a good fit of the two conditional exponential distributions then the incompatibility may have a very little practical impact on the inferences. For example, if the true joint density function of (X, Y) is $f(x, y) \propto exp(-xy/E_o - \epsilon x - \epsilon y)$ where ϵ is an arbitrarily small positive number, then an imputer is likely to choose the two conditional models given above. In this case, $f_{MI}(x, y)$ is nearly the same as $f(x, y)$ depending upon the ϵ.

On the other hand, suppose that the true joint density function of (X, Y) is $f(x, y) \propto exp(-\alpha xy - \beta x - \gamma y)$ with $\alpha > 0, \beta > 0$ and $\gamma > 0$. The two conditional distributions are exponential distributions with $\alpha x + \beta$ as the parameter for $f(y \mid x)$ and $\alpha y + \gamma$ as the parameter of $f(x \mid y)$ (Arnold and Strauss (1988)) . Again, assume that the missing data mechanism is ignorable and that the imputations are carried out under the following sequence of regression models, $m(x \mid y) \sim exponential(\phi_1 + \phi_2 y)$ and $m(y \mid x) \sim exponential(\phi_3 + \phi_4 x)$. These two conditional distributions are not compatible with any joint distribution unless $\phi_2 = \phi_4$. Note that the functional form of the two conditional densities match the true densities, and the two conditional densities are compatible when ($\phi_2 = \phi_4 > 0$, $\phi_1 > 0$ and $\phi_3 > 0$), a subspace of the joint parameter space ($\phi_i > 0, i = 1, 2, 3, 4$) used by the imputer. We may view these two conditionals as "over parameterized" where the joint distribution is embedded within the joint parameter space of the conditional distributions used in the imputation process. For such situations, Theorem 1 given in the next section provides sufficient conditions for the sequential regression imputation approach yields a consistent estimator of the joint density function of (x, y). In fact, many standard conditional regression models satisfy the sufficient conditions.

The rest of the paper is organized as follows. Section 2 provides definition of incompatibility and model validity to classify regression models in the sequential regression approach. Section 3, focusing on bivariate scenario, provides two sufficient conditions for the convergence of the sequential regression approach and resulting in consistent estimators. Section 4 enhances the analytical results given in Sections 3 through a simulation study for incompatible but approximately valid or well fitting model sequences. Section 5 extends the results for multivariate missing data with single variable missing data pattern (that is, any subject is missing at most one variable). Section 6 summarizes the findings, discusses extensions for arbitrary pattern of missing data and the limitation of the sequential regression algorithm.

2 Classification of Regression Model Sequences

Before we establish the convergence and consistency properties, we define the degree or types of incompatibility among the conditionally specified regression model in the sequential regression algorithm. We consider two types of incompatible models, one with reference to the true or actual distribution and another without any reference to the true distribution. The former is of more theoretical interest or when the posited joint distribution is too complicated and an imputer would like to find an approximately valid sequential regression model. The latter is tuned towards selecting the kind of sequential regression models that will lead to convergence.

> **Definition 1 (Weakly Incompatible Model Sequence):** Supposes that the joint density function, $f(y_1, \ldots, y_p)$ has the conditional densities $f(y_i \mid y_{[-i]}, \psi_i)$, $i = 1, 2, \ldots, p$. A regression model $m_i(y_i \mid y_{[-i]}, \theta_i)$ with $\theta_i \in \Theta_i$ is defined to be validly specified for $f(y_i \mid y_{[-i]}, \psi_i)$ if the following condition holds: for any ψ_i, θ_i can be expressed as $(g(\psi_i), \xi_i)$, and there exists $\theta_i^0 = (g(\psi_i), 0) \in \Theta_i$ such that $m_i(y_i \mid y_{[-i]}, \theta_i^0) = f(y_i \mid y_{[-i]}, \psi_i)$.
>
> A sequence of regression models is defined to be weakly incompatible if each regression model in the sequence is validly specified.

For example, both $m_(y \mid x, \theta) \sim N(\theta_{10} + \theta_{11}x, \sigma^2)$ and $m_(y \mid x, \theta) \sim N(\theta_{10} + \theta_{11}x + \theta_{12}x^2, \sigma^2)$ are validly specified models for the conditional density $y \mid x \sim N(2 + x, 1)$. The former is exactly specified, and the latter has an extra term with the parameter $\xi = \theta_{12}$.

> **Definition 2 (Possibly Compatible Models):** A sequence of regression models $m_i(y_i \mid y_{[-i]}, \theta_i), \theta_i \in \Theta_i$ is defined to be possibly compatible, if there exists a target joint density function $p(y_1, \ldots, y_p \mid \theta)$ with conditional density functions, $p_i(y_i \mid y_{[-i]}, \theta_{Y_i|Y_{[-i]}})$ for $i = 1, 2, \ldots, p$ such that the exact functional form of m_i is the same as p_i for some subspace $\Theta_C \subseteq \Theta_1 \times \Theta_2 \times \cdots \times \Theta_p$ and $(\theta_{Y_1|Y_{[-1]}}, \ldots, \theta_{Y_p|Y_{[-p]}})$ can be functionally expressed in terms of $(\theta_1, \theta_2, \ldots, \theta_p)$.

We now provide several examples of conditional regression models commonly used in the sequential regression multivariate imputation algorithm and relate them to above definitions.

The following four examples are possibly compatible models, and they can be weakly incompatible if they correspond to the true conditional densities:

Example 1 (Two Linear Regression Models): Suppose (X, Y) are two continuous variables, and sequential regression imputation assumes two linear regression models:

$$\begin{aligned} m_1 : y \mid x, \theta_1 &\sim N\left(\theta_{10} + \theta_{11}x, \sigma_{12}^2\right), \\ m_2 : x \mid y, \theta_2 &\sim N\left(\theta_{20} + \theta_{21}y, \sigma_{21}^2\right). \end{aligned}$$

The target joint distribution is a bivariate normal distribution. The compatibility condition is $\theta_{11}/\sigma_{12}^2 - \theta_{21}/\sigma_{21}^2 = 0$, $\theta_{11}^2 \neq \sigma_{12}^2/\sigma_{21}^2$ and $\theta_{21}^2 \neq \sigma_{21}^2/\sigma_{12}^2$, where the first equation ensures that $m_1(y \mid x, \theta_1)/m_2(x \mid y, \theta_2)$ can be expressed as $m(y)/m(x)$ and the latter two inequalities ensure that $m(y)$ and $m(x)$ are integrable.

Example 2 (Two Logistic Regression Models): Suppose (X, Y) are two binary variables, and sequential regression imputation assumes two logistic regression models:

$$\begin{aligned} m_1 : y \mid x, \theta_1 &\sim Bernoulli\left[(1 + \exp(-\theta_{10} - \theta_{12}x))^{-1}\right], \\ m_2 : x \mid y, \theta_2 &\sim Bernoulli\left[(1 + \exp(-\theta_{20} - \theta_{21}y))^{-1}\right]. \end{aligned}$$

The target joint distribution is a bivariate Bernoulli distribution and the compatibility condition is $\theta_{12} - \theta_{21} = 0$.

Example 3 (Two Conditional Exponential Models With Intercepts): Suppose (X, Y) are two positive continuous variables, and sequential regression imputation assumes

$$\begin{aligned} m_1 : y \mid x, \theta_1 &\sim Exp(\theta_{10} + \theta_{11}x), \\ m_2 : x \mid y, \theta_2 &\sim Exp(\theta_{20} + \theta_{21}y). \end{aligned}$$

The compatibility condition is $\theta_{11} - \theta_{21} = 0$.

Example 4 (Two Conditional Poisson Models): Suppose (X, Y) are two count variables, and sequential regression imputation assumes two Poisson regression models with canonical links:

$$\begin{aligned} m_1 : y \mid x, \theta_1 &\sim Poisson(\exp(\theta_{10} + \theta_{11}x)), \\ m_2 : x \mid y, \theta_2 &\sim Poisson(\exp(\theta_{20} + \theta_{21}y)). \end{aligned}$$

The compatibility condition is $\theta_{11} - \theta_{21} = 0$ and $\theta_{11} < 0$.

The following two examples are not possibly compatible models:

Example 5 (Two Conditional Exponential Models): Suppose (X, Y) are two positive continuous variables, and sequential regression imputation assumes

$$\begin{aligned} m_1 : y \mid x, \theta_1 &\sim Exp\left(\theta_1 x\right), \\ m_2 : x \mid y, \theta_2 &\sim Exp\left(\theta_2 y\right). \end{aligned}$$

Example 6 (Two Gamma Regression Models): Suppose (X, Y) are two non-negative continuous variables, and sequential regression imputation assumes two Gamma regression models:

$$\begin{aligned} m_1 : y \mid x, \theta_1 &\sim \Gamma(K)^{-1}(\theta_{10} + \theta_{11}x)^{-K} y^{K-1} \exp\{-y(\theta_{10} + \theta_{11}x)^{-1}\}, \\ m_2 : x \mid y, \theta_2 &\sim \Gamma(J)^{-1}(\theta_{20} + \theta_{21}y)^{-J} x^{J-1} \exp\{-x(\theta_{20} + \theta_{21}y)^{-1}\}. \end{aligned}$$

We now define a subclass of possibly compatible model sequence where the parameters are separable between conditional and marginal distributions. This separability of the parameters was first used by Anderson (1958) to develop maximum likelihood estimates for the mean and the covariance matrix of multivariate normal distribution.

Definition 3 (Possibly Compatible Model Sequence With Separable Marginal Parameters): A joint density function $p(y_1, \ldots, y_p \mid \theta)$ is defined to have separable marginal parameters if for any subset Y_M of $Y = \{y_1, \ldots, y_p\}$, $\theta_{Y_C|Y_M}$ is distinctive from θ_{Y_M}, where $Y_C = Y - Y_M$, $\theta_{Y_C|Y_M}$ is from the conditional distribution $p(Y_C \mid Y_M, \theta_{Y_C|Y_M})$ and θ_{Y_M} is from the marginal distribution $p(Y_M \mid \theta_{Y_M})$. Equivalently, separable marginal parameters imply that for the parameterization $(\theta_{Y_M}, \theta_{Y_C|Y_M})$, the parameter space is the product of two independent parameter spaces $\Theta = \Theta_{Y_M} \times \Theta_{Y_C|Y_M}$.

A sequence of possibly compatible regression models is defined to have separable marginal parameters if the target joint density function has separable marginal parameters.

Possibly compatible regression model sequences with separable marginal parameters include the linear regression models and logistic regression models for binary variables from Example 1 and Example 2, and they can be extended to multivariate settings:

Example 1 (Linear Model Sequence):

Two linear regression models: Suppose (X, Y) are two continuous variables, and sequential regression imputation assumes two linear regression models:

$$\begin{aligned} m_1 : y \mid x, \theta_1 &\sim N\left(\theta_{10} + \theta_{11}x, \sigma_{12}^2\right), \\ m_2 : x \mid y, \theta_2 &\sim N\left(\theta_{20} + \theta_{21}y, \sigma_{21}^2\right). \end{aligned}$$

The target joint distribution is a bivariate normal distribution and the necessary compatibility condition is $\theta_{11}/\sigma_{12}^2 - \theta_{21}/\sigma_{21}^2 = 0$. The marginal parameters are separable as follows: $m_X(x) \sim N(\mu_x, \sigma_x^2)$ and $m_1(y \mid x, \theta_1) \sim N(\theta_{10} + \theta_{11}x, \sigma_{12}^2)$; $m_Y(y) \sim N(\mu_y, \sigma_y^2)$ and $m_2(x \mid y, \theta_2) \sim N(\theta_{20} + \theta_{21}y, \sigma_{21}^2)$. Each marginal-conditional decomposition of the joint model has parameters one-to-one mapped to the joint model parameters, and marginal model parameters are distinctive from (or independent of) the conditional model parameters.

Multivariate linear regression model sequence: Suppose $(Y_1, \ldots, Y_p)$ are p-dimension continuous variables, and model sequence consists of

$$m_i(y_i \mid y_{[-i]}, \theta_i) \sim N(\theta_{i0} + \sum_{j \neq i} \theta_{ij} y_j, \sigma_i^2), i = 1, \ldots, p.$$

The target joint distribution is a multivariate normal distribution and the necessary compatibility condition is that for any $i \neq j$, $\theta_{ij}/\sigma_i^2 - \theta_{ji}/\sigma_j^2 = 0$. The marginal parameters are separable as follows: for any subset Y_M of $Y = \{y_1, \ldots, y_p\}$, $m_M(y_M) \sim MVN(\mu_M, \Sigma_M)$ and for any $y_i \in Y_C = Y - Y_M$, $m_i(y_i \mid y_{[-i]}, \theta_i) \sim N(\theta_{i0} + \sum_{j \neq i} \theta_{ij} y_j, \sigma_i^2)$.

Example 2 (Logistic Regression Model Sequence for Binary Variables):

Two logistic regression models: Suppose (X, Y) are two binary variables, and sequential regression imputation assumes two logistic regression models:

$$\begin{aligned} m_1 : y \mid x, \theta_1 &\sim Bernoulli\left[(1 + \exp(-\theta_{10} - \theta_{12} x))^{-1}\right], \\ m_2 : x \mid y, \theta_2 &\sim Bernoulli\left[(1 + \exp(-\theta_{20} - \theta_{21} y))^{-1}\right]. \end{aligned}$$

The target joint distribution is a bivariate Bernoulli distribution and the compatibility condition is $\theta_{12} - \theta_{21} = 0$. The marginal parameters are separable as follows: $m_X(x) \sim Bernoulli(p_X)$ and $m_1(y \mid x, \theta_1) \sim Bernoulli[(1 + \exp(-\theta_{10} - \theta_{12} x))^{-1}]$; and $m_Y(y) \sim Bernoulli(p_Y)$ and $m_2(x \mid y, \theta_2) \sim Bernoulli[(1 + \exp(-\theta_{20} - \theta_{21} y))^{-1}]$.

Multivariate logistic regression model sequence: Suppose $(Y_1, \ldots, Y_p)$ are p-dimension binary variables, and model sequence is for $i = 1, \ldots, p$,

$$m_i(y_i \mid y_{[-i]}, \theta_i) \sim Bernoulli\left[\left(1 + \exp\left(-\theta_{i0} - \sum_{j \neq i} \theta_{ij} y_j - \sum_{j \neq i, k \neq i, k < j} \theta_{ijk} y_j y_k\right)\right)^{-1}\right].$$

The target joint distribution is a multivariate Bernoulli distribution and the compatibility condition is that for any different i, j and k, $\theta_{ij} = \theta_{ji}$ and $\theta_{ijk} = \theta_{jik} = \theta_{kij}$. The marginal parameters are separable as follows: for any subset Y_M of $Y = \{y_1, \ldots, y_p\}$, Y_M follows a multivariate Bernoulli distribution and for any $y_i \in Y_C = Y - Y_M$,

$$m_i(y_i \mid y_{[-i]}, \theta_i) \sim Bernoulli\left[\left(1 + \exp\left(-\theta_{i0} - \sum_{j \neq i} \theta_{ij} y_j - \sum_{j \neq i, k \neq i, k < j} \theta_{ijk} y_j y_k\right)\right)^{-1}\right].$$

Examples in which the target joint distribution does not have separable marginal parameters include the bivariate exponential distribution and the bivariate Poisson distribution from Example 3 and Example 4.

In summary, Definition 1 classifies all regression model sequences into valid and invalid sequences with reference to the true joint density function of the variables being imputed; Definition 2 classifies model sequences into possibly compatible and incompatible sequences regardless of the true underlying joint distribution of the variables. The possibly compatible sequence has a target joint density function within the parameter space; Definition 3 defines a subclass of possibly compatible sequences based on the property of the target joint distribution's marginal parameter property.

3 Bivariate Missing Data

Before we consider the multivariate imputation problem, we consider the bivariate case, mostly for notational simplicity and ease of presentation. We also assume that the missing data mechanism is ignorable as in Rubin (1976) and all the conditional distributions belong to the exponential family. The convergence and consistency are asymptotic properties as the sample size, the number of imputation and the number of iterations or sequential updates all tend towards ∞.

Suppose that (X, Y) follows a joint distribution with the joint density $f_{XY}(x, y \mid \psi)$, the marginal densities $f_X(x \mid \psi_X)$ and $f_Y(y \mid \psi_Y)$, and the conditional densities $f_{Y|X}(y \mid x, \psi_1)$ and $f_{X|Y}(x \mid y, \psi_2)$. Let R denote the response pattern where $R = 0$ consists of complete cases $\{(x_{0i}, y_{0i})\}, i = 1, \ldots, n_0$; $R = 1$ consists of cases with missing X but observed Y, $\{y_{1j}, j = 1, \ldots, n_1\}$; and $R = 2$ consists of cases with missing Y but observed X, $\{x_{2k}, k = 1, \ldots, n_2\}$. The missing data to be imputed consists of $\{x_{1j}, j = 1, \ldots, n_1\}$ when $R = 1$ and $\{y_{2k}, k = 1, \ldots, n_2\}$ when $R = 2$. The total sample size is $n = n_0 + n_1 + n_2$. We also assume that the proportion of missing data will be nontrivial in a sense that as $n \to \infty$, $n_0/n \to \rho$ and $n_1/n \to \rho_1$, where $0 < \rho < 1$ and $0 < \rho_1 < 1 - \rho$. We denote $\Pr(R = 1 \mid X, Y) = g_1(y)$, $\Pr(R = 2 \mid X, Y) = g_2(x)$ and $\Pr(R = 0 \mid X, Y) = 1 - g_1(y) - g_2(x)$, where parameters in g_1 and g_2 are distinct from ψ, the parameters in the complete data model. It is easy to show that $f(x \mid y, R = 1) = f(x \mid y, R \neq 1) = f_{X|Y}(x \mid y, \psi_2)$, and $f(y \mid x, R = 2) = f(y \mid x, R \neq 2) = f_{Y|X}(y \mid x, \psi_1)$.

The sequential regression imputation algorithm assumes a regression model $m_1(y \mid x, \theta_1)$ for y given x and $m_2(x \mid y, \theta_2)$ for x given y respectively. We assume that the regression models are generalized linear models from the exponential family:

$$
\begin{aligned}
m_1(y \mid \phi_1, \delta_1, x) &= \exp\left\{\left[T_1(y)^T \phi_1 - b_1(\phi_1)\right]/a_1(\delta_1) + c_1(y, \delta_1)\right\}, \\
m_2(x \mid \phi_2, \delta_2, y) &= \exp\left\{\left[T_2(x)^T \phi_2 - b_2(\phi_2)\right]/a_2(\delta_2) + c_2(x, \delta_2)\right\},
\end{aligned}
$$

where $\theta_i = (\phi_i, \delta_i), i = 1, 2$ and the link functions h_1 and h_2 connect the conditional means and predictor variables through $h_1^{-1}(\sum_{u=0}^{U} \theta_{1u} h_{1u}(x)) = b_1'(\phi_1)$ and $h_2^{-1}(\sum_{v=0}^{V} \theta_{2v} h_{2v}(y)) = b_2'(\phi_2)$.

At iteration t, the algorithm is executed in two steps:

Step 1: ${\theta_1}^{(t)}$ is estimated by regressing $\{y_{0i}, y_{1j}\}$ on $\{x_{0i}, x_{1j}^{(t-1)}\}$ with model m_1, and the missing values of y, $\{y_{2k}^{(t)}\}$, are drawn from the conditional distribution $m_1(y \mid \{x_{2k}\}, \theta_1^{(t)})$;

Step 2: ${\theta_2}^{(t)}$ is estimated by regressing $\{x_{0i}, x_{2k}\}$ on updated $\{y_{0i}, y_{2k}^{(t)}\}$ with model m_2, and the missing values of X, $\{x_{1j}^{(t)}\}$, are drawn from $m_2(x \mid \{y_{1j}\}, \theta_2^{(t)})$.

To be specific, the above two steps calculate the log-likelihood functions at iteration t for the two models:

$$l_1(\theta_1 \mid X_{obs}, Y_{obs}, X_{mis}^{(t-1)}) = \sum_i \log m_1(y_{0i} \mid x_{0i}, \theta_1) + \sum_j \log m_1(y_{1j} \mid x_{1j}^{(t-1)}, \theta_1),$$
$$l_2(\theta_2 \mid X_{obs}, Y_{obs}, Y_{mis}^{(t)}) = \sum_i \log m_2(x_{0i} \mid y_{0i}, \theta_2) + \sum_k \log m_2(x_{2k} \mid y_{2k}^{(t)}, \theta_2),$$

and estimate the parameters $(\theta_1^{(t)}, \theta_2^{(t)})$ by solving the score equations:

$$s_1(\theta_1 \mid X_{obs}, Y_{obs}, X_{mis}^{(t-1)}) = \partial l_1(\theta_1 \mid X_{obs}, Y_{obs}, X_{mis}^{(t-1)})/\partial\theta_1 = 0,$$
$$s_2(\theta_2 \mid X_{obs}, Y_{obs}, Y_{mis}^{(t)}) = \partial l_2(\theta_2 \mid X_{obs}, Y_{obs}, Y_{mis}^{(t)})/\partial\theta_2 = 0.$$

The completed data set at iteration T consists of $\{(x_{0i}, y_{0i}), (x_{1j}^{(T)}, y_{1j}), (x_{2k}, y_{2k}^{(T)})\}$. Suppose $\theta_1^{(T)}$ and $\theta_2^{(T)}$ are the estimates of θ_1 and θ_2 respectively. We wish to study the properties of these estimates as n and T tends to ∞.

When the sample size is large and with infinite number of imputations, the score equations given above can be approximated by (or tend to) the following equations:

$$\begin{aligned}
\tilde{s}_1(\theta_1 \mid {\theta_2}^{(t-1)}, \psi) &= n_0 \iint \frac{\partial \log m_1(y \mid x, \theta_1)}{\partial\theta_1} f_{XY}(x, y \mid R = 0)\mathrm{d}x\mathrm{d}y \\
&\quad + n_1 \iint \frac{\partial \log m_1(y \mid x, \theta_1)}{\partial\theta_1} m_2(x \mid y, {\theta_2}^{(t-1)})\mathrm{d}x f_Y(y \mid R = 1)\mathrm{d}y. \\
\tilde{s}_2(\theta_2 \mid {\theta_1}^{(t)}, \psi) &= n_0 \iint \frac{\partial \log m_2(x \mid y, \theta_2)}{\partial\theta_2} f_{XY}(x, y \mid R = 0)\mathrm{d}x\mathrm{d}y \\
&\quad + n_2 \iint \frac{\partial \log m_2(x \mid y, \theta_2)}{\partial\theta_2} m_1(y \mid x, {\theta_1}^{(t)})\mathrm{d}y f_X(x \mid R = 2)\mathrm{d}x.
\end{aligned}$$

Then both $\tilde{s}_1(\theta_1^{(t)} \mid \theta_2^{(t-1)}, \psi)$ and $\tilde{s}_2(\theta_1^{(t)} \mid \theta_2^{(t)}, \psi)$ converge to 0 in probability as $n \to \infty$, which lead to an approximate iterative algorithm $\tilde{s}_1(\theta_1^{(t)} \mid \theta_2^{(t-1)}, \psi) = 0$ and $\tilde{s}_2(\theta_2^{(t)} \mid \theta_1^{(t)}, \psi) = 0$.

Therefore, the implicit recursive algorithm $\theta_1^{(t)} = \tilde{s}_1^{-1}(\theta_2^{(t-1)}, \psi), \theta_2^{(t)} = \tilde{s}_2^{-1}(\theta_1^{(t)}, \psi)$ has the convergence property similar to that of the imputation algorithms asymptotically.

Theorem 1. *Suppose that the imputation models are weakly incompatible as defined in the previous section and the conditional distributions satisfy the following usual regularity conditions:*

1. *The density functions m_1 and m_2 are differentiable with respect to θ_1 and θ_2 respectively and the differentiation and integration are interchangeable with respect to (x,θ_1) for m_1 and (y, θ_2) for m_2 respectively*
2. *The mean and the variance of the score functions given above exist under both the posited (m_1, m_2) and the true models $(f_{X|Y}, f_{Y|X})$.*

Then as the sample size n, the number of imputations m and the number of iterations t tend to ∞, the regression models $m_1(y \mid x, \theta_1^{(t)}) \rightarrow f_{Y|X}(y \mid x, \psi_1)$ and $m_2(x \mid y, \theta_2^{(t)}) \rightarrow f_{X|Y}(x \mid y, \psi_2)$.

The proof of the Theorem is given in Appendix 1. To illustrate further, we consider, Examples 1 and 2 described above and assess the convergence properties of the asymptotic iterative algorithm.

Example 1 (Two Linear Regression Models revisited): Suppose the data (X, Y) are generated from a bivariate normal distribution BVN(μ, Σ) with the conditional distributions $y \mid x \sim N(\alpha_{10} + \alpha_{11}x, \tau_{12}^2)$ and $x \mid y \sim N(\alpha_{20} + \alpha_{21}y, \tau_{21}^2)$. where $\alpha_{11}/\tau_{12}^2 = \alpha_{21}/\tau_{21}^2$. Suppose data are missing completely at random: $\pi_0 = pr(R = 0)$, $\pi_1 = pr(R = 1)$ and $\pi_2 = pr(R = 2)$. The imputation model sequence consists of two linear regression models from Example 1.The asymptotic iterative algorithm is calculated in Appendix 2. The estimated regression parameters are shown to converge to $\theta_1^* = (\alpha_{10}, \alpha_{11}, \tau_{12}^2)^T$, and $\theta_2^* = (\alpha_{20}, \alpha_{21}, \tau_{21}^2)^T$. The rate of convergence for the iterative algorithm is $\pi_1\pi_2/\{(\pi_0 + \pi_1)(\pi_0 + \pi_2)\}$.

Example 2 (Two Logistic Regression Models revisited): Suppose the data (X, Y) are generated from a bivariate Bernoulli distribution with $pr(X = 0, Y = 0) = p_{00}$, $pr(X = 0, Y = 1) = p_{01}$, $pr(X = 1, Y = 0) = p_{10}$ and $pr(X = 1, Y = 1) = p_{11} = 1 - p_{00} - p_{01} - p_{10}$, with conditional distributions $y \mid x \sim Bernoulli\{(1 + \exp(-\alpha_{10} - \gamma_{12}x))^{-1}\}$ and $x \mid y \sim Bernoulli\{(1 + \exp(-\alpha_{20} - \gamma_{21}y))^{-1}\}$, where $\gamma_{12} = \gamma_{21}$. Suppose data are missing completely at random: $\pi_0 = pr(R = 0)$, $\pi_1 = pr(R = 1)$ and $\pi_2 = pr(R = 2)$. The imputation model sequence consists of two logistic regression models from Example 2. The asymptotic iterative algorithm is calculated in Appendix 2. The estimated regression parameters are shown to converge to $\theta_1^* = (\alpha_{10}, \gamma)^T$, and $\theta_2^* = (\alpha_{20}, \gamma)^T$ where $\gamma_{12} = \gamma_{21} = \gamma$.

We now show the results for the possibly compatible models, where the posited conditional models may not agree with the true distributions but may be compatible with some joint distribution in a subset of the parameter space. The following theorem provides conditions for the convergence of the sequential regression imputation algorithm.

Theorem 2. *Suppose a sequential regression imputation algorithm uses possibly compatible models $m_1(y \mid x, \theta_1)$ and $m_2(x \mid y, \theta_2)$, with $p_{XY}(x, y \mid \theta_1, \theta_2)$ as the joint distribution only when $\theta = (\theta_1, \theta_2) \in \Theta_C \subset \Theta_1 \times \Theta_2$. If $p_{XY}(x, y \mid \theta_1, \theta_2, \theta \in \Theta_C)$ has separable marginal parameters and (θ_1^*, θ_2^*) is the maximum likelihood estimate of (θ_1, θ_2) from the joint model, then under the same regularity conditions in Theorem 1 with respect to differentiation/integration and the existence of the mean/variance of the score functions , $m_1(y \mid x, \theta_1^{(t)}) \rightarrow p(y \mid x, \theta_1^*)$ and $m_2(x \mid y, \theta_2^{(t)}) \rightarrow p(x \mid y, \theta_2^*)$ as $n,\ t \rightarrow \infty$.*

The proof of this theorem is given in Part 2 of Appendix 1. Note that if the compatibility condition is strictly imposed when θ_1 and θ_2 are estimated at each iteration, then the imputation algorithm is a simplified version of a standard Markov chain with convergence to a stationary joint distribution. However, the sequential regression imputation does not estimate θ_1 and θ_2 simultaneously within one iteration, and the compatibility condition is ignored in the estimation process. For sequences with separable marginal parameters such as Example 1 and Example 2, since $({\theta_1}^*, {\theta_2}^*) \in \Theta_C$ holds inherently, according to Theorem 2, the compatibility condition is approximately satisfied for $({\theta_1}^{(t)}, {\theta_2}^{(t)})$ after a certain number of iterations. However, we will show that this is not always true for possibly compatible sequences without separable marginal parameters.

When a possibly compatible model sequence does not have separable marginal parameters, the marginal distributions $p(x \mid \theta_1, \theta_2)$ and $p(y \mid \theta_1, \theta_2)$ from the target joint distribution also depend on regression parameters, and, hence, the log-likelihood functions from the sequential regression imputation and joint modeling imputation differ. For an heuristic explanation, consider θ_1 from m_1 as an example. For any single observation, the log-density function involving θ_1 is $\log(m_1(y \mid x, \theta_1))$ from the sequential regression model, where as it is $\log(m_1(y \mid x, \theta_1) p(x \mid \theta_1, \theta_2))$ from the joint model. Because the distribution of observed X involves θ_1, in general, the log-likelihood functions of θ_1 from the joint model and the sequential regression differ. Therefore, the two algorithms yield different parameter estimates and imputation results.

To clarify this aspect further, consider the following simulation examples:

Example 4 (Two Poisson Regression Models revisited): For two Poisson imputation models defined in Example 4, the compatibility condition requires that $\theta_{11} = \theta_{21} < 0$, and the joint model is $m(x, y \mid \theta_1, \theta_2, \theta_{11} = \theta_{21} < 0) = c(\theta_{10}, \theta_{20}, \theta_{11}) \exp(\theta_{10} y +$

$\theta_{20}x+\theta_{11}xy)$, where $c(\theta_{10}, \theta_{20}, \theta_{11})$ is the normalizing constant. The log-density function involving $(\theta_{10}, \theta_{11})$ is $-\exp(\theta_{10}+\theta_{11}x)+(\theta_{10}+\theta_{11}x)y$ from the conditionally specified model m_1, and $\log(c(\theta_{10}, \theta_{20}, \theta_{11}))+(\theta_{10}+\theta_{11}x)y$ from the joint model. For three different bivariate count data sets, we applied the same sequential regression imputation algorithm assuming two conditional Poisson regression models from Example 4 (T, the number of iterations, is set as 10000):

(1) We generated the complete data from $Y \sim \text{Poisson}(2.5)$ and $X \mid Y \sim \text{Poisson}(\exp(3-0.3Y))$, and the data are missing completely at random with $n_0 = n_1 = n_2 = 10000$. Sequential regression imputation estimates are approximately $\theta_{11}^{(T)} = -0.1$ and $\theta_{21}^{(T)} = -0.2$. Although both slope estimates are negative, they are not equal, and the compatibility condition is not satisfied.

(2) We generated the complete data from $Y \sim \text{Poisson}(2.5)$ and $X \mid Y \sim \text{Poisson}(\exp(-1+0.3Y))$, and the data are missing completely at random with $n_0 = n_1 = n_2 = 10000$. Sequential regression imputation estimates are approximately $\theta_{11}^{(T)} = 0.2$ and $\theta_{21}^{(T)} = 0.25$. They are neither negative nor equal, and the compatibility condition is not satisfied.

(3) We generated the complete data from a bivariate Poisson distribution $\propto \exp(2y+x-0.3xy)$, with conditionals $Y \mid X \sim \text{Poisson}(2-0.3X)$ and $X \mid Y \sim \text{Poisson}(\exp(1-0.3Y))$, and the data are missing completely at random with $n_0 = n_1 = n_2 = 10000$. Sequential regression imputation estimates are $\theta_{11}^{(T)} = -0.3$ and $\theta_{21}^{(T)} = -0.3$. The imputation results are compatible since both models are correctly specified.

The simulations show that in general the possibly compatible regression model sequences with non-separable marginal parameters do not converge to the joint models (Situations (1) and (2)), unless the conditional distributions are correctly specified (Situation(3)). The practical consequence of these findings is that to yield approximately unbiased results, both conditional distributions have to be as close to the corresponding true conditional distributions as possible to achieve convergence, regardless of compatibility with respect to any joint distribution. This underscores the importance of model diagnostics to check the conditional regression model fit to the data.

When the model sequence is neither weakly incompatible or possibly compatible, then there is no joint model for the sequence to converge to. However, as we showed in Section 1, the sequential regression algorithm can still converge. In general, the estimates from sequential regression imputation algorithms with incompatible models depend on the population distribution, the missing data mechanism and the regression models. It is difficult, if not impossible, to obtain analytical results about the convergence except for some examples. We now describe the results from simulation study designed to study the properties of the sequential regression algorithm for such incompatible but empirically well-fitting regression models.

4 Simulation Studies for a Bivariate Missing Data

One approach to define a well-fitting regression model is through Kullback-Leibler divergence measure. For example, the maximum likelihood estimates of the parameters in the regression model $m_1(y \mid x, \theta_1)$ can be viewed as an asymptotic equivalent to those obtained by minimizing the relative entropy of the regression model, $\iint \log[f(y \mid x)/m_1(y \mid x, \theta_1)]f(x, y)\mathrm{d}x\mathrm{d}y$ or the Kullback-Leibler divergence between the regression model and the true conditional density. Since it is asymmetric and does not satisfy the triangle inequality, it is not a metric. However, the divergence is always positive unless the two distributions are the same, therefore it is often used to describe the discrepancy between the two distributions. We calculate the divergence between the fitted regression model and the true conditional distribution $\int \log[f(y \mid x)/m_1(y \mid x, \theta_1)]f(y \mid x)\mathrm{d}y$ at different values of x, and use the divergence curve to describe the model fitness regarding the true model. For a well-fitting model sequence, when the divergence curve between each regression model and the true conditional model is approximately 0, draws from the fitted regression model can be approximately treated as draws from the true model.

We now use Example 6 to show that a well-fitting incompatible model sequence can be approximately validly specified:

Example 6 (Two Gamma Regression Models revisited):

$$\begin{aligned} y \mid x, \theta_1 &\sim \Gamma(K)^{-1}(\theta_{10} + \theta_{11}x)^{-K}y^{K-1}\exp\{-y(\theta_{10} + \theta_{11}x)^{-1}\}, \\ x \mid y, \theta_2 &\sim \Gamma(J)^{-1}(\theta_{20} + \theta_{21}y)^{-J}x^{J-1}\exp\{-x(\theta_{20} + \theta_{21}y)^{-1}\}. \end{aligned}$$

For the simulation study, we generated data from the following population distribution:

$$\begin{aligned} f_X(x \mid \psi_x) &= \beta^K\Gamma(K)^{-1}x^{-K-1}\exp(-\beta/x), \\ f_{Y|X}(y \mid x, \psi_1) &= \Gamma(J)^{-1}(\alpha x)^{-J}y^{J-1}\exp\{-y(\alpha x)^{-1}\}. \end{aligned}$$

Then X follows a marginal inverse Gamma distribution and Y given X follows a conditional Gamma distribution. The corresponding conditional distribution of X given Y is

$$f_{X|Y}(x \mid y, \psi_2) = \Gamma(K + J)^{-1}(\beta + y/\alpha)^{K+J}x^{-(K+J)-1}\exp\{-(\beta + y/\alpha)x^{-1}\}.$$

The following parameters are chosen for the distributions: $K = 3$, $\beta = 3$, $J = 5$, and $\alpha = 0.25$.

We generated 500 data sets of sample size n=50, 100, 200, 500, 1000 and 10,000 from the bivariate distribution defined by f_X and $f_{Y|X}$ described as above. Some values were set to missing based on the following missing at random mechanism: first, data are divided equally into two random groups. In the first group, y is fully observed and the probability of x being observed is $pr(x \text{ is observed} \mid y) = [1 + \exp(-1 - 0.4y)]^{-1}$; In the second group, x is fully observed and the probability of y being observed is $pr(y \text{ is observed} \mid x) = [1 + \exp(-.5 - 0.2x)]^{-1}$. This sets about 25% of the values of each variable to be missing.

Based on empirical examination of the data, we determined the following four sequential regression imputation algorithms using different sets of reasonable regression models with varying degree of incompatibility to impute the missing values:

1. Algorithm 1 uses a possibly compatible regression model set:

$$\begin{aligned} m_{11}(y^{1/3} \mid x^{1/3}, \theta_1) &= \frac{1}{\sqrt{2\pi\sigma_{12}^2}} \exp\left\{ -\frac{(y^{1/3} - \theta_{10} - \theta_{11}x^{1/3})^2}{\sigma_{12}^2} \right\}, \\ m_{12}(x^{1/3} \mid y^{1/3}, \theta_2) &= \frac{1}{\sqrt{2\pi\sigma_{21}^2}} \exp\left\{ -\frac{(x^{1/3} - \theta_{20} - \theta_{21}y^{1/3})^2}{\sigma_{21}^2} \right\}. \end{aligned}$$

2. Algorithm 2 uses an incompatible regression model set:

$$\begin{aligned} m_{21}(y^{1/3} \mid x^{1/3}, \theta_1) &= \frac{1}{\sqrt{2\pi\sigma_{12}^2}} \exp\left\{ -\frac{(y^{1/3} - \theta_{10} - \theta_{11}x^{1/3})^2}{\sigma_{12}^2} \right\}, \\ m_{22}(x^{1/3} \mid y^{1/3}, \theta_2) &= \frac{1}{\sqrt{2\pi\sigma_{21}^2}} \exp\left\{ -\frac{(x^{1/3} - \theta_{20} - \theta_{21}y^{1/3} - \theta_{22}/y)^2}{\sigma_{21}^2} \right\}. \end{aligned}$$

3. Algorithm 3 uses the incompatible regression model set as defined in Example 6:

$$\begin{aligned} m_{31}(y \mid x, \theta_1) &= \frac{y^{\theta_{12}-1} \exp(-y(\theta_{10} + \theta_{11}x)^{-1})}{\Gamma(\theta_{12})(\theta_{10} + \theta_{11}x)^{\theta_{12}}}, \\ m_{32}(x \mid y, \theta_2) &= \frac{x^{\theta_{22}-1} \exp(-x(\theta_{20} + \theta_{21}y)^{-1})}{\Gamma(\theta_{22})(\theta_{20} + \theta_{21}y)^{\theta_{22}}}. \end{aligned}$$

4. Algorithm 4 uses a weakly incompatible regression model set:

$$\begin{aligned} m_{41}(y \mid x, \theta_1) &= \frac{y^{\theta_{12}-1} \exp(-y(\theta_{10} + \theta_{11}x)^{-1})}{\Gamma(\theta_{12})(\theta_{10} + \theta_{11}x)^{\theta_{12}}}, \\ m_{42}(x \mid y, \theta_2) &= \frac{(\theta_{20} + \theta_{21}y)^{\theta_{22}}}{\Gamma(\theta_{22})} x^{-\theta_{22}-1} \exp((\theta_{20} + \theta_{21}y)/x). \end{aligned}$$

Algorithm 1 and 2 impute the missing values on the cubic root scale, and drawn values are transformed back to the original scale at the end.

We calculated the Kullback-Leibler divergence curves for all regression models based on the complete data, or the "Before Deletion" data. The plots in Figure 1 show the divergence curves for each model from the four sets corresponding to $f_{X|Y}$ and $f_{Y|X}$. From Algorithm 1 to Algorithm 4, the model fitting is gradually improved since the divergence curve is gradually closer to 0 for both conditional densities, and both divergence curves reach 0 for Algorithm 4 as it uses a validly specified model set. In particular, the Kullback-Leibler divergence between fitted m_{32} and the true conditional density

$$\iint \log \frac{f_{X|Y}(x \mid y)}{m_{32}(x \mid y, \hat{\theta}_2^{BD})} f_{XY}(x, y) \mathrm{d}x \mathrm{d}y$$

from Algorithm 3 is uniformly close to 0 (less than 0.05 given any y); Furthermore, in the neighborhood of θ_1^*,

$$\iint \frac{\partial \log m_{31}(y \mid x, \theta_1)}{\partial \theta_1} \left[f_Y(y) m_{32}(x \mid y, \hat{\theta}_2^{BD}) - f_{XY}(x, y) \right] \mathrm{d}x \mathrm{d}y = o(1),$$

which means that fitted m_{31} (based on Y and imputed X from fitted m_{32}) is also close to the true distribution. Therefore, we regard m_{31} and m_{32} in Algorithm 3 as a well-fitting model sequence for (X, Y).

INSERT FIGURE 1 HERE.

Our primary evaluation criterion for imputation performance is the maximum of absolute difference between the empirical joint distribution based on the "Before Deletion" data and the "After Imputation" data at iteration T: $\left\| \widehat{F}_{MI}^{n,T}(x, y) - \widehat{F}_{BD}^{n}(x, y) \right\|_{\infty}$. We evaluated this conservative distance measure at T=5, 10, 20, 100, 500 and 1000 iterations. We fixed the number of imputations at 100. For each of 500 data sets, the distance measure was computed to form a function of n and T.

The averaged empirical joint distribution differences over 500 data sets from all four algorithms with different sample sizes and T=100, 500 and 1000 iterations are summarized in Figure 2. All algorithms using fewer iterations T=5, 10, 20 yielded larger differences with similar patterns, so we excluded them in the figure to achieve better visual effect. The simulation results show that as T and n increases, the empirical joint distribution difference from each algorithm stabilizes. When T and n are sufficiently large, the average (SD) of the differences between the before deletion and multiple imputation empirical distributions is 0.0288 (0.006) for Algorithm 1; 0.0257 (0.006) for Algorithm 2; 0.0188 (0.005) for Algorithm 3 and 0.0155 (0.004) for Algorithm 4. The empirical joint distribution difference decreases from Algorithm 1 to Algorithm 4, indicating that as the model fitting is improved, the performance is improved as well. Both incompatible but better fitting sets from Algorithm 2 and 3 outperform the possibly compatible set with separable marginal parameters from Algorithm 1. The simulation study suggests that the validity of the inferences depends more on the reasonableness of the model fit rather than the model compatibility.

INSERT FIGURE 2 HERE.

5 Multivariate Missing Data

The appeal of sequential regression is the ability to handle missing values in complex multivariate data structure. The sequential regression imputation approach for the p-dimensional data $Y_1, \ldots, Y_p$, assumes $m_i(y_i \mid y_{[-i]}, \theta_i)$, and $\theta_i^{(t)}$ is estimated based on $Y_{i,obs}$ and $Y_{[-i]}^{(t)}$. Although the imputation procedure is similar to bivariate algorithms, complications arise due to complex missingness patterns.

Consider the situation with three variables, (Y_1, Y_2, Y_3), with all possible item missing data patterns. The estimate $\theta_1^{(t)}$ of θ_1 in $m_1(Y_1 \mid Y_2, Y_3, \theta_1)$ is obtained by regressing the observed values of Y_1 on the corresponding subset of Y_2 and Y_3. The predictor subset consists of (Y_2, Y_3) in four missingness groups: 1. both are observed, 2 & 3: one is observed and the other is imputed, and 4: both are imputed. The predictor variables in the four groups are generally distributed differently, and then each group plays a different role in estimating $\theta_1^{(t)}$. For a data set with p variables, there are $2^p - 1$ possible missingness groups, including the complete cases Y_{cc}. It is difficult to establish results in generality given the complexity of the joint distribution of the predictors.

5.1 Single variable missingness

We consider single variable missingness pattern where there is at most one variable missing in any record. There are up to $p + 1$ missingness groups, and we denote $R = i$ for subjects with Y_i missing and $R = 0$ for the fully observed group.

During the estimation of each regression model, the subset of $Y_{[-i]}^{(t)}$ form up to p patterns, and the log-likelihood is

$$l_i(\theta_i \mid Y_{i,obs}, Y_{[-i]}^{(t)}) = l_i(\theta_i \mid Y_{cc}) + \sum_{j<i} l_i(\theta_i \mid Y_{[-j]}, Y_j^{(t)}) + \sum_{j>i} l_i(\theta_i \mid Y_{[-j]}, Y_j^{(t-1)}).$$

If there is no missingness in Y_j, $l_i(\theta_i \mid Y_{[-j]}, Y_j^{(t)})$ is absorbed into $l_i(\theta_i \mid Y_{cc})$, therefore we assume for simplicity that there is missingness in each variable.

The parameter estimate $\theta_i^{(t)}$ is obtained by solving the score equation

$$s_i(\theta_i \mid Y_{i,obs}, Y_{[-i]}^{(t)}) = s_i(\theta_i \mid Y_{cc}) + \sum_{j<i} s_i(\theta_i \mid Y_{[-j]}, Y_j^{(t)}) + \sum_{j>i} s_i(\theta_i \mid Y_{[-j]}, Y_j^{(t-1)}) = 0.$$

Based on $\theta_i^{(t)}$, $Y_{i,mis}$ is drawn from $m_i(Y_{i,mis} \mid Y_{[-i],obs}, \theta_i^{(t)})$, where $Y_{[-i],obs}$ are fully observed. We now show that in terms of convergence properties, sequential regression imputation algorithms for multivariate missing data with single variable missingness are similar to those for bivariate missing data, and conclusions in Section 3 can be extended. For weakly incompatible model sequences, the following theorem is a generalization of Theorem 1 for the bivariate case.

Theorem 3. Suppose p-dimensional data follow a joint population distribution $f(y_1, \ldots, y_p \mid \psi)$ with conditional densities $\{f_i(y_i \mid y_{[-i]}, \psi_i), i = 1, \ldots, p\}$. If the sequential regression imputation algorithm uses a weakly incompatible model sequence $\{m_i(y_i \mid y_{[-i]}, \theta_i), i = 1, \ldots, p\}$ and satisfies the regularity conditions for the differentiation/integration and the mean/variance of the score functions, then for $i = 1, \ldots, p$, $m_i(y_i \mid y_{[-i]}, \theta_i^{(t)}) \rightarrow f_i(y_i \mid y_{[-i]}, \psi_i)$, as $n,\ m,\ t \rightarrow \infty$.

Proof is given in Part 3 of Appendix 1. Also for possibly compatible models, the following is the generalization of Theorem 2.

Theorem 4. Suppose that the sequential regression imputation algorithm uses possibly compatible models $\{m_i(y_i \mid y_{[-i]}, \theta_i), i = 1, \ldots, p\}$ and satisfies the regularity conditions in Theorem 3, with $\theta \in \Theta_C$ as the subspace of $\Theta_1 \times \Theta_2 \times \ldots x\Theta_p$ where $p(y_1, \ldots, y_p \mid \theta_1, \ldots, \theta_p, \theta \in \Theta_C)$ defines the joint distribution. If the model sequence has separable marginal parameters and $(\theta_1^*, \ldots, \theta_p^*) \in \Theta_C$ is the maximum likelihood estimate of $(\theta_1, \ldots, \theta_p)$ based on the joint likelihood, then $m_i(y_i \mid y_{[-i]}, \theta_i^{(t)}) \rightarrow p(y_i \mid y_{[-i]}, \theta_i^*)$, as $n,\ m,\ t \rightarrow \infty$.

For proof, see Appendix 1, Part 4.

In practice, we can also develop well fitting model sequences regardless of model compatibility, where Kullback-Leibler divergence can be used to check the model fitting.

6 Discussion

Multiple imputation through specifications of a sequence of conditional regression models is a convenient approach for handling complex data structures with different types of variables. Several software packages have been developed to implement this approach and are being used in several substantive analyses in various disciplines. However, theoretical properties of this method have not been systematically investigated. One key question is whether using a set of incompatible conditional distributions leads to convergence or stability of the infinite imputation completed data statistics. Recently, Li, Yu and Rubin (LYR) (2012) have raised caution using some theoretical examples. However, these examples differ from the usual sequential regression setup in many ways. We address these examples in light of the results given in this paper.

Example 1 in LYR uses a deterministic set of incompatible conditional normal distributions (that is, the same parameters are used across all updating iterations) to show that different ordering of updating the iterations leads to different results. However, the sequential regression does not use deterministic set of conditionals but the parameters themselves are updated at each iteration. We conducted a simulation study with the complete data (before deletion data sets) of size $n = 7000$ on 3 variables, $Y = (Y_1, Y_2, Y_3)$, from a multivariate normal distribution with mean $(-1, 0, 1)$ and the covariance matrix $(1-\rho)I_3 + \rho J_3$ where I_3 is an identity matrix of order 3 and J_3 is a 3×3 matrix of ones. Some values were deleted such that all possible 7 patterns are represented in the after deletion data. The number of subjects in each pattern was 1000. The missing values were imputed using the following weakly incompatible models: (1) $Y_1|Y_2, Y_3 \sim N(\alpha_o + \alpha_1 Y_2 + \alpha_2 Y_3, \sigma_1^2)$; (2) $Y_2|Y_1, Y_3 \sim N(\beta_o + \beta_1 Y_1 + \beta_2 Y_3, \sigma_2^2)$; and (3) $Y_3|Y_1, Y_2 \sim N(\gamma_o + \gamma_1 Y_1 + \gamma_2 Y_2, \sigma_3^2)$. The imputations were carried out in three different orders (Y_1, Y_2, Y_3), (Y_2, Y_1, Y_3) and (Y_3, Y_1, Y_2). The number of imputations was fixed at 100 and the number of iterations considered were $T = 20, 50, 200, 500$ and 1000. Our results show that the multiple imputation estimates of the mean and the covariance matrix are unbiased for each of the three orders in which imputations were carried out. This shows that order of the imputation is irrelevant and incompatibility does not result in bias as long as each conditional model is validly specified.

Example 3 in LYR uses a grossly misspecified model sequence in the imputation for a bivariate normal data. When the imputation models are misspecified or the missing data mechanism is not ignorable, it is difficult to assess whether it is the property of the method or the effect of misspecification. Even in this case, consider the following situation: suppose the data are missing at random and the imputer uses the model $Y_1|Y_2 \sim N(\alpha_o + \alpha_1 Y_2, \sigma_1^2)$ and $Y_2|Y_1 \sim N(\beta_o + \beta_1 Y_1 + \beta_2 Y_1^2, \sigma_2^2)$ then Theorem 1 applies and the sequential regression imputation algorithm results in the consistent estimator of the joint distribution of (Y_1, Y_2). The key, therefore, is not to fix the parameters across the iterations but revise the estimates based on updating of the imputed values. Thus asymptotically, the observed data tend to pull towards the consistent model when the joint distribution is embedded in the parameter space of the conditional distributions. Thus, our investigations suggest that the sequential regression approach may yield valid results if the conditional distributions fit the data well even though they may not be compatible with any joint distribution.

There are number of limitations in this study. The investigation was restricted to a missing data pattern with any subject missing values on at most one variable. This was mainly to restrict the number of missing data patterns to a manageable number. Further investigations are necessary to assure that the algorithm will converge and provide valid results for more complex missing data pattern. We have performed a limited simulation study to consider more complex pattern of missing data where well fitting incompatible models were used to impute the missing values. The multiple imputation inferences had desirable repeated sampling properties even in this situation. However, establishing the exact conditions for convergence seems to be more complicated and further research is necessary. On the contrary, using a poorly fitting but compatible model sequence led to inferences with undesirable properties. Even this simulation study suggests that an imputer has to choose the models carefully to ensure that each conditional model fits the data well.

Appendix 1

Part 1: Proof of Theorem 1

Proof: Since weakly incompatible model sequences include two cases, we first prove the theorem for exactly specified sequences, and assume that the functional forms of $m_1(y \mid x, \theta_1)$ and $m_2(x \mid y, \theta_2)$ correspond to the true conditional densities $f_{X|Y}(x \mid y, \psi_1)$ and $f_{Y|X}(y \mid x, \psi_2)$ respectively. The asymptotic score functions defining the iterative algorithm can be rewritten as below:

$$\tilde{s}_1(\theta_1 \mid {\theta_2}^{(t-1)}) = \iint \frac{\partial \log m_1(y \mid x, \theta_1)}{\partial \theta_1} \Big\{ n_0 f_{XY}(x, y \mid R=0) + n_1 m_2(x \mid y, {\theta_2}^{(t-1)}) f_Y(y \mid R=1) \Big\} \mathrm{d}x\mathrm{d}y$$

$$\tilde{s}_2(\theta_2 \mid \theta_1^{(t)}) = \iint \frac{\partial \log m_2(x \mid y, \theta_2)}{\partial \theta_2} \Big\{ n_0 f_{XY}(x, y \mid R=0) + n_2 m_1(y \mid x, {\theta_1}^{(t)}) f_X(x \mid R=2) \Big\} \mathrm{d}x\mathrm{d}y.$$

Since the missingness is ignorable, we have

$$\begin{aligned} & n_0 f_{XY}(x, y \mid R=0) + n_1 m_2(x \mid y, \psi_2) f_Y(y \mid R=1) \\ = \;& (n_0 + n_1) f_{XY}(x, y \mid R \neq 2) = (n_0 + n1) m_1(y \mid x, \psi_1) f_X(x \mid R \neq 2) \end{aligned}$$

and

$$\begin{aligned} & n_0 f_{XY}(x, y \mid R=0) + n_2 m_1(y \mid x, \psi_1) f_X(x \mid R=2) \\ = \;& (n_0 + n_2) f_{XY}(x, y \mid R \neq 1) = (n_0 + n_2) m_2(x \mid y, \psi_2) f_Y(y \mid R \neq 1). \end{aligned}$$

It is then easy to show that (ψ_1, ψ_2) satisfies the asymptotic score equations

$$\begin{aligned} \tilde{s}_1(\psi_1 \mid \psi_2) &= (n_0 + n_1) \iint \frac{\partial \log m_1(y \mid x, \theta_1)}{\partial \theta_1} m_1(y \mid x, \psi_1) f_X(x \mid R \neq 2) \mathrm{d}x\mathrm{d}y \Big|_{\theta_1 = \psi_1} = 0; \\ \tilde{s}_2(\psi_2 \mid \psi_1) &= (n_0 + n_2) \iint \frac{\partial \log m_2(x \mid y, \theta_2)}{\partial \theta_2} m_2(x \mid y, \psi_2) f_Y(y \mid R \neq 1) \mathrm{d}x\mathrm{d}y \Big|_{\theta_2 = \psi_2} = 0. \end{aligned}$$

Therefore, as n, $t \to \infty$, $(\theta_1^{(t)}, \theta_2^{(t)}) \to (\psi_1, \psi_2)$, which leads to that $m_1(y \mid x, \theta_1^{(t)}) \to f_{Y|X}(y \mid x, \psi_1)$ and $m_2(x \mid y, \theta_2^{(t)}) \to f_{X|Y}(x \mid y, \psi_2)$.

We now prove the theorem for validly specified model sequences with extra terms compared to the true conditional densities. Suppose that without loss of any generality we introduce a parameterization $\theta_1 = (\zeta_1, \xi_1)$ and $\theta_2 = (\zeta_2, \xi_2)$ such that $m_1(y \mid x, \zeta_1 = \psi_1, \xi_1 = 0) = f_{Y|X}(y \mid x, \psi_1)$ and $m_2(x \mid y, \zeta_2 = \psi_2, \xi_2 = 0) = f_{X|Y}(x \mid y, \psi_2)$. We need to show that $\theta_1^* = (\psi_1, 0)$ and $\theta_2^* = (\psi_2, 0)$ are the convergent point of the asymptotic iterative algorithm.

Given $\theta_2^* = (\psi_2, 0)$,

$$\tilde{s}_1(\theta_1 \mid \theta_2^*) = (n_0 + n_1) \iint \frac{\partial \log m_1(y \mid x, \theta_1)}{\partial \theta_1} f_{Y|X}(y \mid x, \psi_1) f_X(x \mid R \neq 2) \mathrm{d}x\mathrm{d}y.$$

Since maximizing the likelihood is equivalent to minimizing the relative entropy of the regression model regarding the true distribution, to find the solution to $\tilde{s}_1(\theta_1 \mid \theta_2^*) = 0$ is equivalent to minimize $\iint \log[f_{Y|X}(y \mid x, \psi_1)/m_1(y \mid x, \theta_1)] f_{Y|X}(y \mid x, \psi_1) f_X(x \mid R \neq 2)\mathrm{d}x\mathrm{d}y$. Since the relative entropy has non-negative values and its minimum 0 is reached if and only if $m_1(y \mid x, \theta_1 = (\psi_1, 0)) = f_{Y|X}(y \mid x, \psi_1)$. Therefore, the asymptotic score equation $\tilde{s}_1(\theta_1 \mid \theta_2) = 0$ holds at (θ_1^*, θ_2^*). The similar arguments apply to m_2, and we also have $\tilde{s}_2(\theta_2^* \mid \theta_1^*) = 0$.

Part 2: Proof of Theorem 2

Proof: To determine the target to which the approximate algorithm converges, we first apply the joint model $m_{XY}(x, y \mid \theta_{XY})$ to analyze the incomplete data, where $\theta_{XY} = (\theta_1, \theta_2, \theta \in \Theta_C)$. Since the joint model has separable marginal parameters, suppose that without loss of generality we have two parameterizations $\theta_{XY} = (\theta_1, \theta_X)$ and $\theta_{XY} = (\theta_2, \theta_Y)$ for the joint model. We use Expectation-Maximization algorithm to obtain the maximum likelihood estimate $\theta_{XY}^* = (\theta_1^*, \theta_2^*)$. The expectation step calculates

$$Q(\theta_{XY} \mid \theta_{XY}^{(t-1)}) = \sum_i \log m_{XY}(x_{0i}, y_{0i} \mid \theta_{XY}) + \sum_j \int \log m_{XY}(x_{1j}, y_{1j} \mid \theta_{XY}) m_2(x \mid y_{1j}, \theta_{XY}^{(t-1)})\mathrm{d}x$$
$$+ \sum_k \int \log m_{XY}(x_{2k}, y_{2k} \mid \theta_{XY}) m_1(y \mid x_{2k}, \theta_{XY}^{(t-1)})\mathrm{d}y,$$

and the maximization step finds the parameter which maximizes the expected log-likelihood:

$$\theta_{XY}^{(t)} = \arg\max\nolimits_{\theta_{XY}} Q(\theta_{XY} \mid \theta_{XY}^{(t-1)}).$$

The expected step can be approximated by an asymptotic quantity

$$\begin{aligned}
\tilde{Q}(\theta_{XY} \mid \theta_{XY}^{(t-1)}) = & \lim_{n\to\infty} Q(\theta_{XY} \mid \theta_{XY}^{(t-1)}) \\
= & \; n_0 \iint \log m_{XY}(x, y \mid \theta_{XY}) f_{XY}(x, y \mid R = 0)\mathrm{d}x\mathrm{d}y \\
& + n_1 \iint \log m_{XY}(x, y \mid \theta_{XY}) m_2(x \mid y, \theta_{XY}^{(t-1)})\mathrm{d}x f_Y(y \mid R = 1)\mathrm{d}y \\
& + n_2 \iint \log m_{XY}(x, y \mid \theta_{XY}) m_1(y \mid x, \theta_{XY}^{(t-1)})\mathrm{d}y f_X(x \mid R = 2)\mathrm{d}x,
\end{aligned}$$

and the maximization step maximizes the asymptotic quantity.

Since ${\theta_{XY}}^*$ is the convergent point to the asymptotic Expectation-Maximization algorithm, for both parameterizations, score equations hold at the convergent point because the marginal parameters are separable:

$$\left.\frac{\partial \tilde{Q}(\theta_1, {\theta_X}^* \mid {\theta_{XY}}^*)}{\partial \theta_1}\right|_{{\theta_1}^*} = 0, \quad \left.\frac{\partial \tilde{Q}({\theta_1}^*, \theta_X \mid {\theta_{XY}}^*)}{\partial \theta_X}\right|_{{\theta_X}^*} = 0;$$

and

$$\left.\frac{\partial \tilde{Q}(\theta_2, {\theta_y}^* \mid {\theta_{XY}}^*)}{\partial \theta_2}\right|_{{\theta_2}^*} = 0, \ \left.\frac{\partial \tilde{Q}({\theta_2}^*, \theta_y \mid {\theta_{XY}}^*)}{\partial \theta_y}\right|_{{\theta_Y}^*} = 0.$$

We now show that the maximum likelihood estimate ${\theta_{XY}}^*$ is also the fixed point of the asymptotic sequential regression imputation algorithm by demonstrating

$$\begin{aligned} \tilde{s}_1({\theta_1}^* \mid {\theta_2}^*) &= 0, \\ \tilde{s}_2({\theta_1}^* \mid {\theta_2}^*) &= 0. \end{aligned}$$

From the Expectation-Maximization algorithm, we assume that the probability functions are absolute continuous, and we interchange the differential and integral sign. Then

$$\begin{aligned}
& \frac{\partial \tilde{Q}(\theta_1, {\theta_X}^* \mid {\theta_{XY}}^*)}{\partial \theta_1} \\
= \; & n_0 \iint \frac{\partial \log m_{XY}(x, y \mid \theta_1, \theta_X^*)}{\partial \theta_1} f_{XY}(x, y \mid R=0)\mathrm{d}x\mathrm{d}y \\
& + n_1 \iint \frac{\partial \log m_{XY}(x, y \mid \theta_1, \theta_X^*)}{\partial \theta_1} m_2(x \mid y, {\theta_2}^*)\mathrm{d}x f_Y(y \mid R=1)\mathrm{d}y \\
& + n_2 \iint \frac{\partial \log m_{XY}(x, y \mid \theta_1, \theta_X^*)}{\partial \theta_1} m_1(y \mid x, \theta_1)\mathrm{d}y f_X(x \mid R=2)\mathrm{d}x \\
= \; & n_0 \iint \left(\frac{\partial \log m_1(y \mid x, \theta_1)}{\partial \theta_1} + \frac{\partial \log m_X(x, \theta_X^*)}{\partial \theta_1}\right) f_{XY}(x, y \mid R=0)\mathrm{d}x\mathrm{d}y \\
& + n_1 \iint \left(\frac{\partial \log m_1(y \mid x, \theta_1)}{\partial \theta_1} + \frac{\partial \log m_X(x, \theta_X^*)}{\partial \theta_1}\right) m_2(x \mid y, {\theta_2}^*)\mathrm{d}x f_Y(y \mid R=1)\mathrm{d}y \\
& + n_2 \iint \left(\frac{\partial \log m_1(y \mid x, \theta_1)}{\partial \theta_1} + \frac{\partial \log m_X(x, \theta_X^*)}{\partial \theta_1}\right) m_1(y \mid x, \theta_1)\mathrm{d}y f_X(x \mid R=2)\mathrm{d}x \\
= \; & n_0 \iint \frac{\partial \log m_1(y \mid x, \theta_1)}{\partial \theta_1} f_{XY}(x, y \mid R=0)\mathrm{d}x\mathrm{d}y \\
& + n_1 \iint \frac{\partial \log m_1(y \mid x, \theta_1)}{\partial \theta_1} m_2(x \mid y, {\theta_2}^*)\mathrm{d}x f_Y(y \mid R=1)\mathrm{d}y \\
& + n_2 \iint \frac{\partial \log m_1(y \mid x, \theta_1)}{\partial \theta_1} m_1(y \mid x, \theta_1)\mathrm{d}y f_X(x \mid R=2)\mathrm{d}x \\
= \; & \tilde{s}_1(\theta_1 \mid {\theta_2}^*) + n_2 \iint \frac{\partial \log m_1(y \mid x, \theta_1)}{\partial \theta_1} m_1(y \mid x, \theta_1)\mathrm{d}y f_X(x \mid R=2)\mathrm{d}x.
\end{aligned}$$

Then the asymptotic score equations holds at ${\theta_{XY}}^*$:

$$\begin{aligned}
& \tilde{s}_1({\theta_1}^* \mid {\theta_2}^*) \\
= \; & \left.\frac{\partial \tilde{Q}(\theta_1, {\theta_X}^* \mid {\theta_{XY}}^*)}{\partial \theta_1}\right|_{{\theta_1}^*} - n_2 \iint \frac{\partial \log m_1(y \mid x, \theta_1)}{\partial \theta_1} m_1(y \mid x, \theta_1)\mathrm{d}y f_X(x \mid R=2)\mathrm{d}x\bigg|_{{\theta_1}^*} \\
= \; & 0.
\end{aligned}$$

Similarly, $\tilde{s}_2({\theta_2}^* \mid {\theta_1}^*) = 0$ can also be obtained.

Part 3: Proof of Theorem 3

Proof: As in Theorem 1, we first prove the theorem for exactly specified model sequences, where $m_i(y_i \mid y_{[-i]}, \theta_i = \psi_i) = f_i(y_i \mid y_{[-i]}, \psi_i)$ for $i = 1, \ldots, p$. We need to show that for each regression model, the asymptotic score equation $\tilde{s}_i(\psi_i \mid \psi_{[-i]})$ holds.

Denoting $u_i(y_i \mid y_{[-i]}) = \partial \log(m_i(y_i \mid y_{[-i]}, \theta_i))/\partial\theta_i$ and n_i the sample size of each missingness group, then the asymptotic function for the ith model given $\theta^*_{[-i]} = \psi_{[-i]}$ is as

$$\tilde{s}_i(\theta_i \mid \psi_{[-i]}) = \tilde{s}_i(\theta_i \mid R = 0) + \sum_{j \neq i} \tilde{s}_i(\theta_i \mid \psi_j, R = j),$$

where

$$\tilde{s}_i(\theta_i \mid R = 0) = \int \cdots \int u_i(y_i \mid y_{[-i]}) n_0 f(y_1, \ldots, y_p \mid R = 0) \mathrm{d}y_1 \cdots \mathrm{d}y_p,$$

and for $j \neq i$,

$$\begin{aligned} \tilde{s}_i(\theta_i \mid \psi_j, R = j) &= \int \cdots \int u_i(y_i \mid y_{[-i]}) n_j m_j(y_j \mid y_{[-j]}, \psi_j) f(y_{[-j]} \mid R = j) \mathrm{d}y_1 \cdots \mathrm{d}y_p \\ &= \int \cdots \int u_i(y_i \mid y_{[-i]}) n_j f(y_1, \ldots, y_p \mid R = j) \mathrm{d}y_1 \cdots \mathrm{d}y_p. \end{aligned}$$

Since the missingness is ignorable, we have

$$\sum_{j=0, j\neq i}^{p} n_j f(y_1, \ldots, y_p \mid R = j) = (n - n_i) f(y_1, \ldots, y_p \mid R \neq i) = (n - n_i) m_i(y_i \mid y_{[-i]}, \psi_i) f(y_{[-i]} \mid R \neq i),$$

and the asymptotic function can be rewritten as below:

$$\tilde{s}_i(\theta_i \mid \psi_{[-i]}) = (n - n_i) \int \cdots \int u_i(y_i \mid y_{[-i]}) m_i(y_i \mid y_{[-i]}, \psi_i) f(y_{[-i]} \mid R \neq i) \mathrm{d}y_1 \cdots \mathrm{d}y_p,$$

and it is easy to show that the asymptotic score equation holds:

$$\tilde{s}_i(\psi_i \mid \psi_{[-i]}) = (n - n_i) \int \cdots \int u_i(y_i \mid y_{[-i]}) m_i(y_i \mid y_{[-i]}, \psi_i) f(y_{[-i]} \mid R \neq i) \mathrm{d}y_1 \cdots \mathrm{d}y_p \bigg|_{\theta_i = \psi_i} = 0.$$

Therefore, as n, $t \to \infty$, $\theta_i^{(t)} \to \psi_i$, which leads to that $m_i(y_i \mid y_{[-i]}, \theta_i^{(t)}) \to f_i(y_i \mid y_{[-i]}, \psi_i)$, for $i = 1, \ldots, p$.

We now prove the theorem for validly specified model sequences with extra terms compared to the true conditional densities. As in Theorem 2, suppose that without loss of any generality we introduce a parameterization $\theta_i = (\zeta_i, \xi_i)$ such that $m_i(y_i \mid y_{[-i]}, \zeta_i = \psi_i, \xi_i =$

$0) = f_i(y_i \mid y_{[-i]}, \psi_i)$. We need to show that $\{\theta_i^* = (\psi_i, 0), i = 1, \ldots, p\}$ is the convergent point of the asymptotic iterative algorithm.

Given that $\theta_j^* = (\psi_j, 0)$ for $j \neq i$,

$$\tilde{s}_i(\theta_i \mid \theta^*_{[-i]}) = (n - n_i) \int \cdots \int u_i(y_i \mid y_{[-i]}) f_i(y_i \mid y_{[-i]}, \psi_i) f(y_{[-i]} \mid R \neq i) \mathrm{d}y_1 \cdots \mathrm{d}y_p.$$

As in Theorem 2, to find the solution to $\tilde{s}_i(\theta_i \mid \theta^*_{[-i]}) = 0$ is equivalent to minimize $\int \cdots \int \log[f_i(y_i \mid y_{[-i]}, \psi_i)/m_i(y_i \mid y_{[-i]}, \theta_i)] f_i(y_i \mid y_{[-i]}, \psi_i) f(y_{[-i]} \mid R \neq i) \mathrm{d}y_1 \cdots \mathrm{d}y_p$. Since the relative entropy has non-negative values and its minimum 0 is reached if and only if $m_i(y_i \mid y_{[-i]}, \theta_i = (\psi_i, 0)) = f_i(y_i \mid y_{[-i]}, \psi_i)$. Therefore, the asymptotic score equation $\tilde{s}_i(\theta_i \mid \theta_{[-i]}) = 0$ holds at $(\theta_1^*, \ldots, \theta_p^*)$ for any $i = 1, \ldots, p$.

Part 4: Proof of Theorem 4

Proof: As in Theorem 2, we first apply the joint model $p(y_1, \ldots, y_p \mid \theta)$ to analyze the incomplete data to determine the convergent point. Denote $\tilde{Q}(\theta \mid \theta^{(t-1)})$ the expected log-likelihood from the Expectation-Maximization algorithm, where $\theta = (\theta_1, \ldots, \theta_p) \in \Theta_C$. For $i = 1, \ldots, p$, denote $\theta_{M,[-i]}$ the parameter of the marginal joint model of $Y_{[-i]}$ from the joint model, then because of the separability of marginal parameters, $\theta_{M,[-i]}$ is distinctive from θ_i and $\theta = (\theta_i, \theta_{M,[-i]})$ is a parameterization of the joint model. Since $\theta^* = (\theta_1^*, \ldots, \theta_p^*)$ is the maximum likelihood estimate, $\partial \tilde{Q}(\theta \mid \theta^*)/\partial\theta|_{\theta=\theta^*} = 0$. On the other hand, because marginal separability ensures that

$$\partial \log p(y_1, \ldots, y_p \mid \theta_i, \theta_{M,[-i]})/\partial\theta_i = \partial(\log(m_i(y_i \mid y_{[-i]}, \theta_i)))/\partial\theta_i = u_i(y_i \mid y_{[-i]}),$$

we have

$$\begin{aligned}
\partial \tilde{Q}(\theta_i, \theta^*_{M,[-i]} \mid \theta^*)/\partial\theta_i \quad &= \int \cdots \int u_i(y_i \mid y_{[-i]}) n_0 f(y_1, \ldots, y_p \mid R = 0) \mathrm{d}y_1 \cdots \mathrm{d}y_p \\
&+ \int \cdots \int u_i(y_i \mid y_{[-i]}) \sum_{j=1}^{p} \left[n_j m_j(y_j \mid y_{[-j]}, \theta_j^*) f(y_{[-j]} \mid R = j) \right] \mathrm{d}y_1 \cdots \mathrm{d}y_p \\
&= \tilde{s}_i(\theta_i \mid \theta^*_{[-i]}) \\
&+ n_i \int \cdots \int u_i(y_i \mid y_{[-i]}) m_i(y_i \mid y_{[-i]}, \theta_i^*) f(y_{[-i]} \mid R = i) \mathrm{d}y_1 \cdots \mathrm{d}y_p.
\end{aligned}$$

Therefore, the asymptotic score equations hold at θ^* as for all $i = 1, \ldots, p$,

$$\begin{aligned}
& \tilde{s}_i(\theta_i^* \mid \theta^*_{[-i]}) \\
= \quad & \left[\partial \tilde{Q}(\theta_i, \theta^*_{M,[-i]} \mid \theta^*)/\partial\theta_i - n_i \int \cdots \int u_i(y_i \mid y_{[-i]}) m_i(y_i \mid y_{[-i]}, \theta_i^*) f(y_{[-i]} \mid R = i) \mathrm{d}y_1 \cdots \mathrm{d}y_p \right] \Bigg|_{\theta_i^*} \\
= \quad & 0.
\end{aligned}$$

Therefore, as $n, t \to \infty$, $\theta_i^{(t)} \to \theta_i^*$, which leads to that $m_i(y_i \mid y_{[-i]}, \theta_i^{(t)}) \to p_i(y_i \mid y_{[-i]}, \theta_i^*)$, for $i = 1, \ldots, p$.

Appendix 2

Examples

Example 1 (Two Linear Regression Models revisited): Suppose the data follow a bivariate normal distribution $(x, y)^T \sim N(\mu, \Sigma)$, where $\mu = (\mu_x, \mu_y)^T$ and

$$\Sigma = \begin{pmatrix} \tau_x^2 & \rho\tau_x\tau_y \\ \rho\tau_x\tau_y & \tau_y^2 \end{pmatrix}$$

and its conditional distributions are

$$\begin{aligned} y \mid x & \sim N(\alpha_{10} + \alpha_{11}x, \tau_{12}^2), \\ x \mid y & \sim N(\alpha_{20} + \alpha_{21}y, \tau_{21}^2) \end{aligned}$$

with $\alpha_{11}/\tau_{12}^2 = \alpha_{21}/\tau_{21}^2$. The missing data mechanism is assumed to be missing completely at random: $\pi_0 = Pr$(both X and Y are observed), $\pi_1 = Pr$(Y is observed and X is missing) and $\pi_2 = Pr$(X is observed and Y is missing).

The estimated regression parameters converge to $\theta_1^* = (\alpha_{10}, \alpha_{11}, \tau_{12}^2)^T$, and $\theta_2^* = (\alpha_{20}, \alpha_{21}, \tau_{21}^2)^T$. Based on the approximate iterative algorithm

$$\begin{aligned} \left(\theta_{11}^{(t)}, \theta_{10}^{(t)}, \sigma_{12}^{2\ (t)}\right)^T &= \tilde{s}_1^{-1}\left(\theta_{21}^{(t-1)}, \theta_{20}^{(t-1)}, \sigma_{21}^{2\ (t-1)}\right), \\ \left(\theta_{21}^{(t)}, \theta_{20}^{(t)}, \sigma_{21}^{2\ (t)}\right)^T &= \tilde{s}_2^{-1}\left(\theta_{11}^{(t)}, \theta_{10}^{(t)}, \sigma_{12}^{2\ (t)}\right), \end{aligned}$$

the Jacobian matrices $D\tilde{s}_1^{-1}$ and $D\tilde{s}_2^{-1}$ are calculated as follows:

$$\begin{aligned} D\tilde{s}_1^{-1}(\theta_2^*) &= r_1 \begin{pmatrix} \frac{\alpha_{11}}{\alpha_{21}} - 2\alpha_{11}^2 & 0 & -\frac{\alpha_{11}}{\tau_x^2} \\ -\frac{\alpha_{11}}{\alpha_{21}}\mu_x + 2\alpha_{11}^2\mu_x - \alpha_{11}\mu_y & -\alpha_{11} & -\frac{\alpha_{11}}{\tau_x^2}\mu_x \\ -2(1-\alpha_{11}\alpha_{21})\alpha_{11}\tau_y^2 & 0 & \alpha_{11}^2 \end{pmatrix}, \\ D\tilde{s}_2^{-1}(\theta_1^*) &= r_2 \begin{pmatrix} \frac{\alpha_{21}}{\alpha_{11}} - 2\alpha_{21}^2 & 0 & -\frac{\alpha_{21}}{\tau_y^2} \\ -\frac{\alpha_{21}}{\alpha_{11}}\mu_y + 2\alpha_{21}^2\mu_y - \alpha_{21}\mu_x & -\alpha_{21} & -\frac{\alpha_{21}}{\tau_y^2}\mu_y \\ -2(1-\alpha_{21}\alpha_{11})\alpha_{21}\tau_x^2 & 0 & \alpha_{21}^2 \end{pmatrix}, \end{aligned}$$

where $r_1 = \pi_1(\pi_0 + \pi_1)^{-1}$ and $r_2 = \pi_2(\pi_0 + \pi_2)^{-1}$.

The eigenvalues of $r_1^{-1} r_2^{-1} D\tilde{s}_1^{-1} \times D\tilde{s}_2^{-1}$ can be solved by the following characteristic equation:

$$(\lambda - 1)\left(\lambda - \rho^2\right)\left(\lambda - \rho^4\right) = 0.$$

The eigenvalues of the rate matrix $D\tilde{s}^{-1}(\theta^*)$ are $r_1 r_2$, $r_1 r_2 \rho^2$ and $r_1 r_2 \rho^4$. Since $0 \leq \rho^2 \leq 1$ holds for any bivariate normal data, the largest eigen value of $D\tilde{s}^{-1}(\theta^*)$ is $\pi_1\pi_2(\pi_0 + \pi_1)^{-1}(\pi_0 + \pi_2)^{-1}$, which is the global rate of convergence for the iterative algorithm.

- (Two Logistic Regression Models revisited) Suppose the data (X, Y) come from a bivariate Bernoulli distribution with $pr(X = 0, Y = 0) = p_{00}$, $pr(X = 0, Y = 1) = p_{01}$, $pr(X = 1, Y = 0) = p_{10}$ and $pr(X = 1, Y = 1) = p_{11} = 1 - p_{00} - p_{01} - p_{10}$, where the corresponding conditional distributions are

$$\begin{aligned} y \mid x &\sim Bernoulli\left[(1 + \exp(-\alpha_{10} - \gamma_{12}x))^{-1}\right], \\ x \mid y &\sim Bernoulli\left[(1 + \exp(-\alpha_{20} - \gamma_{21}y))^{-1}\right]. \end{aligned}$$

The parameters from the conditional distributions satisfy the compatibility condition $\gamma_{12} = \gamma_{21}$. The missing data mechanism is assumed to be missing completely at random: $\pi_0 = Pr(\text{both } X \text{ and } Y \text{ are observed})$, $\pi_1 = Pr(Y \text{ is observed and } X \text{ is missing})$ and $\pi_2 = Pr(X \text{ is observed and } Y \text{ is missing})$.

The estimated regression parameters converge to $\theta_1^* = (\alpha_{10}, \gamma_{12})^T$, and $\theta_2^* = (\alpha_{20}, \gamma_{21})^T$. Based on the approximate iterative algorithm

$$\begin{aligned} \left(\theta_{12}^{(t)}, \theta_{10}^{(t)}\right)^T &= \tilde{s}_1^{-1}\left(\theta_{21}^{(t-1)}, \theta_{20}^{(t-1)}\right), \\ \left(\theta_{21}^{(t)}, \theta_{20}^{(t)}\right)^T &= \tilde{s}_2^{-1}\left(\theta_{12}^{(t)}, \theta_{10}^{(t)}\right), \end{aligned}$$

the Jacobian matrices $D\tilde{s}_1^{-1}$ and $D\tilde{s}_2^{-1}$ are calculated as follows:

$$\begin{aligned} D\tilde{s}_1^{-1}\left(\theta_2^*\right) &= \frac{\pi_1}{\pi_0 + \pi_1}\begin{pmatrix} 1 & 0 \\ -\frac{p_{11}}{p_{01}+p_{11}} & \frac{p_{00}}{p_{10}+p_{00}} - \frac{p_{11}}{p_{01}+p_{11}} \end{pmatrix}, \\ D\tilde{s}_2^{-1}\left(\theta_1^*\right) &= \frac{\pi_2}{\pi_0 + \pi_2}\begin{pmatrix} 1 & 0 \\ -\frac{p_{11}}{p_{10}+p_{11}} & \frac{p_{00}}{p_{00}+p_{01}} - \frac{p_{11}}{p_{10}+p_{11}} \end{pmatrix}. \end{aligned}$$

The eigenvalues of $D\tilde{s}_1^{-1} \times D\tilde{s}_2^{-1}$ are $\pi_1\pi_2(\pi_0+\pi_1)^{-1}(\pi_0+\pi_2)^{-1}$ and

$$\frac{\pi_1}{\pi_0+\pi_1}\frac{\pi_2}{\pi_0+\pi_2}\left[\frac{p_{00}}{p_{10}+p_{00}}-\frac{p_{11}}{p_{01}+p_{11}}\right]\left[\frac{p_{00}}{p_{00}+p_{01}}-\frac{p_{11}}{p_{10}+p_{11}}\right].$$

Then the eigenvalue of $D\tilde{s}^{-1}(\theta^*)$ with largest absolute value is $\pi_1\pi_2(\pi_0+\pi_1)^{-1}(\pi_0+\pi_2)^{-1}$, which is the asymptotic global rate of convergence for the iterative algorithm.

Example 2 (Conditional Exponential): Suppose (X, Y) are two positive continuous variables, and two imputation models are

$$\begin{aligned} m_1 : y \mid x, \theta_1 &\sim Exp\left(\theta_1 x\right), \\ m_2 : x \mid y, \theta_2 &\sim Exp\left(\theta_2 y\right). \end{aligned}$$

At iteration t, we apply m_1 on $\{(x_{0i}, y_{0i}), (x_{1j}^{(t-1)}, y_{1j})\}$ to estimate $\theta_1^{(t)}$ by maximizing the following likelihood:

$$L_1^{(t)}(\theta_1 \mid (x_{0i}, y_{0i}), (x_{1j}^{(t-1)}, y_{1j})) \propto \theta_1^{n_0+n_1}\exp\{-\theta_1(\sum_i x_{0i}y_{0i} + \sum_j x_{1j}^{(t-1)}y_{1j})\}.$$

By solving the score equation $\partial lnL_1^{(t)}/\partial\theta_1 = 0$, we have

$$\theta_1^{(t)} = \frac{n_0+n_1}{\sum_i x_{0i}y_{0i} + \sum_j x_{1j}^{(t-1)}y_{1j}}. \tag{1}$$

Then the missing values of Y are drawn from $y_{2k}^{(t)} \mid x_{2k} \sim \theta_1^{(t)}x_{2k}\exp\{-\theta_1^{(t)}x_{2k}y\}$. Since $x_{1j}^{(t-1)} \mid y_{1j} \sim Exp(\theta_2^{(t-1)}y_{1j})$, by central limit theory we have $p\lim\sum_j x_{1j}^{(t-1)}y_{1j} = n_1/\theta_2^{(t-1)}$. We also denote $E_0 = E[x_{0i}y_{0i}] = E[XY \mid R = 0]$, and then $p\lim\sum_i x_{0i}y_{0i} = n_0E_0$. Therefore, the following relation holds by applying the asymptotic approximations on Eq. (1):

$$\theta_1^{(t)} \approx \frac{n_0+n_1}{n_0E_0 + n_1/\theta_2^{(t-1)}}. \tag{2}$$

Similarly, model 2 is applied on $\{(x_{0i}, y_{0i}), (x_{2k}, y_{2k}^{(t)})\}$ to obtain the parameter estimation

$$\theta_2^{(t)} = \frac{n_0+n_2}{\sum_i x_{0i}y_{0i} + \sum_k x_{2k}y_{2k}^{(t)}} \tag{3}$$

and the approximate relation

$$\theta_2^{(t)} \approx \frac{n_0+n_2}{n_0E_0 + n_2/\theta_1^{(t)}}. \tag{4}$$

Eq. (2) and (4) define an approximate iterative relation between θ_1 and θ_2. The fixed point to this approximate algorithm can be calculated by solving the two equations:

$$\theta_1^* = \theta_2^* = \frac{1}{E_0}. \tag{5}$$

By taking Taylor expansion of Eq. (2) and (4) at the fixed point, we have

$$\begin{aligned} \theta_1^{(t)} - \theta_1^* &\approx \frac{\theta_1^{*2}}{n_0 + n_1} \frac{n_1}{\theta_2^{*2}} (\theta_2^{(t-1)} - \theta_2^*), \\ \theta_2^{(t-1)} - \theta_2^* &\approx \frac{\theta_2^{*2}}{n_0 + n_2} \frac{n_2}{\theta_1^{*2}} (\theta_1^{(t-1)} - \theta_1^*). \end{aligned}$$

Therefore, the parameters around the fixed point are updated approximately according to

$$\theta_1^{(t)} - \theta_1^* \approx \frac{n_1}{n_0 + n_1} \frac{n_2}{n_0 + n_2} (\theta_1^{(t-1)} - \theta_1^*), \tag{6}$$

$$\theta_2^{(t)} - \theta_2^* \approx \frac{n_1}{n_0 + n_1} \frac{n_2}{n_0 + n_2} (\theta_2^{(t-1)} - \theta_2^*). \tag{7}$$

The derivative $n_1 n_2 (n_0 + n_1)^{-1} (n_0 + n_2)^{-1}$ is the rate of convergence for the approximate iterative algorithm.

References

[1] B. C. Arnold, E. Castillo, and J. M. Sarabia. Conditionally specified distributions: an introduction (with discussion). *Statistical Science*, 16(3):249–274, 2001.

[2] B. C. Arnold and S. J. Press. Compatible conditional distributions. *Journal of the American Statistical Association*, 84:152–156, 1989.

[3] B. C. Arnold and D. J. Strauss. Bivariate distributions with conditionals in prescribed exponential families. *Journal of the Royal Statistical Society B*, 53(2):365–375, 1991.

[4] A. Gelman and T. E. Raghunathan. Discussion of arnold et al. conditionally specified distributions. *Statistical Science*, 16(3):249–274, 2001.

[5] A. Gelman and T. P. Speed. Characterizing a joint probability distribution by conditionals. *Journal of the Royal Statistical Society B*, 55(1):185–188, 1993.

[6] A. B. Kennickell. Imputation of the 1989 survey of consumer finances: Stochastic relaxation and multiple imputation. *Proc. Sec. Surv. Res. Meth., Am. Statist. Assoc.*, pages 1–10, 1991.

[7] F. Li, Y. Yu, and D. B. Rubin. Imputing missing data by fully conditional models: some cautionary examples and guidelines. 2012.

[8] T. E. Raghunathan, J. M. Lepkowski, J. VanHoewyk, and P. Solenberger. A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, 27:85–95, 2001.

[9] P. Royston. Multiple imputation of missing values: Update. *Stata Journal*, 5(2):188–201, 2005.

[10] D. B. Rubin. Inference and missing data. *Biometrika*, 63:581–590, 1976.

[11] S. van Buuren, J. P. L. Brand, C. G. M. Groothuis-Oudshoorn, and D. B. Rubin. Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, 76(12):1049–1064, 2006.

[12] S van Buuren and C. Oudshoorn. Flexible multivariate imputation by mice. Technical report, TNO-rapport PG 99.054. TNO Prevention and Health, Leiden., 1999.

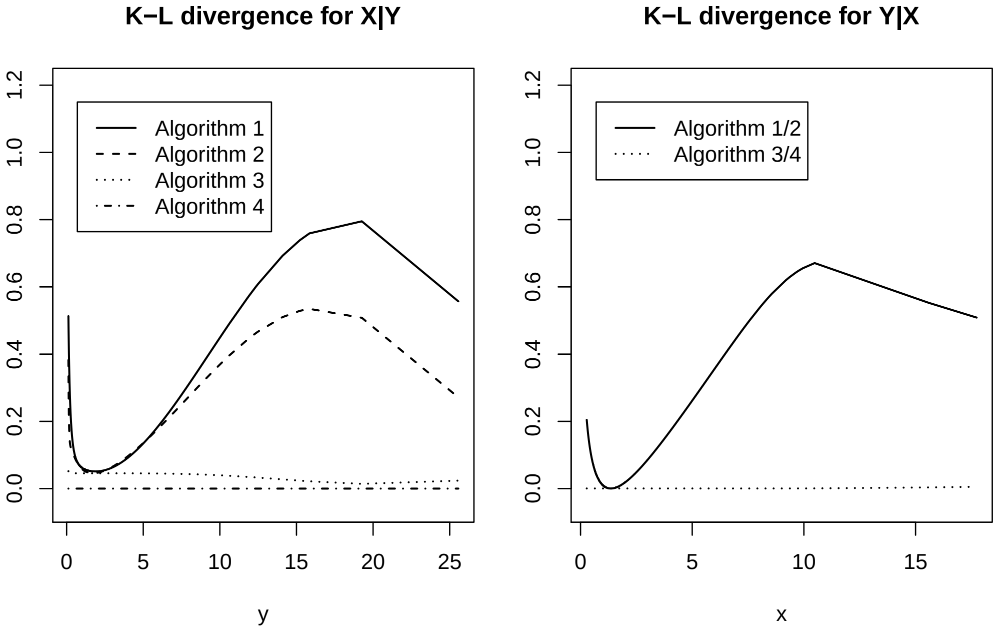

Figure 1: Kullback-Leibler divergence curves between fitted regression models and true conditional densities of four sets of models for Example 6.

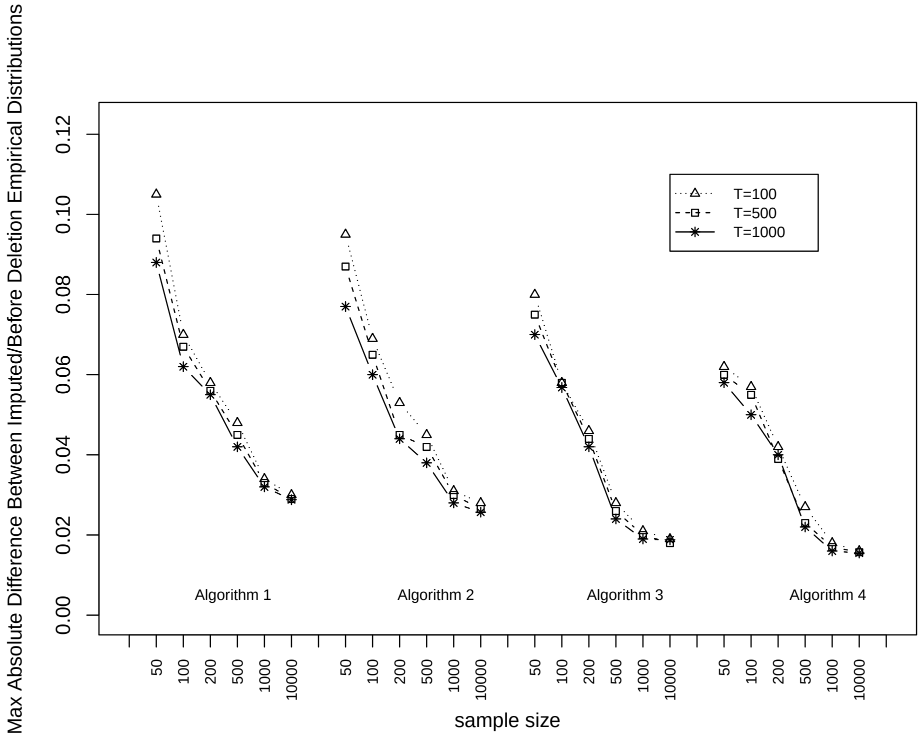

Figure 2: Maximum of absolute difference between empirical distributions based on multiply imputed data and before deletion data, $\left\|\widehat{F}_{MI}^{n,T}(x,y) - \widehat{F}_{BD}^{n}(x,y)\right\|_{\infty}$, from four imputation algorithms plotted as a function of sample size n=50, 100, 200, 500, 1000 and 10000 and the number of iterations $T = 100, 500, 1000$ for Example 6.