Assessment and Improvement of a Sequential Regression Multivariate Imputation Algorithm

by

Jian Zhu

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2016

Doctoral Committee:

Professor Trivellore E. Raghunathan, Chair
Professor Michael R. Elliott
Professor Susan A. Murphy
Associate Professor Lu Wang

This thesis work is dedicated to my wife Jianing and my daughter Yilin.

ACKNOWLEDGEMENTS

I am extremely thankful to my advisor, Professor Trivellore Raghunathan (Raghu). Without his long-term support, I would have never been able to complete my dissertation research. His philosophy of life and statistics are truly inspirational to me, and I greatly appreciate the advice he has given me throughout the years. I will always look upon on him as my role model in my future work.

I am very lucky to have a great and impressive committee. Each member, Professor Trivellore Raghunathan, Professor Michael Elliott, Professor Lu Wang and Professor Susan Murphy, has given me much great advice. I am very grateful for their feedback on shaping my research goals and revising my thesis.

I am very thankful to Dr. Hiroko Dodge, who gave me the opportunity to work with her at Michigan Alzheimer's Disease Center. I have gained a lot of first-hand experience in medical related research through extensive collaboration work with her and her teams.

I would also like to thank the Department of Biostatistics at University of Michigan for its long-term support.

Last but not least, I would like to thank my entire family.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

CHAPTER I

Introduction

1.1 Missing Data Background

In many statistical studies, especially in large complex studies with many types of variables, data can be missing due to various reasons such as attrition, refusal, partial response (Little and Rubin, 2002), and study designs such as matrix sampling (Raghunathan and Grizzle, 1995) and data merging (Wang, Song and Wang, 2015). The missing values often create statistical challenges for analysts in obtaining valid inferences.

One challenge is raised by different missingness patterns. When data are presented in a wide format (rows correspond to subjects and columns correspond to variables), the data matrix displays three missingness patterns: 1. univariate pattern, 2. monotone pattern and 3. arbitrary pattern (Schafer and Graham, 2002). In a univariate pattern, missing values occur only on one variable and all other variables are fully observed; in a monotone pattern with ordered variables, once a variable is missing, then all succeeding variables are missing; the arbitrary pattern is the most general pattern in which different sets of variables can be missing for different subjects. This thesis focuses on the arbitrary pattern in multivariate datasets. In addition, we consider a single variable missing pattern, a special case of the arbitrary pattern in

which data are missing on up to one variable in any row and the missing variable may be different across subjects. Unlike the first two patterns, which can be handled by simple methods, the arbitrary pattern and its special case may require more sophisticated approaches.

Another challenge is raised by different mechanisms that generate the missing values. Take a naive analysis approach for example, which restricts the analysis to subjects providing complete data on variables relevant for the analysis. This approach is the default method in many statistical software packages, and it is valid if the complete cases or available cases are representative of all cases. However, this strong assumption may not be true in general, and, even if true, this approach may lead to loss of efficiency. It is more likely that the inferences from the complete-case analysis will not be valid; for example, the parameter estimates may be biased.

Rubin (1976), a groundbreaking paper, proposed a missing data analysis framework with the definition of three types of missingness mechanisms: missing completely at random (MCAR), missing at random (MAR) and not missing at random (NMAR). Furthermore, when data are MAR and the population or substantive model parameters are distinct from the parameters in the missingness mechanism, the likelihood-based inferences do not require the exact form of missingness mechanism. Within this framework, researchers have developed many methods to properly analyze incomplete data.

The likelihood-based methods generally adopt either frequentist maximum likelihood (ML) methods or fully Bayesian approaches. The Expectation-Maximization (EM) algorithm (Dempster, Laird and Rubin, 1977), and Monte Carlo simulation techniques have made implementation of these methods computationally possible for a selected set of models. The frequentist-based ML methods usually rely on large

sample size whereas the fully Bayesian methods can handle small sample sizes. The parameters of interest can be drawn from the joint posterior distribution along with the missing values by rejection sampling methods or Monte Carlo Markov Chain (MCMC) algorithms such as Gibbs sampler (Geman and Geman, 1984).

Inverse probability weighting (Little and Rubin, 2002) is another way to adjust the complete case analysis, and is especially useful for unit non-response in survey settings. An alternative is the propensity prediction method (Little and An, 2004), which uses the propensity score as a covariate instead of weights.

Both maximum likelihood and Bayesian-based methods are computationally intensive and require development of specialized codes. On the other hand, inverse probability methods can handle special patterns of missing data (such as univariate pattern or monotone pattern). Rubin (1977, 1987) proposed the multiple imputation approach, which capitalizes on the complete data software programs. The multiple imputation approach uses the observed data to estimate the distribution of the missing data, and imputes the missing values from the predictive distribution. To properly estimate the uncertainty due to filling in the missing values, multiple completed data sets are created by the same imputation algorithm, and analysis results from each completed data set are combined to obtain the final inference following Rubin's (1977) formula.

An advantage of this approach, especially when data are missing arbitrarily in a complex study, is that once the missing data are imputed, one can analyze the completed datasets as if there were no missing values by simply using existing software packages. If the same data are analyzed by various researchers, a well-imputed data source can ensure that analysis results by different researchers are consistent. This approach also can be viewed as a small simulation approximation for the fully

Bayesian analysis when the model used for the imputation is also the model for the analysis.

Numerous approaches have been developed for creating multiple imputations. Some common imputation methods include nonparametric methods such as Hot Deck imputation (Andridge and Little, 2010), and model-based methods based on a joint distribution of all variables with missing values conditional on all the variables that have no missing values. However, when a study has a complex data structure with many types of variables, skip patterns, structural dependencies etc., both Hot-deck and joint model approaches are difficult, if not impossible, to implement.

1.2 Sequential Regression Multivariate Imputation (SRMI)

When joint modeling becomes difficult in practice, an alternative is to use a sequential regression multivariate imputation (SRMI) algorithm. This approach assumes a set of univariate regression models for each missing variable and has gained popularity due to its flexibility and ease of implementation. First proposed in Kennickell (1991), this approach is also commonly known as multiple imputation by chained equations (MICE) or multiple imputation by fully conditional specification (FCS).

The algorithm can be described as follows. Let the data set on n subjects consist of q variables with no missing values arranged in an $n \times q$ matrix, U. Let $Y_1, Y_2, \ldots, Y_p$ be p variables with some missing values. Sequential regression multivariate imputation is an iterative approach for imputing the missing values in Y_j through random draws based on fitting a regression model $Pr(Y_j|U, Y_{[-j]})$ where $Y_{[-j]}$ is all the variables with missing values except Y_j (Van Buuren and Oudshoorn, 1999, Raghunathan et al., 2001). Specifically, imputations for Y_j at iteration t are drawn from the predictive

distribution, $Pr(Y_j|U, Y_1^{(t)}, Y_2^{(t)}, \ldots, Y_{j-1}^{(t)}, Y_{j+1}^{(t-1)}, \ldots, Y_p^{(t-1)})$ where $Y_l^{(s)}$ is the filled in Y_l at iteration s (observed or imputed). Each predictive distribution corresponds to an appropriate regression model.

At each iteration, imputation involves two steps: (1) the regression model is fit to the observed values of the variable being imputed and all other variables (observed or imputed), and the parameters are drawn from the approximate posterior distribution; and (2) the draws from the regression model given the drawn parameters and all other variables are used as imputations for the missing values. The software IVEWARE (Raghunathan et al., 2001) implements this approach using a fully Bayesian approach (that is, Steps 1 and 2). 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 PROC MI in SAS (2008), MICE (Van Buuren and Oudshoorn, 1999) and MI (Su et al., 2011) in R, STATA (Royston, 2005) and SPSS (SPSS, 2009).

The sequential regression approach has two major practical advantages over other model-based imputation methods. Firstly, it enables handling of complex data structures by focusing on a set of regression models with a univariate outcome. The flexible selection of regression models also enables better prediction of the missing values based on other variables, and the regression models are more intuitive to analysts than a joint model. Secondly, individual regression models can easily account for study designs such as skip patterns, logical constraints, bounds for imputed values and consistency requirements.

1.2.1 Issues

A theoretical weakness of this approach is that the specifications of fully conditional distributions for a set of variables do not guarantee the existence of a joint

distribution. Theoretical assessment of SRMI is limited in the literature. In Liu et al. (2014), the convergence properties were studied when the joint distribution exists; however, this may not be the case for general SRMI by common GLMs. Empirical studies have shown that a few iterations are sufficient to utilize the predictive power of the observed covariates in creating imputations (Van Burren, 2007), but such examples are often limited to linear regression models. Similar simulation studies (Collins, Schafer and Kam, 2001) have proposed guidance for SRMI such as recommending the use of the most inclusive policy; therefore, careful examination for general model specification is needed.

When the missingness is general and models other than linear regression models are used, SRMI may not work well in some cases. Li, Yu and Rubin (2012) have demonstrated the need for caution through the use of some theoretical examples. This thesis studies one motivating trivariate count variable example extensively. In particular, a simulation study shows that with poorly fit model sequences, one SRMI algorithm for this example may diverge. Therefore, the regular SRMI needs to be modified to avoid such behavior.

1.3 Thesis Layout

In order to address the issues above, this thesis work investigates the convergence properties of SRMI in various cases, and proposes modifications for improvement. It consists of the following chapters.

Chapter 2 assesses the convergence properties of SRMI for the simple case where each subject may be missing a value on at most one variable. We define several classes of generalized linear regression model sequences according to their model compatibility and validity properties. We also establish two sufficient conditions that

allow the algorithm to perform well. For all types of compatible and incompatible model sequences, we conduct simulation studies to evaluate their convergence and performance. We then use the results to develop criteria for the choice of imputation models. This chapter was published as Zhu and Raghunathan (2015).

Chapter 3 proposes a modified block sequential regression multivariate imputation (BSRMI) approach to divide the data into blocks when imputing each variable based on missing data patterns and tune the regression models through compatibility restrictions. This is extremely helpful to avoid divergence when it is difficult to get well-fitting models across all records with missing values for a general pattern of missing data. We establish two sufficient conditions for the convergence of the algorithm, and study the repeated sampling properties of inferences using several simulated data sets.

Chapter 4 extends the imputation model selection to quasi-likelihood regression models in both SRMI and BSRMI, when it is difficult to identify a well-fitting generalized linear model sequence. We examine the performance of the modified approach through simulation studies. We demonstrate that the new approach can be used to improve repeated sampling properties due to their improved model prediction.

Finally, Chapter 5 summarizes the findings and discusses limitations and future work.

CHAPTER II

Convergence Properties of SRMI for Single Variable Missing Patterns

Abstract

A sequential regression or chained equations imputation approach uses a Gibbs sampling type iterative algorithm which 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 is 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. This chapter develops conditions for convergence and assesses 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. The sequence of regression models are assumed to be empirically good fit for the data chosen by the imputer based on appropriate model diagnostics. The results are used to develop criteria for the choice of regression models.

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

2.1 Introduction

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 uses 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 $Y_{[-i]}^{(t)} = \{Y_1^{(t)}, \ldots, Y_{i-1}^{(t)}, Y_{i+1}^{(t-1)}, \ldots, Y_p^{(t-1)}\}$, 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, the first step may be skipped and the maximum likelihood or any other consistent estimate of θ_i may be used 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 $\{Y_{i,obs}, Y^{(t)}_{[-i]}\}$ (typically the maximum likelihood estimate $\widehat{\theta}^{(t)}_i$) in Step 2.

The sequential regression 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 models. 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 a set of regression models with a univariate outcome. 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 skip 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 (that is, Steps 1 and 2). 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 PROC MI in SAS (2008), MICE in the R package(Van Buuren and Oudshoorn, 1999) and in STATA (Royston, 2005). The recent issue of the Journal of Statistical Software (2012) has published

several articles on this approach.

A theoretical weakness of this approach is that the specifications of conditional distributions for a set of variables 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. This problem was also discussed in the context of spatial analysis (Besag, 1974), and the 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.

In the sequential regression imputation context, Van Buuren et al. (2006) showed through simulations that incompatibility caused minimal effects in some cases. Drechsler and Rassler (2008) showed that choosing poorly fitting incompatible models may lead to biased estimation. From a theoretical perspective, Li et al. (2012) used incompatible regression models with fixed parameters to illustrate that model incompatibility may or may not necessarily imply algorithmic incompatibility (convergence). However, they fix the parameters across all the iterations but in the sequential regression approach parameters change with iterations. That is, the sequential regression approach is a Markov type process. Liu and Gelman (2013) established technical conditions for the convergence of the sequential regression approach if the stationary joint distribution exists. In practice, however, the models may be chosen where the stationary joint distribution may not exist. Hence, we need to investigate the

convergence properties under more general conditions.

The incompatibility may not lead to divergence can be illustrated using the following bivariate example. Suppose that the data set with 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 $m_1(y \mid x, \theta_1) \sim exponential(\theta_1 x)$ and $m_2(x \mid y, \theta_2) \sim exponential(\theta_2 y)$ 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 $exponential(\widehat{\theta}_1^{(t)} x)$ 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 $exponential(\widehat{\theta}_2^{(t)} y)$ 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 2.1 given in the next section provides sufficient conditions for the sequential regression imputation approach to yield 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 chapter is organized as follows. Section 2.2 provides definition of incompatibility and model validity to classify regression models in the sequential regression approach. Section 2.3, focusing on bivariate scenario, provides two sufficient conditions for the convergence of the sequential regression approach and obtain consistent estimators. Section 2.4 enhances the analytical results given in Sections 3 through a simulation study for incompatible but approximately valid or well fitting model sequences. Section 2.5 considers the convergence properties under nonignorable missing data mechanisms. Section 2.6 extends the results for multivariate missing data with single variable missing data pattern (that is, any subject is missing at most one variable). Section 2.7 summarizes the findings, discusses extensions for arbitrary pattern of missing data and the limitation of the sequential regression algorithm.

2.2 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 $(\theta_1, \theta_2, \ldots, \theta_p)$.

Two linear regression models, $m_1(y|x, \theta_1) \sim N(\theta_{1o} + \theta_{11}x, \sigma_{12}^2)$ and $m_2(x|y, \theta_2) \sim N(\theta_{20} + \theta_{21}y, \sigma_{21}^2)$ are weakly compatible if $\theta_{11}/\sigma_{12}^2 - \theta_{21}/\sigma_{21}^2 = 0$, $\theta_{11}^2 \neq \sigma_{12}^2/\sigma_{21}^2$ 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. They are also possibly compatible if the target joint distribution is a bivariate normal distribution.

A subclass of possibly compatible model sequence with separate parameters for marginal and conditional distributions is defined below. Anderson (1958) used the separable parameters to develop maximum likelihood estimates for the mean and the covariance matrix of the multivariate normal distribution with a monotone pattern of missing data.

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.

As an example, suppose that $(Y_1, \ldots, Y_p)$ is a p-dimensional continuous variable, and the 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 $y_i \in Y_C = Y - Y_M$, $m_i(y_i \mid y_{[-i]}, \theta_i) \sim N(\theta_{i0} + \sum_{j \neq i} \theta_{ij} y_j, \sigma_i^2)$.

Suppose that $(Y_1, \ldots, Y_p)$ is a p-dimensional binary variable, and model sequence consists of

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

where $i = 1, 2, \ldots, p$. 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 $y_i \in Y_C = Y - Y_M$,

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

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.

2.3 Bivariate Missing Data

Before we consider the multivariate imputation problem, we consider the bivariate case, mostly for notational simplicity and ease of presentation. For now, we assume that the missing data mechanism is ignorable as in Rubin (1976) and all

the conditional distributions belong to the exponential family. We consider non-ignorable missing data mechanisms in Section 2.5. 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:

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

where $\theta_i = (\phi_i, \delta_i), i = 1, 2$ and the link functions h_1 and h_2 connect the condi-

tional 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 2.1. *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 $m_1(y \mid x, \theta_1^{(t)}) \to f_{Y|X}(y \mid x, \psi_1)$ and $m_2(x \mid y, \theta_2^{(t)}) \to f_{X|Y}(x \mid y, \psi_2)$.

The proof of the Theorem is given in Appendix 2.1. To illustrate further, we

consider two examples to assess the convergence properties of the asymptotic iterative algorithm.

Example 2.1 (Two Linear Regression Models): Suppose the data (X, Y) are generated from a bivariate normal distribution $\text{BVN}(\mu, \Sigma)$ 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 asymptotic iterative algorithm is calculated in Appendix 2.2. 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 2.2 (Two Logistic Regression Models): 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 asymptotic iterative algorithm is calculated in Appendix 2.2. 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 2.2. *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 2.1 with respect to differentiation/integration and the existence of the mean/variance of the score functions ,* $m_1(y \mid x, \theta_1^{(t)}) \to p(y \mid x, \theta_1^*)$ *and* $m_2(x \mid y, \theta_2^{(t)}) \to p(x \mid y, \theta_2^*)$ *as* $n,\ t \to \infty$.

The proof of this theorem is given in Part 2 of Appendix 2.1. 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 2.1 and Example 2.2, since $({\theta_1}^*, {\theta_2}^*) \in \Theta_C$ holds inherently, according to Theorem 2.2, 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 2.3 (Two Poisson Regression Models): Consider, $m_1(y|x, \theta_1) \sim Poisson(exp(\theta_{10} + \theta_{11}x))$ and $m_2(x|y, \theta_2) \sim Poisson(exp(\theta_{20} + \theta_{21}y))$. The compatibility condition requires that $\theta_{11} = \theta_{21} < 0$, and the corresponding joint distribution 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 (T, the number of iterations, is set as 10000):

(1) We generated the complete data from $Y \sim$ Poisson(2.5) and $X \mid Y \sim$ 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 $Y \sim$ Poisson(2.5) and $X \mid Y \sim$ 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.

Suppose X is a binary variable and Y is a continuous variable, and sequential regression imputation assumes the following regression models:$m_1(y \mid x, \theta_1) \sim N(\theta_{10} + \theta_{11}x, \sigma_{12}^2)$ and, $m_2(x \mid y, \theta_2) \sim Bernoulli\left[(1 + \exp(-\theta_{20} - \theta_{21}y))^{-1}\right]$.

The target joint distribution exists under the compatibility condition $\theta_{11}/\sigma_{12}^2 - \theta_{21} = 0$. Although the parameter of the marginal distribution of X can be separated from the conditional distribution of $Y|X$, the parameters of the marginal distribution of Y cannot be separated from the conditional distribution of $X|Y$. Therefore, this is a possibly compatible model sequence with non-separable marginal parameters.

We generated 100 replicates of data of sample size 1000 from a population with

$x \sim Bernoulli(p = 0.7)$, $y|x \sim N(1 - 2x, 1)$. To generate the missing values, we divided the data into two random halves. In the first group, y is observed and the probability of x being observed is $pr(x \text{ is observed} \mid y) = [1 + \exp(-0.45y + 2)]^{-1}$ and in the second group, x is observed and the probability of y being observed is $pr(y \text{ is observed} \mid x) = [1 + \exp(-0.5x - 0.3)]^{-1}$. SRMI algorithm using linear and logistic regression model sequence is applied on each replicate to obtain 10 multiply imputed data sets. For each regression parameter, for instance θ_{10}, the average of the estimate, $\theta_{10}(r) = \sum_{m=1}^{10} \sum_{t=501}^{1000} \theta_{10}^{(t)}(m, r)/500/10$, after a burn-in of 500 iterations is calculated for the rth replicate, and then the mean of the 100 averaged estimates parameters and the corresponding standard deviation are obtained. These values for various parameters are $\bar{\theta}_{10} = 1.01(sd = 0.08)$, $\bar{\theta}_{11} = -2.00(sd = 0.10)$, $\bar{\sigma}_{12} = 1.00(sd = 0.03)$, $\bar{\theta}_{20} = 0.88(sd = 0.17)$, $\bar{\theta}_{21} = -2.04(sd = 0.21)$. Most importantly the left side of the compatibility condition $\theta_{11}/\sigma_{12}^2 - \theta_{21}$ has a mean of 0.03 and standard deviation of 0.07 from the 100 replicates. The results show that since linear and logistic regression model sequence is validly specified for the data set, unbiased estimates of the regression parameters and compatibility conditions are asymptotically reached during SRMI procedure as implied by Theorem 2.1.

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 2.1, 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.

2.4 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 an example to show that a well-fitting incompatible model sequence can be approximately validly specified:

Example 2.4 (Two Gamma Regression Models):

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

For the simulation study, we generated data from the following population distri-

bution:

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

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 this single data set, we determined the following four sequential regression imputation methods using different sets of reasonable regression models with varying degree of incompatibility to impute the missing values:

1. Sequence 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. Sequence 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. Sequence 3 uses the incompatible regression model set:

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

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

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

Sequences 1 and 2 impute the missing values on the cube root scale, and then transformed to the original scale.

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 2.1 show the divergence curves for each model from the four sets corresponding to $f_{X|Y}$ and $f_{Y|X}$. From Sequence 1 to Sequence 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 Sequence 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 Sequence 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 Sequence 3 as a well-fitting model sequence for (X, Y). The choice of these 4 sequences enables us to demonstrate that better fitting model sequences yield better imputation results regardless of model compatibility/incompatibility.

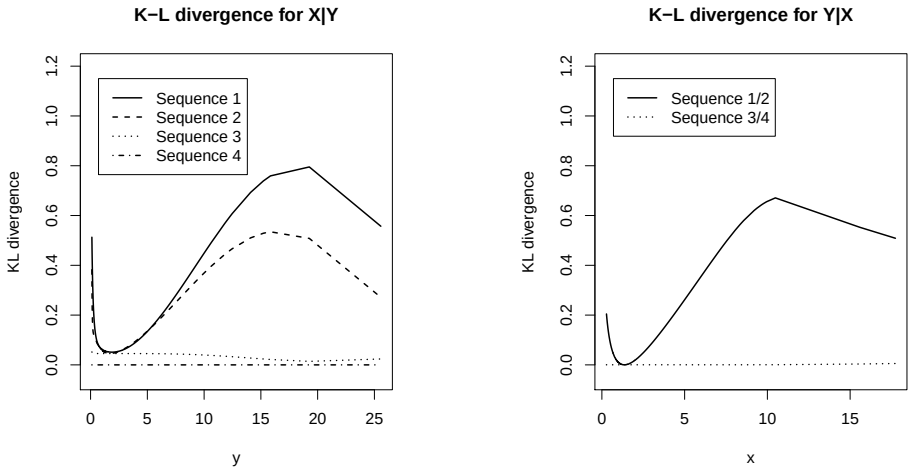

Figure 2.1: Kullback-Leibler divergence curves between fitted regression models and true conditional densities of four sets of models for Example 2.4.

All these models are reasonably well fitting models and cover the span of incompatibility that includes: (1) Possibly compatible sequence with separable marginal

parameters; (2) Possibly compatible sequence with non-separable marginal parameters; (3) Incompatible model sequence; and, finally, (4) Specified possibly compatible sequence with non-separable marginal parameters; Model sequence 3 is a very well-fitting but incompatible model sequence that can be chosen by imputers because of the skewness of the residuals. The last model sequence may not likely to be chosen in practice but it is from the true conditional distributions; This enables us to demonstrate that convergence can be reached for all types of model sequences.

We also want to illustrate a gradual improvement strategy for practical purpose through the first two algorithms: after conducting proper transformation such as Box-Cox transformation determined by the available data, imputers may start with a simple linear regression model sequence for the transformed data, and then gradually improve the linear model sequence by fine-tuning it based on further model diagnosis, such as adding extra nonlinear terms in the models; in our example, residual plots suggest that the extra non-linear term $1/y$ in model m_{22} improves model fitting compared to model m_{12}.

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: $\left\| \widehat{F}^{n,T}_{MI}(x,y) - \widehat{F}^{n}_{BD}(x,y) \right\|_{\infty}$. 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 2.2. All sequences 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 sequence 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 Sequence 1; 0.0257 (0.006) for Sequence 2; 0.0188 (0.005) for Sequence 3 and 0.0155 (0.004) for Sequence 4. The empirical joint distribution difference decreases from Sequence 1 to Sequence 4, indicating that as the model fitting is improved, the performance is improved as well. Both incompatible but better fitting sets from Sequences 2 and 3 outperform the possibly compatible set with separable marginal parameters from Sequence 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.

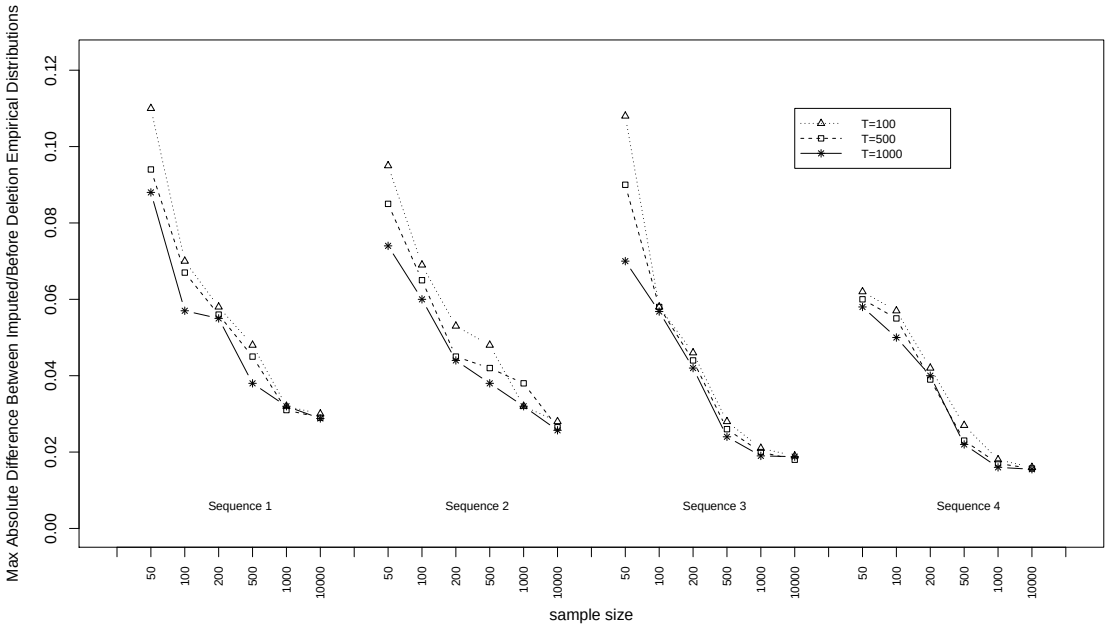

Figure 2.2: Maximum of absolute difference between empirical distributions based on multiply imputed data and before deletion data, $\left\|\widehat{F}^{n,T}_{MI}(x,y) - \widehat{F}^{n}_{BD}(x,y)\right\|_{\infty}$, from four imputation algorithms plotted as a function of sample size n=50, 100, 200, 500, 1000 and 10000 and the number of iterations $T = 100, 500, 1000$ for Example 2.4.

2.5 Non-ignorable Missing Data Mechanism

The sequential regression approach discussed earlier in this chapter ignores the missing data mechanism and, hence, we limited our discussion to missing at random mechanism. Nonetheless, it may be important to know what would happen if a user were to use this approach when the mechanism is nonignorable.

- When data are missing not at random, the meaning of the validly specified model is not clear and the modeling task will be different than the usual sequential regression approach. Theorem 2.1 clearly requires a validly specified model and we believe that with a nonignorable missing data mechanism, the sequential regression model ignoring the missing data mechanism cannot be a validly specified model.

- It should be noted that even if one were to formulate a joint distribution for the variables, ignore the missing data mechanism when actually it is nonignorable then the algorithm (for example, Gibbs Sampling) may converge to something that has very little meaning. This may be true also for SRMI. We investigate further through a simulation study described later.

- Theorem 2.2 on the other hand is valid, as the construction and proof of Theorem 2.2 do not rely on missing data mechanism but depend only on the separable marginal parameter property of the target joint distribution.

To investigate further, we extended our simulation study in section 2.4 by changing only the missing data mechanism. We divided data into two equal random groups. In the first group, y is fully observed and the probability of x being observed is $pr(x \text{ is observed} \mid y) = [1 + \exp(15 + 0.1y - 5x)]^{-1}$; In the second group, x is fully observed and the probability of y being observed is $pr(y \text{ is observed} \mid x) =$

$[1 + \exp(12 - 6y + 0.1x)]^{-1}$. This results in about 45% of x values and 35% of the y values to be missing. Both x and y are missing Not at Random. Figure 2.3 shows a plot of the maximum difference between before deletion and after imputation empirical CDFs; while the difference measure still converges for all algorithms as the sample size and number of iterations increase, none of the difference measures is close to 0, indicating biased imputation results due to ignoring the missing data mechanism.

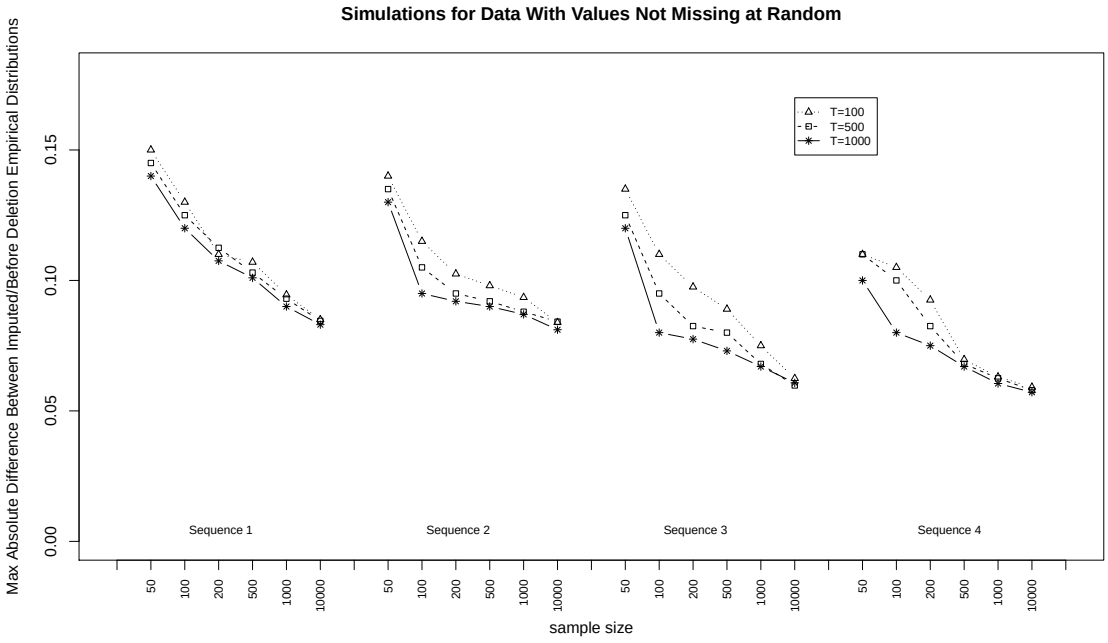

Figure 2.3: Maximum of absolute difference between empirical distributions based on multiply imputed data and before deletion data, $\left\| \widehat{F}^{n,T}_{MI}(x,y) - \widehat{F}^{n}_{BD}(x,y) \right\|_{\infty}$, from four imputation algorithms plotted as a function of sample size n=50, 100, 200, 500, 1000 and 10000 and the number of iterations $T = 100, 500, 1000$ when the data are not missing at random in Example 2.4.

2.6 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 $Y^{(t)}_{[-i]}$. 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.

2.6.1 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 $Y^{(t)}_{[-i]}$ form up to p patterns, and the log-likelihood is

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

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

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

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 2.3 can be extended. The following theorems are a generalization of Theorems 2.1 and 2.2 for the bivariate case.

Theorem 2.3. 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 $i = 1, \ldots, p$, $m_i(y_i \mid y_{[-i]}, \theta_i^{(t)}) \to f_i(y_i \mid y_{[-i]}, \psi_i)$, as $n,\ m,\ t \to \infty$.

Theorem 2.4. 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 2.3, 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 $m_i(y_i \mid y_{[-i]}, \theta_i^{(t)}) \to p(y_i \mid y_{[-i]}, \theta_i^*)$, as $n,\ m,\ t \to \infty$.

Proof of both theorems are extensions of proofs for Theorems 2.1 and 2.2 and these are included in the supplementary material available on the website or from the authors.

2.7 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 chapter.

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 may be 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 2.1 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 2.1

Part 1: Proof of Theorem 2.1

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:

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

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 $m_1(y \mid x, \theta_1^{(t)}) \to f_{Y|X}(y \mid x, \psi_1)$ 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 2.2

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:

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

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:

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

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.

Appendix 2.2

Examples

Example 2.1 (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: $\pi_0 = Pr(\text{both } X \text{ and } Y \text{ are observed})$, $\pi_1 = Pr(Y \text{ is observed and } X \text{ is missing})$ and $\pi_2 = Pr(X \text{ is observed and } Y \text{ is missing})$.

The estimated regression parameters converge to $\theta_1^* = (\alpha_{10}, \alpha_{11}, \tau_{12}^2)^T$, 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:

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

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

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.

- Example 2.2 (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:

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

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

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.

2.8 Supplemental Materials

Proof of Theorem 2.3

Proof: As in Theorem 2.1, 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:

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

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 2.2, 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 2.2, 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 $i = 1, \ldots, p$.

Proof of Theorem 2.4

Proof: As in Theorem 2.2, 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 $i = 1, \ldots, p$,

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

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 $i = 1, \ldots, p$.

CHAPTER III

Block Sequential Regression Multivariate Imputation Algorithm (BSRMI)

Abstract

Multiple imputation using sequential regression (chained equations, fully conditional specifications) is a popular approach for handling missing values in a complex data structure with many types of variables, structural dependencies among the variables and bounds on plausible imputation values. The sequential regression approach is a Gibbs style algorithm with iterative draws from the posterior predictive distribution of missing values for any given variable, conditional on all observed values and imputed values of all other variables. As this approach only requires the conditional distribution of each variable with missing values, this collection may not be compatible with any joint distribution of the variables with missing values. However, many theoretical investigations and empirical studies have shown that this approach produces valid inferences (from the repeated sampling perspective) if the regression models fit the data well and convergence is obtained under a broad set of regularity conditions. Sometimes it is not possible to get well-fitting models across all missing data patterns, which leads to difficulties in implementing this approach. We propose a modification where the data is divided into blocks for each variable based on miss-

ing data patterns and the regression models are tuned through a set of compatibility restrictions. This increases the number of models to be fit but reduces the model complexity. We establish regularity conditions for the convergence of the algorithm, and study the repeated sampling properties of inferences using several simulated data sets. Throughout, we assume that the missing data mechanism is ignorable.

Key Words: Missing data; Multiple imputation; Chained equations, Compatible conditionals; Conditional specifications; Block specifications;

3.1 Introduction

Let the data set on n subjects consist of q variables with no missing values arranged in an $n \times q$ matrix, U (if all variables in the data set have missing values then U is considered as a column of 1s, representing the intercept term). Let $Y_1, Y_2, \ldots, Y_p$ be p variables with some missing values. The sequential regression (chained equations, fully conditional specifications) imputation approach, first proposed in Kennickell (1991), is an iterative approach for imputing the missing values in Y_j through random draws based on fitting a regression model $Pr(Y_j|U, Y_{[-j]})$ where $Y_{[-j]}$ is all the variables with missing values except Y_j (Van Buuren and Oudshoorn, 1999, Raghunathan et al., 2001). Specifically, imputations for Y_j at iteration t are drawn from the predictive distribution, $Pr(Y_j|U, Y_1^{(t)}, Y_2^{(t)}, \ldots, Y_{j-1}^{(t)}, Y_{j+1}^{(t-1)}, \ldots, Y_p^{(t-1)})$ where $Y_l^{(s)}$ is the filled in Y_l at iteration s (observed or imputed). Each predictive distribution corresponds to an appropriate regression model. At each iteration, imputation involves two steps: (1) the regression model is fit to the observed values of the variable being imputed and all other variables (observed or imputed), and the parameters are drawn from the approximate posterior distribution; and (2) the draws from the regression model given the drawn parameters and all other variables are used as im-

putations for the missing values. We assume that the regression models are a good fit determined by appropriate model diagnostics.

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 a set of regression models with a univariate outcome. 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 skip patterns, logical constraints, bounds for imputed values and consistency requirements. Empirical studies, however, have shown that a few iterations are sufficient to utilize the predictive power of the observed covariates in creating imputations.

A theoretical weakness of this approach is that the specifications of conditional distributions for a set of variables do not guarantee the existence of a compatible joint distribution. The convergence properties of these algorithms have been investigated in Liu et al. (2014), Hughes et al. (2014) and Chapter 2 (Zhu and Raghunathan, 2015). Ultimately, developing well-fitting regression models is critical for obtaining valid inferences using this procedure.

Finding a set of well-fitting regression models (within the parametric framework) may be a challenging task when several variables are simultaneously missing. Consider the following trivariate count data generated using the following model assumptions,

$$
\begin{aligned}
y_1 &\sim \text{Poisson}(\lambda = \exp(\alpha_{10})), \\
y_2 \mid y_1 &\sim \text{Poisson}(\lambda = \exp(\alpha_{20} + \alpha_{21} y_1)), \\
y_3 \mid y_1, y_2 &\sim \text{Poisson}(\lambda = \exp(\alpha_{30} + \alpha_{31} y_1 + \alpha_{32} y_2)),
\end{aligned}
$$

where $\alpha_{10} = \log(2),\ \alpha_{20} = -1,\ \alpha_{21} = 0.3,\ \alpha_{30} = -2,\ \alpha_{31} = 0.5$ and $\alpha_{32} = 0.1$.

Suppose that the variables are missing at random with a general pattern of missing data mechanism, which creates 7 possible missing data patterns as follows (✓ denotes observed and ? denotes missing):

P	Y_1	Y_2	Y_3
1	✓	✓	✓
2	✓	✓	?
3	✓	?	✓
4	✓	?	?
5	?	✓	✓
6	?	✓	?
7	?	?	✓

Given the count type of variables, one might attempt to use an SRMI approach assuming the following three Poisson regression models:

$$
\begin{aligned}
m_1(y_1 \mid y_2, y_3) &\sim \text{Poisson}(\lambda = \exp(\theta_{10} + \theta_{12} y_2 + \theta_{13} y_3)), \\
m_2(y_2 \mid y_1, y_3) &\sim \text{Poisson}(\lambda = \exp(\theta_{20} + \theta_{21} y_1 + \theta_{23} y_3)), \\
m_3(y_3 \mid y_1, y_2) &\sim \text{Poisson}(\lambda = \exp(\theta_{30} + \theta_{31} y_1 + \theta_{32} y_2)).
\end{aligned}
$$

It is easy to verify that the algorithm quickly breaks down due to extremely large values imputed within the first few iterations. It is also easy to check that these

are very poor fitting models for certain conditional distributions. Specifically, the model diagnostics will show that m_1 and m_2 are very poor fitting models(especially m_1). It is not easy to improve the model fitting within this Poisson family, given the constraint that the mean and the variance functions are the same. It is also not easy to find other members of the Generalized Linear Model (GLM) family that fits the data well. It is desirable, therefore, to modify the algorithm to ensure convergence even for poor-fitting models, and then gradually improve the model fit to achieve better imputation results without concerns about convergence.

Li et al. (2012) distinguish between model and algorithm incompatibilities. The former does not necessarily lead to the latter, which is directly linked to divergence. When data are missing simultaneously for several variables, the poor-fitting models can become algorithmically incompatible. To be specific, suppose that only Y_1 and Y_2 are missing among all the records. The two imputation models (Y_1 given Y_2 and Y_3) and (Y_2 given Y_1 and Y_3) are poorly chosen so that they are algorithmically incompatible for all values of the parameters, and hence at any iteration, jointly, the drawn values of Y_1 and Y_2 in those records may become incompatible with observed values of Y_3, reaching to divergence. This issue does not arise for data missing up to one variable in any record, since the drawn values to complete any record are always from univariate conditional models given observed values of all other variables, and in that case model incompatibility does not necessarily lead to algorithm incompatibility (Zhu and Raghunathan, 2015).

We propose a modification of the algorithm where previously imputed variables obtained under poorly fitting models have less influence on the subsequent variables. Difficulties in getting well-fitting models are due to differences in the information base across the patterns for developing models. For instance, to obtain missing

values in Y_1, for the subjects in Pattern 5, the information is in the regression model $Pr(Y_1|Y_2, Y_3)$, for the subjects in Pattern 6, it is $Pr(Y_1, Y_3|Y_2)$ and, for Pattern 7, $Pr(Y_1, Y_2|Y_3)$. The modified algorithm proposes imputation of missing values in Y_1 in Pattern 5 using $m_{1.23}(Y_1|Y_2, Y_3)$, using $m_{1.2}(Y_1|Y_2)$ for Pattern 6 and using $m_{1.3}(Y_1|Y_3)$ for Pattern 7. For imputing the missing Y_2 values, we need three models, $m_{2.13}$ for Pattern 3, $m_{2.1}$ for Pattern 4 and $m_{2.3}$ for Pattern 7. Although this increases the number of models, it allows for fine tuning the model to the available information and, thus, reducing the chances of algorithmic incompatibility. Note that under the MAR mechanism, the data across all patterns with observed values of the variable being imputed are used to fit the models, though models used to impute the missing values in each pattern can be different.

The proposed block sequential multivariate imputation (BSRMI) algorithm can be viewed as a generalization of the ordered monotone blocks approach in Li et al. (2014), which reorganizes observations to obtain ordered monotone blocks in the pattern of missing data and then uses compatible models to obtain imputations. BSRMI enforces compatibility conditions for convergence described in Zhu and Raghunathan (2015) within each missingness block and leads to overall convergence, as proved later.

The rest of the chapter is organized as follows. Section 3.2 introduces the proposed algorithm for a general pattern of missing data. Section 3.3 provides some theoretical results of the proposed algorithm convergence and validity, especially when the model sequence is validly specified. Section 3.4 revisits the motivation example, and examines performance of the proposed algorithm through a simulation study. Section 3.5 summarizes the findings, and discusses the advantages and limitations of the proposed sequential regression algorithm.

3.2 BSRMI Algorithm

Without loss of any generality, we assume that all imputation models include the set of fully observed variables U as predictors and suppress this hereafter to simplify notations. Suppose the data set consists of p variables, $Y_1, \ldots, Y_p$, with missing values to be imputed in that order. This is an arbitrary order and we later show that the algorithm is invariant with respect to the choice of the order provided all are well-fitting models. Denote $j_- = \{1, \ldots, j-1\}$, $j_+ = \{j+1, \ldots, p\}$, and $[-j] = \{1, \ldots, j-1, j+1, \ldots, p\}$ as subsets of the indices $\{1, 2, \ldots, p\}$. A missing data pattern P is a made-up group of subjects with data records where a set of variables are completely missing and others are fully observed. Suppose the data set has a total of K unique missing patterns (the maximum possible value of K is $2^p - 1$ where subjects missing all the values are the unit non-respondents and are not considered here). In pattern $P = k$, $k = 1, \ldots, K$, let p_k be the number of missing variables, with ordered indices

$$\mathcal{M}_k = \{I_1, \ldots, I_{p_k}\} \subset \{1, \ldots, p\}.$$

The corresponding observed set is an ordered set

$$\mathcal{O}_k = \{1, \ldots, p\} \backslash \mathcal{M}_k.$$

Let $\boldsymbol{Y}_{[\mathcal{M}_k]} = \{Y_i\}_{i \in \mathcal{M}_k}$ and $\boldsymbol{Y}_{[\mathcal{O}_k]} = \{Y_i\}_{i \in \mathcal{O}_k}$ denote the sets of missing and observed variables in pattern $P = k$ respectively, and $(\boldsymbol{Y}^{mis}_{[\mathcal{M}_k]}, \boldsymbol{Y}^{obs}_{[\mathcal{O}_k]})$ denote the missing and observed data sets in that pattern.

To impute missing values in the pattern $P = k$, we need to specify a sequence of p_k regression models, $m_{I_1.\mathcal{O}_k}$, $m_{I_2.I_1,\mathcal{O}_k}, \ldots,$ $m_{I_{p_k}.I_1,I_2,\ldots,\mathcal{O}_k}$. Though these are the models used in imputing the missing values for the subjects in the particular pattern,

each model is fit using all the observed values of the dependent variable across all patterns, and predictors may be observed or imputed. Thus, some imputed values of other variables enter into the model fitting process.

In general, for any variable Y_j, if it is missing in pattern k, (that is, $\{k = 1, \ldots, K \mid j \in \mathcal{M}_k\}$) then the imputations are draws from the regression model

$$m_{j.\overline{\mathcal{M}_k(j_+)}} = m(Y_j \mid \mathbf{Y}_{[\overline{\mathcal{M}_k(j_+)}]}, \theta_{j.\overline{\mathcal{M}_k(j_+)}}),$$

where

$$\mathcal{M}_k(j_+) \doteq \{I_i \in \mathcal{M}_k, I_i > j\}$$

defines the indices of missing variables succeeding Y_j,

$$\overline{\mathcal{M}_k(j_+)} \doteq [-j] \backslash \mathcal{M}_k(j_+)$$

defines the indices of all other variables excluding missing variables succeeding Y_j and $\theta_{j.\overline{\mathcal{M}_{k(j_+)}}}$ is the parameter in the regression model with prior density $\pi(\theta_{j.\overline{\mathcal{M}_{k(j_+)}}})$[1] The model index subscript $j.\overline{\mathcal{M}_k(j_+)}$ indicates that each model regresses Y_j on all other variables except the set of missing variables after Y_j; in other words, imputation of Y_j depends on all preceding variables and all succeeding observed variables in the sequence.

Note that when the values of $\mathbf{Y}_{[\mathcal{M}_k]}$ are drawn from the above ordered models with some draw of the parameter values $\{\theta^*_{I_i.\overline{\mathcal{M}_k(I_{i+})}}\}$, the observed and drawn values follow a joint distribution defined by

$$d(\mathbf{Y}_{[\mathcal{M}_k]}, \mathbf{Y}_{[\mathcal{O}_k]} \mid P = k) = \left(\prod_{i=1}^{p_k} m\left(Y_{I_i} \mid \mathbf{Y}_{[\mathcal{M}_k(I_{i-})]}, \mathbf{Y}_{[\mathcal{O}_k]}, \theta^*_{I_i.\overline{\mathcal{M}_k(I_{i+})}}\right) \right) \times f(\mathbf{Y}_{[\mathcal{O}_k]} \mid P = k).$$

The number of models needed may be reduced if multiple patterns with missing Y_j share the same missing variable set $\mathcal{M}^*(j_+)$ after Y_j. To be specific, such patterns

[1]The bar is slightly different from a complement sign in set theory and the defined set does not include j. If a complement sign is used then j will be included.

can be combined to create one imputation block for Y_j:

$$B_{j.\overline{\mathcal{M}_k^*(j_+)}} = \{k \mid \mathcal{M}_k(j_+) = \mathcal{M}^*(j_+)\},$$

and only one model needs to be specified. We also define an estimation block for Y_j combining all patterns with observed Y_j: $B_{j0} = \{k \mid \mathcal{M}_k \subset [-j]\}$. The combination of similar patterns into blocks can reduce the total number of models needed to implement the algorithm.

Given the above notations, the Block-Specific Sequential Regression Imputation algorithm is executed as follows:

Iteration $t = 1$**:** The first iteration of the block-specific SRMI algorithm fills in the missing values with initial values. We choose a conditionally specified model sequence, $m(Y_1 \mid \theta_1)$ and $m(Y_j \mid \boldsymbol{Y}_{[j_-]}, \theta_{j.[j_-]})$, for $j = 2, \ldots, p$, to determine the initial values. Let $Y_j^{(t)} = \{Y_{j,obs}, Y_{j,mis}^{(t)}\}$, for $j = 1, \ldots, p$.

For the first variable Y_1 :

1. $m(Y_1 \mid \theta_1)$ is fit to $Y_{1,obs}$, and $\hat{\theta}_1$ is drawn from its approximate posterior distribution;
2. $Y_{1,mis}^{(1)}$ is drawn from $m(Y_1 \mid \hat{\theta}_1)$.

For the jth variable Y_j, $j = 2, \ldots, p$:

1. $m(Y_j \mid \boldsymbol{Y}_{[j_-]}, \theta_{j.[j_-]})$ is fit to the data in $\boldsymbol{Y}_{[B_{j0}]}^{(1)}$, and $\theta_{j.[j_-]}$ is drawn from its approximate posterior distribution;
2. $Y_{j,mis}^{(1)}$ is drawn from $m(Y_j \mid \{Y_i^{(1)}\}_{i=1}^{j-1}, \hat{\theta}_{j.[j_-]})$.

For $t = 2, \ldots, T$:

For the jth variable Y_j, $j = 1, \ldots, p-1$, patterns with missing Y_j are grouped to form different blocks $B_{j.\overline{\mathcal{M}_k^*(j_+)}}$ determined by the missing pattern $\mathcal{M}_k^*(j_+)$ for variables after Y_j.

1. Imputation models specified for each corresponding block

$$m_{j.\overline{\mathcal{M}_k^*(j_+)}} = m(Y_j \mid \boldsymbol{Y}_{[\overline{\mathcal{M}_k^*(j_+)}]}, \theta_{j.\overline{\mathcal{M}_k^*(j_+)}})$$

   are fit to the data in B_{j0}, which is

$$D_{j0}^{(t)} = \{Y_1^{(t)}, \ldots, Y_{j-1}^{(t)}, Y_j, Y_{j+1}^{(t-1)}, \ldots, Y_p^{(t-1)} \mid Y_j \text{ is observed}\},$$

   and each $\theta^{(t)}_{j.\overline{\mathcal{M}_k^*(j_+)}}$ is drawn from its approximate posterior distribution;

2. In each block $B_{j.\overline{\mathcal{M}_k^*(j_+)}}$, $Y_{j,mis}^{(t)}$ is drawn from $m(Y_j \mid \boldsymbol{Y}^{(t)}_{[\overline{\mathcal{M}_k^*(j_+)}]}, \theta^{(t)}_{j.\overline{\mathcal{M}_k^*(j_+)}})$, where $\boldsymbol{Y}^{(t)}_{[\overline{\mathcal{M}_k^*(j_+)}]} = \{\boldsymbol{Y}^{(t)}_{[j-]}, \boldsymbol{Y}_{[\mathcal{O}_k(j+)]}\}$.

Note that only values from the previous iterations are involved in the estimation and imputation for Y_1; only one model needs to be specified for the pth variable Y_p, since there is no missing variable after it in any pattern.

Unlike SRMI, BSRMI is not fully conditionally specified. The key advantage of BSRMI over SRMI is that within one iteration, the imputed values for any record are drawn from a compatible joint distribution but not fully conditional specifications. To be concrete, consider the motivating example; index sets are displayed next to their corresponding missing patterns as follows:

P	Y_1	Y_2	Y_3		$\mathcal{M}$	$\mathcal{O}$
1	✓	✓	✓		$\varnothing$	{1,2,3}
2	✓	✓	?		{3}	{1,2}
3	✓	?	✓		{2}	{1,3}
4	✓	?	?		{2,3}	{1}
5	?	✓	✓		{1}	{2,3}
6	?	✓	?		{1,3}	{2}
7	?	?	✓		{1,2}	{3}

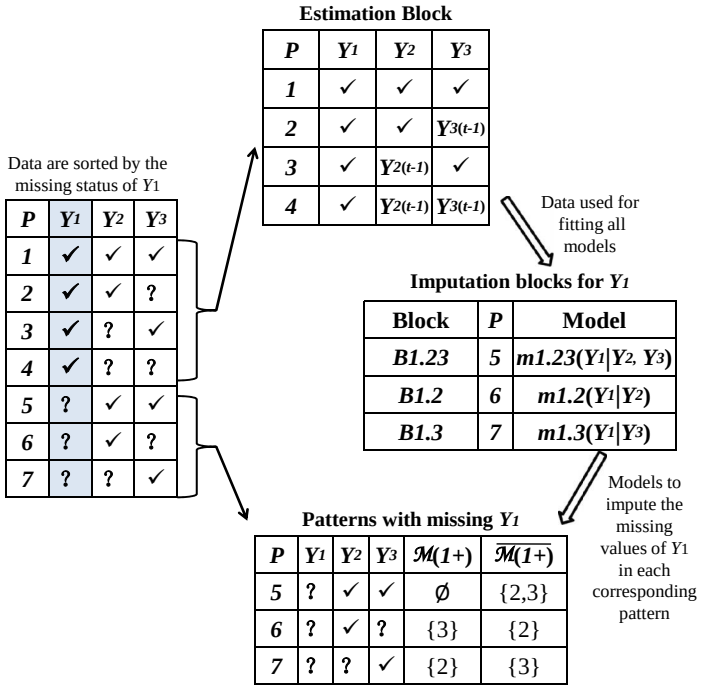

Figure 3.1: Imputing missing values for Y_1 at iteration t.

To impute Y_1 at iteration $t = 2, \ldots, T$:

Imputation models are specified for all missing patterns with missing Y_1. Each pattern belongs to an imputation block determined by predictor index set $\overline{\mathcal{M}(1_+)}$ in that pattern. The models are fit to the estimation block B_{10}, which consists of completed data in patterns $1-4$, and $Y_1^{(t)}$ is drawn from the approximate posterior distribution in each imputation block.

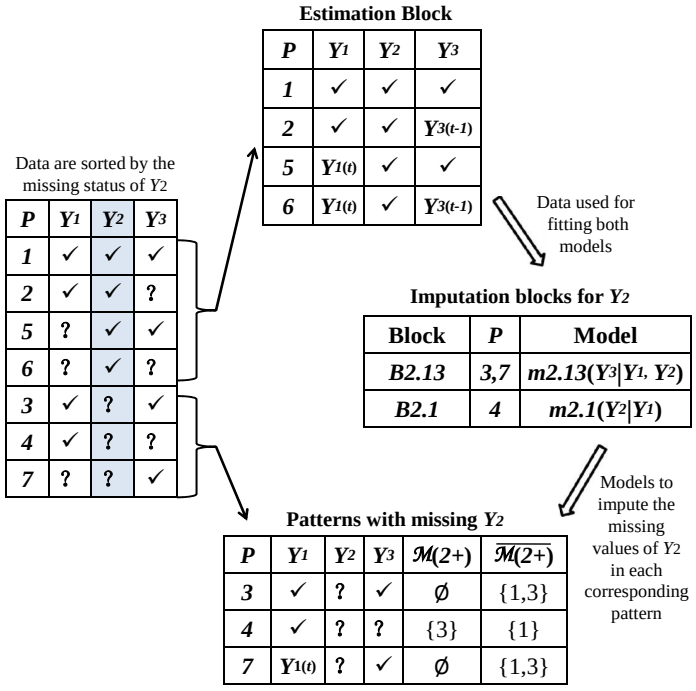

Figure 3.2: Imputing missing values for Y_2 at iteration t.

To impute Y_2: Similarly, imputation models are specified for all missing patterns with missing Y_2. Each pattern belongs to an imputation block determined by predictor index set $\overline{\mathcal{M}(2_+)}$ in that pattern. The models are fit to the estimation block B_{20}, which consists of completed data, and $Y_2^{(t)}$ is drawn from the approximate posterior distribution in each imputation block.

To impute Y_3:

Since Y_3 is the last variable to be imputed in the sequence, all patterns with missing values of Y_3 are grouped into one imputation block for Y_3:

Block	P	Model
$B_{3.12}$	$2, 4, 6$	$m_{3.12}(Y_3 \mid Y_1, Y_2)$

The model is fitted to the completed data in the estimation block B_{30}:

P	Y_1	Y_2	Y_3
1	✓	✓	✓
3	✓	$Y_2^{(t)}$	✓
5	$Y_1^{(t)}$	✓	✓
7	$Y_1^{(t)}$	$Y_2^{(t)}$	✓

$Y_3^{(t)}$ is drawn from the approximate posterior distribution in the imputation block given $Y_1^{(t)}$ and $Y_2^{(t)}$.

Note that if a different imputation order is chosen then all imputation blocks and corresponding models will differ. An example with four variables is provided in the Appendix.

3.3 Convergence Properties

3.3.1 Notation

Suppose that the actual data generation mechanism for $Y_1, \ldots, Y_p$ is a joint distribution $f(y_1, \ldots, y_p \mid \psi)$. The corresponding marginal and conditional distributions are denoted as follows: $\forall\ \mathcal{I}, \mathcal{J} \subset \{1, \ldots, p\}$ and $\mathcal{I} \cap \mathcal{J} = 0$,

$$f(\mathbf{Y}_{[\mathcal{I}]} \mid \psi_{\mathcal{I}}),$$

$$f(\mathbf{Y}_{[\mathcal{I}]} \mid \mathbf{Y}_{[\mathcal{J}]}, \psi_{\mathcal{I}|\mathcal{J}}).$$

Without loss of any generality, we assume that the data set consists of all possible missing patterns, and that the missingness is non-trivial in any pattern; that is, for any pattern $P = k$, the pattern sample size ratio n_k/n tends to a non-zero fraction as total sample size n tends to infinity. We assume that the missing data mechanism is ignorable as in Rubin (1976); that is, $\forall\ \mathcal{M}_{k1} \subseteq \mathcal{M}_k$, conditioning on the observed variables $\boldsymbol{Y}_{[\mathcal{O}_k]}$, the response probability does not depend on the missing variables $\boldsymbol{Y}_{[\mathcal{M}_{k1}]}$:

$$\Pr(P = k \mid \boldsymbol{Y}_{[\mathcal{M}_{k1}]}, \boldsymbol{Y}_{[\mathcal{O}_k]}, \phi) = \Pr(P = k \mid \boldsymbol{Y}_{[\mathcal{O}_k]}, \phi),$$

where ϕ is distinct from population parameter ψ.

The following two properties for an ignorable missing data mechanism are useful in discussing the convergence properties of BSRMI.

1. The the complete data model in pattern $P = k$ follows the conditional distributions given below: For any $\mathcal{M}_{k1}, \mathcal{M}_{k2} \subset \mathcal{M}_k$ and $\mathcal{M}_{k1} \cap \mathcal{M}_{k2} = 0$,

$$\begin{aligned} & \Pr(\boldsymbol{Y}_{[\mathcal{M}_{k1}]} \mid \boldsymbol{Y}_{[\mathcal{M}_{k2}]}, \boldsymbol{Y}_{[\mathcal{O}_k]}, P = k) \\ = & \frac{\Pr(P = k \mid \boldsymbol{Y}_{[\mathcal{M}_{k1}]}, \boldsymbol{Y}_{[\mathcal{M}_{k2}]}, \boldsymbol{Y}_{[\mathcal{O}_k]}) \times \Pr(\boldsymbol{Y}_{[\mathcal{M}_{k1}]}, \boldsymbol{Y}_{[\mathcal{M}_{k2}]}, \boldsymbol{Y}_{[\mathcal{O}_k]})}{\Pr(P = k \mid \boldsymbol{Y}_{[\mathcal{M}_{k2}]}, \boldsymbol{Y}_{[\mathcal{O}_k]}) \times \Pr(\boldsymbol{Y}_{[\mathcal{M}_{k2}]}, \boldsymbol{Y}_{[\mathcal{O}_k]})} \\ = & f(\boldsymbol{Y}_{[\mathcal{M}_{k1}]} \mid \boldsymbol{Y}_{[\mathcal{M}_{k2}]}, \boldsymbol{Y}_{[\mathcal{O}_k]}). \end{aligned}$$

   As a special case, $\forall\ j \in \mathcal{M}_k$,

$$\Pr(Y_j \mid \boldsymbol{Y}_{[\mathcal{M}_k(j_-)]}, \boldsymbol{Y}_{[\mathcal{O}_k]}, P = k) = f(Y_j \mid \boldsymbol{Y}_{[\mathcal{M}_k(j_-)]}, \boldsymbol{Y}_{[\mathcal{O}_k]}).$$

2. Given the complete data, $\forall\ j = 1, \ldots, p$ and $\mathcal{I} \subseteq \{1, \ldots, j-1, j+1, \ldots, p\}$,

$$\begin{aligned} & \Pr(Y_j \text{ is observed} \mid Y_j, \boldsymbol{Y}_{[\mathcal{I}]}) = 1 - \Pr(Y_j \text{ is missing} \mid Y_j, \boldsymbol{Y}_{[\mathcal{I}]}) \\ = & 1 - \Pr(Y_j \text{ is missing} \mid \boldsymbol{Y}_{[\mathcal{I}]}) = \Pr(Y_j \text{ is observed} \mid \boldsymbol{Y}_{[\mathcal{I}]}). \end{aligned}$$

Therefore,

$$\begin{aligned} & \Pr(Y_j \mid \boldsymbol{Y}_{[\mathcal{I}]}, Y_j \text{ is observed}) \\ = & \frac{\Pr(Y_j \text{ is observed} \mid Y_j, \boldsymbol{Y}_{[\mathcal{I}]}) \times \Pr(Y_j, \boldsymbol{Y}_{[\mathcal{I}]})}{\Pr(Y_j \text{ is observed} \mid \boldsymbol{Y}_{[\mathcal{I}]}) \times \Pr(\boldsymbol{Y}_{[\mathcal{I}]})} \\ = & f(Y_j \mid \boldsymbol{Y}_{[\mathcal{I}]}). \end{aligned}$$

Under Property 1, all subsequent conditional imputation model specifications are the same across all patterns. Under property 2, the block with the observed values of a variable can be used to estimate its correctly specified conditional distribution, provided that the missing values of other variables are generated from the actual population conditional distributions.

3.3.2 Asymptotic Approximation of BSRMI

For any variable Y_j, $j = 1, \ldots, p$, patterns with missing Y_j are grouped to form different blocks $B_{j.\overline{\mathcal{M}_k^*(j_+)}}$ that are determined by the missing pattern, $\mathcal{M}_k^*(j_+)$, for variables after Y_j. Imputation models $m_{j.\overline{\mathcal{M}_k^*(j_+)}} = m(Y_j \mid \boldsymbol{Y}_{[\overline{\mathcal{M}_k^*(j_+)}]}, \theta_{j.\overline{\mathcal{M}_k^*(j_+)}})$ are fit to the data in B_{j0}, and each $\theta^{(t)}_{j.\overline{\mathcal{M}_k^*(j_+)}}$ is drawn from its approximate posterior distribution. Then in each block $B_{j.\overline{\mathcal{M}_k^*(j_+)}}$, $Y^{(t)}_{j,mis}$ is drawn from $m(Y_j \mid \boldsymbol{Y}^{(t)}_{[\overline{\mathcal{M}_k^*(j_+)}]}, \theta^{(t)}_{j.\overline{\mathcal{M}_k^*(j_+)}})$.

We simplify the algorithm by using maximum likelihood estimates instead of draws for the parameter estimates in the regression models. This may underestimate the variability but will not affect consistency of the estimates or the convergence properties of the estimates based on the infinite number of imputations.

The estimation block $B_{j0} = \{b \mid b \subseteq [-j]\}$ consists of all patterns where Y_j is observed. We use $\sum_{P=b}$ to denote summation across all records in pattern $P = b$; then the above procedure calculates the log-likelihood function at iteration t for

model $m_{j.\overline{\mathcal{M}_k(j_+)}}$ defined for pattern $P = k$:

$$\begin{aligned} l(\theta_{j.\overline{\mathcal{M}_k(j_+)}} \mid Y_{j,obs}, \boldsymbol{Y}^{(t)}_{[j_-]}, \boldsymbol{Y}^{(t-1)}_{[\mathcal{O}_k(j_+)]}) &= \sum_{B_{j0}} \log[m(Y_j \mid \boldsymbol{Y}^{(t)}_{[j_-]}, \boldsymbol{Y}^{(t-1)}_{[\mathcal{O}_k(j_+)]}, \theta_{j.\overline{\mathcal{M}_k(j_+)}})] \\ &= \sum_{b \in B_{j0}} \sum_{P=b} \log[m(Y_j \mid \boldsymbol{Y}^{(t)}_{[j_-]}, \boldsymbol{Y}^{(t-1)}_{[\mathcal{O}_k(j_+)]}, \theta_{j.\overline{\mathcal{M}_k(j_+)}})] \end{aligned}$$

and estimate the parameter $\theta^{(t)}_{j.\overline{\mathcal{M}_k(j_+)}}$ by solving the score equation

$$s(\theta_{j.\overline{\mathcal{M}_k(j_+)}} \mid Y_{j,obs}, \boldsymbol{Y}^{(t)}_{[j_-]}, \boldsymbol{Y}^{(t-1)}_{[\mathcal{O}_k(j_+)]}) = \partial l(\theta_{j.\overline{\mathcal{M}_k(j_+)}} \mid Y_{j,obs}, \boldsymbol{Y}^{(t)}_{[j_-]}, \boldsymbol{Y}^{(t-1)}_{[\mathcal{O}_k(j_+)]}) / \partial \theta_{j.\overline{\mathcal{M}_k(j_+)}} = 0.$$

Note that the set of variables $\boldsymbol{Y}_{[\mathcal{O}_k(j_+)]}$ are fully observed in pattern $P = k$, but they may be partially missing in patterns from the estimation block; hence $\boldsymbol{Y}^{(t-1)}_{[\mathcal{O}_k(j_+)]}$ is used to denote the values of those variables in the estimation block.

The algorithm stops at the Tth iteration, and the completed data set consists of $\{Y^{(T)}_j\}$ with $\{\theta^{(T)}_{j.\overline{\mathcal{M}_k(j_+)}}\}$. Convergence and consistency are asymptotic properties of these estimates as the sample size, the number of imputations and the number of iterations or sequential updates all tend towards ∞.

When both the sample size and number of imputations are large, components in the score equation given above can be approximated by their asymptotic functional forms. Data in any pattern b in B_{j0} of sample size n_b at iteration t consists of the observed values $\boldsymbol{Y}_{[\mathcal{O}_b]}$ including Y_j and imputed values $\boldsymbol{Y}^{(t)}_{[\mathcal{M}_b(j_-)]}$ and $\boldsymbol{Y}^{(t-1)}_{[\mathcal{M}_b(j_+)]}$, where $\forall\ i \in \mathcal{M}_b(j_-)$,

$$Y^{(t)}_i \sim m(Y_i \mid \boldsymbol{Y}^{(t)}_{[\mathcal{M}_b(i_-)]}, \boldsymbol{Y}_{[\mathcal{O}_b]}, \theta^{(t)}_{i.\overline{\mathcal{M}_b(i_+)}}),$$

and $\forall\ s \in \mathcal{M}_b(j_+)$,

$$Y^{(t-1)}_s \sim m(Y_s \mid \boldsymbol{Y}^{(t-1)}_{[\mathcal{M}_b(s_-)]}, \boldsymbol{Y}_{[\mathcal{O}_b]}, \theta^{(t-1)}_{s.\overline{\mathcal{M}_b(s_+)}}).$$

The joint distribution of $\boldsymbol{Y}_{[\mathcal{O}_b]}$ and $\boldsymbol{Y}^{(t-1)}_{[\mathcal{M}_b(j_+)]}$ in pattern $P = b$ is

$$\begin{aligned}
& d(\boldsymbol{Y}_{[\mathcal{O}_b]}, \boldsymbol{Y}^{(t-1)}_{[\mathcal{M}_b(j_+)]} \mid \theta^{(t-1)}_{[-j]}, P = b) \\
= & \int d(\boldsymbol{Y}^{(t-1)}_{[\mathcal{M}_b(j_+)]} \mid \boldsymbol{Y}^{(t-1)}_{[\mathcal{M}_b(j_-)]}, \boldsymbol{Y}_{[\mathcal{O}_b]}, P = b) \times d(\boldsymbol{Y}^{(t-1)}_{[\mathcal{M}_b(j_-)]} \mid \boldsymbol{Y}_{[\mathcal{O}_b]}, P = b) \\
& \times f(\boldsymbol{Y}_{[\mathcal{O}_b]} \mid P = b)\ \mathrm{d}\boldsymbol{Y}^{(t-1)}_{[\mathcal{M}_b(j_-)]} \\
= & \int \prod_{i \in \mathcal{M}_b(j_-)} m(Y_i^{(t-1)} \mid \boldsymbol{Y}^{(t-1)}_{[\mathcal{M}_b(i_-)]}, \boldsymbol{Y}_{[\mathcal{O}_b]}, \theta^{(t-1)}_{i.\overline{\mathcal{M}_b(i_+)}}) \\
& \times \prod_{s \in \mathcal{M}_b(j_+)} m(Y_s^{(t-1)} \mid \boldsymbol{Y}^{(t-1)}_{[\mathcal{M}_b(s_-)]}, \boldsymbol{Y}_{[\mathcal{O}_b]}, \theta^{(t-1)}_{s.\overline{\mathcal{M}_b(s_+)}}) \\
& \times f(\boldsymbol{Y}_{[\mathcal{O}_b]} \mid P = b)\ \mathrm{d}\boldsymbol{Y}^{(t-1)}_{[\mathcal{M}_b(j_-)]}
\end{aligned}$$

Based on the property that $d(\boldsymbol{Y}^{(t)}_{[\mathcal{M}_b(j_-)]} \mid \boldsymbol{Y}_{[\mathcal{O}_b]}, P = b) \perp\!\!\!\perp d(\boldsymbol{Y}^{(t-1)}_{[\mathcal{M}_b(j_+)]} \mid \boldsymbol{Y}_{[\mathcal{O}_b]}, P = b)$, the overall density function for data in that pattern for Y_i at iteration t is

$$\begin{aligned}
& d(\boldsymbol{Y}^{(t)}_{[\mathcal{M}_b(j_-)]}, \boldsymbol{Y}_{[\mathcal{O}_b]}, \boldsymbol{Y}^{(t-1)}_{[\mathcal{M}_b(j_+)]} \mid \theta^{(t)}_{[j-]}, \theta^{(t-1)}_{[-j]}, P = b) \\
= & \left[\prod_{i \in \mathcal{M}_b(j_-)} m(Y_i^{(t)} \mid \boldsymbol{Y}^{(t)}_{[\mathcal{M}_b(i_-)]}, \boldsymbol{Y}_{[\mathcal{O}_b]}, \theta^{(t)}_{i.\overline{\mathcal{M}_b(i_+)}}) \right] \times d(\boldsymbol{Y}_{[\mathcal{O}_b]}, \boldsymbol{Y}^{(t-1)}_{[\mathcal{M}_b(j_+)]} \mid \theta^{(t-1)}_{[-j]}, P = b)
\end{aligned}$$

Hence, the component in the score function can be approximated by

$$\begin{aligned}
& \tilde{s}_b(\theta_{j.\overline{\mathcal{M}_k(j_+)}} \mid \theta^{(t)}_{[j-]}, \theta^{(t-1)}_{[-j]}) \\
= & \ n_b \int \frac{\partial \log m(Y_j \mid \boldsymbol{Y}^{(t)}_{[j-]}, \boldsymbol{Y}^{(t-1)}_{[\mathcal{O}_k(j_+)]}, \theta_{j.\overline{\mathcal{M}_k(j_+)}})}{\partial \theta_{j.\overline{\mathcal{M}_k(j_+)}}} \\
& \times d(\boldsymbol{Y}^{(t)}_{[\mathcal{M}_b(j_-)]}, \boldsymbol{Y}_{[\mathcal{O}_b]}, \boldsymbol{Y}^{(t-1)}_{[\mathcal{M}_b(j_+)]} \mid \theta^{(t)}_{[j-]}, \theta^{(t-1)}_{[-j]}, P = b) \quad \mathrm{d}\{\boldsymbol{Y}^{(t)}_{[\mathcal{M}_b(j_-)]}, \boldsymbol{Y}_{[\mathcal{O}_b]}, \boldsymbol{Y}^{(t-1)}_{[\mathcal{M}_b(j_+)]}\},
\end{aligned}$$

and the score function can be approximated by

$$\tilde{s}(\theta_{j.\overline{\mathcal{M}_k(j_+)}} \mid \theta^{(t)}_{[j-]}, \theta^{(t-1)}_{[-j]}) = \sum_{b \in B_{j0}} \tilde{s}_b(\theta_{j.\overline{\mathcal{M}_k(j_+)}} \mid \theta^{(t)}_{[j-]}, \theta^{(t-1)}_{[-j]}).$$

Let θ_j denote the vector of model parameters across all the models of Y_j in all blocks. The score vector function $\tilde{s}(\theta^{(t)}_j \mid \theta^{(t)}_{[j-]}, \theta^{(t-1)}_{[-j]})$ converges to 0 in probability as $n \to \infty$, which leads to an approximate iterative algorithm $\tilde{s}(\theta^{(t)}_j \mid \theta^{(t)}_{[j-]}, \theta^{(t-1)}_{[-j]}) =$

0. Therefore, the implicit recursive algorithm $\theta_j^{(t)} = \tilde{s}^{-1}(\theta_{[j-]}^{(t)}, \theta_{[-j]}^{(t-1)})$ has the same convergence property as that of the imputation algorithm.

Theorem 3.1. *Suppose* $\forall\ k = 1, \ldots, K,\ j = 1, \ldots, p,$ *all BSRMI imputation models* $m_{j.\overline{\mathcal{M}_k(j+)}} : m(Y_j \mid \boldsymbol{Y}_{\overline{[\mathcal{M}_k(j+)]}}, \theta_{j.\overline{\mathcal{M}_k(j+)}})$ *are validly specified conditional distributions satisfying the following regularity conditions:*

1. *The density functions are differentiable with respect to* θ_{il} *and the order of differentiation and integration are interchangeable.*
2. *The mean and the variance of the score functions given above exist under both the posited* $\{m_{il}\}$ *and the true population density* $f(Y_1, \ldots, Y_p)$.

Then as the sample size n*, the number of imputations* M *and the number of iterations* t *tend to* ∞*, the regression models* $m(Y_j \mid \boldsymbol{Y}_{[j-]}^{(t)}, \boldsymbol{Y}_{[\mathcal{O}_k(j+)]}^{(t-1)}, \theta_{j.\overline{\mathcal{M}_k(j+)}}^{(t)}) \rightarrow f(Y_j \mid \boldsymbol{Y}_{[j-]}, \boldsymbol{Y}_{[\mathcal{O}_k(j+)]}, \psi)$.

The proof of the theorem is given in Appendix 3.1.

Chapter 2 (Zhu and Raghunathan, 2015) defined a class of fully conditionally specified models as possibly compatible model sequence with separable marginal parameters. For SRMI in the single variable missingness settings, such sequences yield consistent results similar to the imputation under the target joint model. We now extend this in the context of BSRMI.

Theorem 3.2. *Suppose* $\forall\ k = 1, \ldots, K,\ j = 1, \ldots, p,$ *all BSRMI imputation models* $m_{j.\overline{\mathcal{M}_k(j+)}} : m(Y_j \mid \boldsymbol{Y}_{\overline{[\mathcal{M}_k(j+)]}}, \theta_{j.\overline{\mathcal{M}_k(j+)}})$ *follow the exact functional forms of conditional distributions derived from a joint distribution with separable marginal parameters. Suppose the following regularity conditions are met:*

1. *The density functions are differentiable with respect to* θ_{il} *and the order of differentiation and integration are interchangeable.*

2. The mean and the variance of the score functions given above exist under both the posited $\{m_{il}\}$ and the true population density $f(Y_1, \ldots, Y_p)$.

Then as the sample size n, the number of imputations M and the number of iterations t tend to ∞, the regression models $m(Y_j \mid \mathbf{Y}^{(t)}_{[j-]}, \mathbf{Y}^{(t-1)}_{[\mathcal{O}_k(j_+)]}, \theta^{(t)}_{j.\overline{\mathcal{M}_k(j_+)}})$ converge to the corresponding conditional distributions from the estimated joint distribution.

The proof of the theorem is given in Appendix 3.1.

3.4 Simulation Studies

3.4.1 Trivariate Poisson Case

Consider the trivariate count data generated from the trivariate Poisson distribution discussed in Section 3.1. We generated 500 data sets of sample size $n = 500$, 1,000, 5,000 and 10,000. To create missing values, each data set was divided into six random groups:

1. In the first group of size n_1, y_2 and y_3 are observed and

$$\text{pr}(y_1 \text{ is set to be missing} \mid y_2, y_3) = [1 + \exp(1.6 - 0.8y_2 - 0.7y_3)]^{-1};$$

2. In the second group of size n_2, y_1 and y_3 are observed and

$$\text{pr}(y_2 \text{ is set to be missing} \mid y_1, y_3) = [1 + \exp(2.1 - 0.8y_1 - 0.8y_3)]^{-1};$$

3. In the third group of size n_3, y_1 and y_2 are observed and

$$\text{pr}(y_3 \text{ is set to be missing} \mid y_1, y_2) = [1 + \exp(2.3 - 0.9y_1 - 0.8y_2)]^{-1};$$

4. In the fourth group of size n_4, y_3 is observed and

$$\text{pr}(y_1 \text{ and } y_2 \text{ are set to be missing} \mid y_3) = [1 + \exp(1 - 1.2y_3)]^{-1};$$

5. In the fifth group of size n_5, y_1 is observed and

$$\text{pr}(y_2 \text{ and } y_3 \text{ are set to be missing} \mid y_1) = [1 + \exp(1.7 - 1.5y_1)]^{-1};$$

6. In the sixth group of size n_6, y_2 is observed and

$$\text{pr}(y_1 \text{ and } y_3 \text{ are set to be missing} \mid y_2) = [1 + \exp(1.8 - 1.4y_2)]^{-1}.$$

We consider the following two regression modeling specifications:

1. **Poor/Linear Imputation Algorithms:** Poisson regression models with linear predictors without any interaction.

2. **Improved Imputation Algorithms:** We generated a single complete data set of size 500 and set the values to missing using the above mechanism. Based on empirical examination of this data set, we determined several reasonably fitting models for each variable for each of the six orders of imputation (permutations of (1,2 3)). This may be considered as a real analysis of a single data set. The models developed were then used on the 500 data sets.

3. For comparison purposes, we also consider SRMI using the correctly specified model sequences, through deriving and drawing from the actual conditional distributions $f(y_1|y_2, y_3), f(y_2|y_1, y_3)$ and using the correct $f(y_3|y_1, y_2)$.

The following are the reasonably fitting models for all conditional distributions needed, if the order of the variables is Y_1, Y_2 and Y_3.

1. For y_1, we need three regression models, $m_{1.23}$, $m_{1.2}$ and $m_{1.3}$. Our data analysis of the single data set suggested the following models would be a good fit:

(a) $m_{1.23}(y_1 \mid y_2, y_3, \theta_{10}) \sim \text{Poisson}(\lambda = \exp(\theta_{100} + \theta_{101}y_2^{1/3} + \theta_{102}1/(1+y_2)^3 + \theta_{103}y_3^{(1/3)} + \theta_{104}1/(1+y_3)^3 + \theta_{105}\log(y3+1) * 1_{(y3<5)} + \theta_{106}y_2^{1/3} * y_3^{1/3}))$;

(b) $m_{1.2}(y_1 \mid y_2, \theta_{12}) \sim \text{Poisson}(\lambda = \exp(\theta_{120} + \theta_{121} y_2^{1/3} + \theta_{122} 1/(1+y_2)^3))$

(c) $m_{1.3}(y_1 \mid y_3, \theta_{13}) \sim \text{Poisson}(\lambda = \exp(\theta_{130} + \theta_{131} y_3^{(1/3)} + \theta_{132} 1/(1+y_3)^3 + \theta_{133} \log(y3+1) * 1_{(y3<5)}))$;

2. For y_2, the two models were $m_{2.13}(y_2 \mid y_1, y_3, \theta_{20}) \sim \text{Poisson}(\lambda = \exp(\theta_{200} + \theta_{201} y_1 + \theta_{203} y_3))$; and $m_{2.1}(y_2 \mid y_1, \theta_{21}) \sim \text{Poisson}(\lambda = \exp(\theta_{210} + \theta_{211} y_1))$;

3. For y_3, $m_{3.12}(y_3 \mid y_1, y_2, \theta_3) \sim \text{Poisson}(\lambda = \exp(\theta_{30} + \theta_{31} y_1 + \theta_{32} y_2))$.

We created $M = 20$ multiple imputations for each data set using iteration numbers $T = 10, 100$ and 1000. To monitor the convergence status of each algorithm, we compare the multiple imputation estimates of $(\alpha_{10}, \alpha_{20}, \alpha_{21},\ \alpha_{30}, \alpha_{31}, \alpha_{32})$ for each of the three imputation algorithms (poor-fitting, well-fitting and actual). We also compare the Kullback-Leibler divergence between the "before deletion data" estimate and "after imputation data" estimates of the joint distribution of (Y_1, Y_2, Y_3).

Table 3.1 provides the Monte-Carlo biases, differences between the means of the 500 parameter estimates (from 500 data sets) and the true values for each parameter. We use the means of estimates from the before-deletion data as the reference. The maximum biases are in the estimates of α_{31} (about 25%) and α_{30} (about 16%), and others are relatively small. However, for each parameter, the biases are similar across all orders. This indicates that given reasonably fitting models, the effect of imputation order is small.

The Kullback-Leibler divergence between before-Deletion and after-imputation data distributions also shows that the algorithms converge to the same results as the sample size, number of iterations and number of imputations increase. It further shows that better results are obtained using the improved imputation model results. In this problem, it is often difficult to obtain well-fitting models within the Poisson

Table 3.1: Bias in the parameter estimates based on BSRMI using the improved algorithm under different orders, Replicates=500, Imputations=20, Sample size=10,000

	Order					
	(1,2,3)	(1,3,2)	(2,1,3)	(2,3,1)	(3,2,1)	(3,1,2)
$\alpha_{10} = \log(2)$	-0.017	-0.019	-0.016	-0.015	-0.029	-0.026
$\alpha_{20} = -1$	0.061	0.102	0.071	0.092	0.113	0.082
$\alpha_{21} = 0.3$	-0.019	-0.018	-0.041	-0.038	-0.022	-0.021
$\alpha_{30} = -2$	0.331	0.357	0.312	0.333	0.321	0.324
$\alpha_{31} = 0.5$	-0.115	-0.121	-0.135	-0.142	-0.113	-0.112
$\alpha_{32} = 0.1$	0.012	0.013	0.024	0.026	0.022	0.024

family because the model diagnostics indicate that mean and variance functions are not the same for many conditional distributions.

3.4.2 Complex Case

For our second simulation study, we consider the situation discussed in Chen et al. (2011) where SRMI based on routinely chosen (and rather poor-fitting) imputation models performed badly compared to some other imputation methods, such predictive mean matching. The simulated data has the following fully observed variables: a normal variable Y_1, a binary variable Y_2 related to Y_1, and a normal variable Y_3 related to Y_1 and Y_2. The partially missing variables are a Poisson count variable Y_4 related to Y_2, a normal variable Y_5 related to $\{Y_2,Y_3,Y_4\}$, and a binary variable Y_6 related to $\{Y_1,Y_2,Y_4\}$. Some values were deleted using an ignorable missing data mechanism. We used the same parameters as in Chen et al. (2011) to compare the bias, mean square error and confidence coverage. As in the trivariate Poisson case, we used a simulated data set with missing values to develop well-fitting models for both SRMI and BSRMI. For poor-fitting models, we used the standard GLM models (linear, logistic or Poisson with just main effects). Table 3.2 provides the results for SRMI and BSRMI using poor- and improved- fitting models. Clearly, improved model results are markedly better than the corresponding routine or poor-

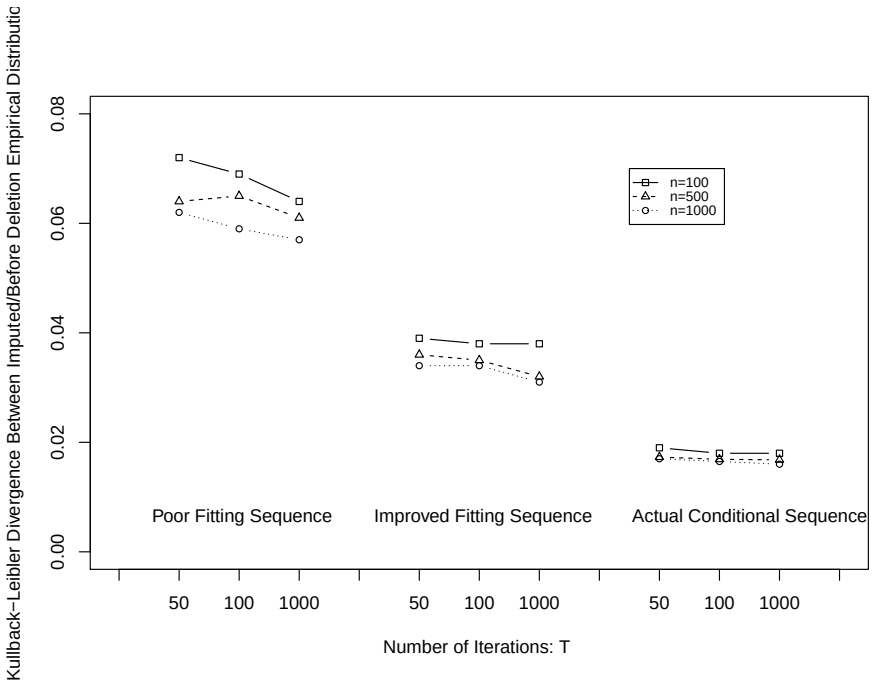

Figure 3.3: Kullback-Leibler divergence between empirical distributions, based on multiply imputed data $\widehat{F}^{n,T}_{MI}(x,y)$ and before deletion data $\widehat{F}^{n}_{BD}(x,y)$ and averaged over 500 replicates of data sets, from three imputation algorithms plotted as a function of sample size n=500, 1000, and 5000 and the number of iterations $T = 50, 100, 1000$ when the data are missing at random. The three imputation algorithms impute the missing values in the order of Y_1, Y_2 and Y_3: the poor sequence assumes linear predictors only, the improved sequence assumes appropriate non-linear terms of the predictors, and the validly specified sequence assumes true conditional models from the data population.

fitting models. Furthermore, the repeated sampling properties of BSRMI estimates are better than the SRMI estimates.

While it is difficult to find the perfectly fitting imputation models in some cases, including this example, we believe that proper model diagnostics are always needed for any imputation procedure.

Table 3.2: Performance of four multiple imputation algorithms on the example in Chen et al. (2011), R=500, M=5

	MICE[2]			SRMI Poor			BSRMI Poor			SRMI Improved			BSRMI Improved		
	Bias	SE	CR	Bias	SE	CR	Bias	SE	CR	Bias	SE	CR	Bias	SE	CR
$\beta_{41} = 0$	-0.02	0.09	94	-0.03	0.08	96	-0.04	0.08	93	0.01	0.11	96	0.00	0.09	94
$\beta_{42} = 1$	-0.08	0.16	93	-0.04	0.15	94	-0.01	0.15	100	-0.14	0.18	88	-0.11	0.17	97
$\beta_{43} = 0$	0.04	0.04	86	0.00	0.04	94	0.00	0.04	96	-0.05	0.06	72	-0.01	0.04	94
$\beta_{51} = 0$	-0.08	0.22	99	-0.11	0.26	92	-0.16	0.26	93	0.01	0.23	98	0.01	0.24	100
$\beta_{52} = -1$	-1.07	0.53	72	-0.48	0.53	92	-0.39	0.54	100	-0.43	0.50	96	-0.42	0.49	100
$\beta_{53} = 1$	1.13	0.32	**5**	0.79	0.21	**6**	0.78	0.22	**14**	0.15	0.22	**94**	0.18	0.20	**94**
$\beta_{54} = 0$	0.15	0.13	90	-0.01	0.13	98	-0.03	0.13	86	0.03	0.13	100	0.02	0.11	100
$\beta_{55} = 1$	-0.44	0.13	**15**	-0.35	0.07	**2**	-0.36	0.07	**7**	-0.15	0.09	**76**	-0.14	0.08	**84**
$\beta_{61} = -1$	-0.01	0.27	94	0.00	0.27	94	0.03	0.27	93	-0.00	0.28	92	0.00	0.27	94
$\beta_{62} = 0$	-0.11	0.56	97	-0.02	0.23	94	-0.04	0.24	93	0.03	0.24	94	0.07	0.24	97
$\beta_{63} = 0$	-0.00	0.20	97	-0.23	0.64	92	-0.19	0.63	86	-0.01	0.66	94	-0.02	0.66	90
$\beta_{64} = 1$	0.02	0.30	96	-0.15	0.29	88	-0.18	0.30	96	-0.14	0.30	88	-0.14	0.30	97
$\beta_{65} = 0$	-0.00	0.05	98	-0.02	0.06	96	0.02	0.06	96	0.00	0.06	92	-0.01	0.07	94
$\beta_{66} = -1$	0.01	0.30	97	0.23	0.32	90	0.25	0.33	93	0.13	0.34	92	0.12	0.35	90

[2]Results from Chen et al. (2011)

3.5 Discussion

When several variables have missing values, then poorly fitting models used in a regular SRMI may lead to algorithmic incompatibility, resulting in divergence of the algorithm and invalid multiple imputation estimates. We used the trivariate Poisson example to illustrate such a situation. Thus, a modification to assure convergence leads to BSRMI, which models a collection of the joint distribution of the subset of variables (within blocks), and thus assures compatibility within patterns. This approach retains the flexibility in SRMI but avoids its pitfalls.

Similar to regular SRMI, the performance of block SRMI algorithms depends on the model fit. Our simulation examples show that better fitting imputation models lead to better inference. It allows an imputer to focus on improving the goodness-of-fit and predictive power of the imputation models. The trade-off is the increase in the number of models to be specified and estimated in the sequence.

The increase in the number of models specified in the sequence also has its advantage, as it allows imputers to further improve the imputation models to get better imputation results. The block SRMI is also order dependent if model specification varies dramatically in different sequences. One may choose an order that results in the best fitting models. However, if the model sequences are well fitting for both sequences, our simulation results also show that the effect of ordering is minimal.

We assume an ignorable missing data mechanism. The Block SRMI applied on nonignorable missing data will also converge but the performance is hard to assess due to the influence of the missing mechanism. However, the block-specific SRMI can easily adapt the nonignorable missing data into its framework due to its flexibility of specification. For example, by formulating models that differ by pattern, a

pattern mixture model analog of BSRMI could be developed. We suspect that if the imputation models and missingness mechanism are validly specified in the nonignorable block-specific SRMI, it will yield valid inferences. Further research is needed to extend BSRMI for nonignorable missing data mechanisms.

Appendix 3.1

Part 1: Proof of Theorem 3.1

Proof: First, consider the case where the BSRMI imputation models are exactly specified. That is, the regression models $m_{j.\overline{\mathcal{M}_k(j_+)}} : m(Y_j \mid \boldsymbol{Y}_{[\overline{\mathcal{M}_k(j_+)}]}, \theta_{j.\overline{\mathcal{M}_k(j_+)}})$ follow the exact functional forms of corresponding $f(Y_j \mid \boldsymbol{Y}_{[\overline{\mathcal{M}_k(j_+)}]}, \psi_{j.\overline{\mathcal{M}_k(j_+)}})$, where $\psi_{j.\overline{\mathcal{M}_k(j_+)}}$ is the corresponding conditional parameter set derived from the population parameter ψ.

When the BSRMI imputation models are validly specified, there exists a parameterization for all models: $\theta_{j.\overline{\mathcal{M}_k(j_+)}} = (\zeta_{j.\overline{\mathcal{M}_k(j_+)}}, \xi_{j.\overline{\mathcal{M}_k(j_+)}})$ such that the regression models $m_{j.\overline{\mathcal{M}_k(j_+)}} : m(Y_j \mid \boldsymbol{Y}_{[\overline{\mathcal{M}_k(j_+)}]}, \zeta_{j.\overline{\mathcal{M}_k(j_+)}}, \xi_{j.\overline{\mathcal{M}_k(j_+)}} = 0)$ follow the exact functional forms of corresponding $f(Y_j \mid \boldsymbol{Y}_{[\overline{\mathcal{M}_k(j_+)}]}, \psi_{j.\overline{\mathcal{M}_k(j_+)}})$.

The proof for correctly specified models is divided into two steps. First, we show that given $\theta^{(t-1)}_{j.\overline{\mathcal{M}_k(j_+)}} = \psi_{j.\overline{\mathcal{M}_k(j_+)}}$, $\forall\ 1 \le j \le p,\ \forall\ 1 \le k \le K$, the completed data follow the asymptotic distribution as the before deletion values in any pattern. At iteration $t-1$, $\forall\ b$ with $\mathcal{M}_b = \{I_1, \ldots, I_{p_b}\}$, the missing pattern $P = b$ is completed with observed values of $\boldsymbol{Y}_{[\mathcal{O}_b]}$ and imputed values of $\boldsymbol{Y}_{[\mathcal{M}_b]}$ drawn from below:

$$
\begin{aligned}
& f\left(Y^{(t-1)}_{I_1} \mid \boldsymbol{Y}_{[\mathcal{O}_b]}, \psi_{I_1.\overline{\mathcal{M}_b(I_{1+})}}\right), \\
& \ldots \\
& f\left(Y^{(t-1)}_{I_i} \mid Y^{(t-1)}_{I_1}, \ldots, Y^{(t-1)}_{I_{i-1}}, \boldsymbol{Y}_{[\mathcal{O}_b]}, \psi_{I_i.\overline{\mathcal{M}_b(I_{i+})}}\right), \\
& \ldots \\
& f\left(Y^{(t-1)}_{I_{p_b}} \mid Y^{(t-1)}_{I_1}, \ldots, Y^{(t-1)}_{I_{p_b-1}}, \boldsymbol{Y}_{[\mathcal{O}_b]}, \psi_{I_{p_b}.\overline{\mathcal{M}_b(I_{p_b+})}}\right).
\end{aligned}
$$

Then the asymptotic density function for $\boldsymbol{Y}_{[\mathcal{M}_b]}^{(t-1)} \mid \boldsymbol{Y}_{[\mathcal{O}_b]}$ is

$$\prod_{1\le i\le p_b} f\left(Y_{I_i} \mid Y_{I_1},\ldots,Y_{I_{i-1}},\boldsymbol{Y}_{[\mathcal{O}_b]},\psi_{I_i.\overline{\mathcal{M}_b(I_{i+})}}\right),$$

and the overall asymptotic joint density function for the completed data in pattern b is

$$\begin{aligned} & m(\boldsymbol{Y}_{[\mathcal{M}_b]},\boldsymbol{Y}_{[\mathcal{O}_b]} \mid P=b) \\ = \;& f(\boldsymbol{Y}_{[\mathcal{O}_b]} \mid P=b)\times \prod_{1\le i\le p_b} f\left(Y_{I_i} \mid Y_{I_1},\ldots,Y_{I_{i-1}},\boldsymbol{Y}_{[\mathcal{O}_b]},\psi_{I_i.\overline{\mathcal{M}_b(I_{i+})}}\right) \\ = \;& f(\boldsymbol{Y}_{[\mathcal{O}_b]} \mid P=b)\times f(\boldsymbol{Y}_{[\mathcal{M}_b]} \mid \boldsymbol{Y}_{[\mathcal{O}_b]}). \end{aligned}$$

As a special case of property 1,

$$f(\boldsymbol{Y}_{[\mathcal{M}_b]} \mid \boldsymbol{Y}_{[\mathcal{O}_b]},P=b)=f(\boldsymbol{Y}_{[\mathcal{M}_b]} \mid \boldsymbol{Y}_{[\mathcal{O}_b]}),$$

then the asymptotic distribution for completed and before-deletion are the same:

$$m(\boldsymbol{Y}_{[\mathcal{M}_b]},\boldsymbol{Y}_{[\mathcal{O}_b]} \mid P=b)=f(\boldsymbol{Y}_{[\mathcal{M}_b]},\boldsymbol{Y}_{[\mathcal{O}_b]} \mid P=b).$$

Next, we prove by induction that given the imputed values based on $\{\theta_{j.\overline{\mathcal{M}_k(j_+)}}^{(t-1)} = \psi_{j.\overline{\mathcal{M}_k(j_+)}}\}$ as described above, the parameter estimation of the imputation model $m(Y_j \mid \boldsymbol{Y}_{[\overline{\mathcal{M}_k(j_+)}]},\theta_{j.\overline{\mathcal{M}_k(j_+)}}))$ based on values from the estimation block B_{j0} yields $\theta_{j.\overline{\mathcal{M}_k(j_+)}}^{(t)} = \psi_{j.\overline{\mathcal{M}_k(j_+)}}$. For $j=1$, at iteration t, step 1 above indicates that the completed data in the estimation block $D_{10}^{(t)} = \{Y_1, Y_2^{(t-1)},\ldots,Y_p^{(t-1)} \mid Y_1 \text{ is observed}\}$ follow the same distribution as the before-deletion data in B_{10} which is $f(\boldsymbol{Y} \mid Y_1 \text{ is observed})$; then the asymptotic score vector function for Y_1 specified for any

pattern $P = k$ with missing Y_1 is

$$\begin{aligned}
&\tilde{s}(\theta_{1.\overline{\mathcal{M}_k(1_+)}} \mid \theta^{(t-1)}_{[-1]}) \\
= \;& n_{B_{10}} \int \frac{\partial \log m(Y_1 \mid \boldsymbol{Y}^{(t-1)}_{[\mathcal{O}_k(1_+)]}, \theta_{1.\overline{\mathcal{M}_k(1_+)}})}{\partial \theta_{1.\overline{\mathcal{M}_k(1_+)}}} \times d(\boldsymbol{Y}^{(t-1)} \mid Y_1 \text{ is observed}) \quad \mathrm{d}\boldsymbol{Y}^{(t-1)} \\
= \;& n_{B_{10}} \int \frac{\partial \log f(Y_1 \mid \boldsymbol{Y}_{[\mathcal{O}_k(1_+)]}, \theta_{1.\overline{\mathcal{M}_k(1_+)}})}{\partial \theta_{1.\overline{\mathcal{M}_k(1_+)}}} \times f(\boldsymbol{Y} \mid Y_1 \text{ is observed}) \quad \mathrm{d}\boldsymbol{Y} \\
= \;& n_{B_{10}} \int \left[\int \frac{\partial \log f(Y_1 \mid \boldsymbol{Y}_{[\mathcal{O}_k(1_+)]}, \theta_{1.\overline{\mathcal{M}_k(1_+)}})}{\partial \theta_{1.\overline{\mathcal{M}_k(1_+)}}} \times f(Y_1 \mid \boldsymbol{Y}_{[\mathcal{O}_k(1_+)]}, \psi_{1.\overline{\mathcal{M}_k(1_+)}}) \mathrm{d}Y_1 \right] \\
& \times f(\boldsymbol{Y}_{[\mathcal{O}_k(1_+)]} \mid Y_1 \text{ is observed}) \quad \mathrm{d}\boldsymbol{Y}_{[\mathcal{O}_k(1_+)]},
\end{aligned}$$

and therefore $\tilde{s}(\theta^{(t)}_{1.\overline{\mathcal{M}_k(1_+)}} = \psi_{1.\overline{\mathcal{M}_k(1_+)}} \mid \theta^{(t-1)}_{[-1]}) = 0$ for all k.

Following a similar approach as in step 1, it can be shown that $D^{(t)}_{20}$ also follows the before-deletion distribution in B_{20}, which is $f(\boldsymbol{Y} \mid Y_2 \text{ is observed})$. Then for $j = 2, \ldots, p$, assume $\tilde{s}(\theta^{(t)}_{j-1} = \psi_{j-1} \mid \theta^{(t)}_{[(j-1)-]}, \theta^{(t-1)}_{[-(j-1)]}) = 0$ and $D^{(t)}_{j0}$ follows the before-deletion distribution; we now show that $\tilde{s}(\theta^{(t)}_j = \psi_j \mid \theta^{(t)}_{[j-]}, \theta^{(t-1)}_{[-j]}) = 0$.

Note that,

$$\begin{aligned}
&\tilde{s}(\theta_{j.\overline{\mathcal{M}_k(j_+)}} \mid \theta^{(t)}_{[j-]}, \theta^{(t-1)}_{[-j]}) \\
= \;& n_{B_{j0}} \int \frac{\partial \log m(Y_j \mid \boldsymbol{Y}^{(t)}_{[j-]}, \boldsymbol{Y}^{(t-1)}_{[\mathcal{O}_k(j_+)]}, \theta_{j.\overline{\mathcal{M}_k(j_+)}})}{\partial \theta_{j.\overline{\mathcal{M}_k(j_+)}}} \times d(\boldsymbol{Y}^{(t)}_{[j-]}, \boldsymbol{Y}^{(t-1)}_{[\mathcal{O}_k(j_+)]} \mid B_{j0}) \; \mathrm{d}\{\boldsymbol{Y}^{(t)}_{[j-]}, \boldsymbol{Y}^{(t-1)}_{[\mathcal{O}_k(j_+)]}\} \\
= \;& n_{B_{j0}} \int \frac{\partial \log f(Y_j \mid \boldsymbol{Y}_{[j-]}, \boldsymbol{Y}_{[\mathcal{O}_k(j_+)]}, \theta_{j.\overline{\mathcal{M}_k(j_+)}})}{\partial \theta_{j.\overline{\mathcal{M}_k(j_+)}}} \times f(\boldsymbol{Y} \mid Y_j \text{ is observed}) \quad \mathrm{d}\boldsymbol{Y}, \\
= \;& n_{B_{j0}} \int \left[\int \frac{\partial \log f(Y_j \mid \boldsymbol{Y}_{[j-]}, \boldsymbol{Y}_{[\mathcal{O}_k(j_+)]}, \theta_{j.\overline{\mathcal{M}_k(j_+)}})}{\partial \theta_{j.\overline{\mathcal{M}_k(j_+)}}} \times f(Y_j \mid \boldsymbol{Y}_{[j-]}, \boldsymbol{Y}_{[\mathcal{O}_k(j_+)]}, \psi_{j.\overline{\mathcal{M}_k(j_+)}}) \mathrm{d}Y_j \right] \\
& \times f(\boldsymbol{Y}_{[j-]}, \boldsymbol{Y}_{[\mathcal{O}_k(j_+)]} \mid Y_j \text{ is observed}) \quad \mathrm{d}\{\boldsymbol{Y}_{[j-]}, \boldsymbol{Y}_{[\mathcal{O}_k(j_+)]}\},
\end{aligned}$$

and then $\tilde{s}(\theta^{(t)}_{j.\overline{\mathcal{M}_k(j_+)}} = \psi_{j.\overline{\mathcal{M}_k(j_+)}} \mid \theta^{(t)}_{[j-]}, \theta^{(t-1)}_{[-j]}) = 0$.

We now consider the validly specified model sequences with extra terms compared to the true conditional densities. We need to show that $\theta^*_{j.\overline{\mathcal{M}_k(j_+)}} = (\psi_{j.\overline{\mathcal{M}_k(j_+)}}, \xi_{j.\overline{\mathcal{M}_k(j_+)}} =$

0) is the convergent point for each $\theta_{j.\overline{\mathcal{M}_k(j_+)}}$. Given $\theta^*_{[-j]} = (\psi_{[-j]}, 0)$,

$$
\begin{aligned}
& \tilde{s}(\theta_{j.\overline{\mathcal{M}_k(j_+)}} \mid (\psi_{[j-]}, 0), (\psi_{[-j]}, 0)) \\
= \;& \tilde{s}(\zeta_{j.\overline{\mathcal{M}_k(j_+)}}, \xi_{j.\overline{\mathcal{M}_k(j_+)}} \mid (\psi_{[j-]}, 0), (\psi_{[-j]}, 0)) \\
= \;& n_{B_{j0}} \int \left[\int \frac{\partial \log m_{j.\overline{\mathcal{M}_k(j_+)}}(Y_j \mid \boldsymbol{Y}_{[j-]}, \boldsymbol{Y}_{[\mathcal{O}_k(j_+)]}, \theta_{j.\overline{\mathcal{M}_k(j_+)}})}{\partial \theta_{j.\overline{\mathcal{M}_k(j_+)}}} \times f(Y_j \mid \boldsymbol{Y}_{[j-]}, \boldsymbol{Y}_{[\mathcal{O}_k(j_+)]}, \psi_{j.\overline{\mathcal{M}_k(j_+)}}) \mathrm{d}Y_j \right] \\
& \qquad \times f(\boldsymbol{Y}_{[j-]}, \boldsymbol{Y}_{[\mathcal{O}_k(j_+)]} \mid Y_j \text{ is observed}) \quad \mathrm{d}\{\boldsymbol{Y}_{[j-]}, \boldsymbol{Y}_{[\mathcal{O}_k(j_+)]}\}.
\end{aligned}
$$

Since maximizing the likelihood is equivalent to minimizing the relative entropy of the posited regression model relative to the true distribution, finding the solution for $\tilde{s}(\theta_{j.\overline{\mathcal{M}_k(j_+)}} \mid (\psi_{[j-]}, 0), (\psi_{[-j]}, 0)) = 0$ is equivalent to minimizing $\iint \log[f(Y_j \mid \boldsymbol{Y}_{[j-]}, \boldsymbol{Y}_{[\mathcal{O}_k(j_+)]}, \psi_{j.\overline{\mathcal{M}_k(j_+)}}) / m_{j.\overline{\mathcal{M}_k(j_+)}}(Y_j \mid \boldsymbol{Y}_{[j-]}, \boldsymbol{Y}_{[\mathcal{O}_k(j_+)]}, \theta_{j.\overline{\mathcal{M}_k(j_+)}})] \times f(Y_j \mid \boldsymbol{Y}_{[j-]}, \boldsymbol{Y}_{[\mathcal{O}_k(j_+)]}, \psi_{j.\overline{\mathcal{M}_k(j_+)}}) \mathrm{d}Y_j \times f(\boldsymbol{Y}_{[j-]}, \boldsymbol{Y}_{[\mathcal{O}_k(j_+)]} \mid Y_j \text{ is observed}) \quad \mathrm{d}\{\boldsymbol{Y}_{[j-]}, \boldsymbol{Y}_{[\mathcal{O}_k(j_+)]}\}$.

The relative entropy has non-negative values and its minimum 0 is reached if and only if $m_{j.\overline{\mathcal{M}_k(j_+)}}(Y_j \mid \boldsymbol{Y}_{[j-]}, \boldsymbol{Y}_{[\mathcal{O}_k(j_+)]}, \theta_{j.\overline{\mathcal{M}_k(j_+)}} = (\psi_{j.\overline{\mathcal{M}_k(j_+)}}, 0)) = f(Y_j \mid \boldsymbol{Y}_{[j-]}, \boldsymbol{Y}_{[\mathcal{O}_k(j_+)]}, \psi_{j.\overline{\mathcal{M}_k(j_+)}})$. Therefore, the asymptotic score equation $\tilde{s}(\theta_{j.\overline{\mathcal{M}_k(j_+)}} \mid (\psi_{[j-]}, 0), (\psi_{[-j]}, 0)) = 0$ holds at $(\theta^*_{j.\overline{\mathcal{M}_k(j_+)}}, \theta^*_{[-j]})$. Similar arguments apply to all other models.

Part 2: Proof of Theorem 3.2

Proof: To determine the target to which the approximate algorithm converges, we first apply the target joint model $m(y_1, \ldots, y_p \mid \theta)$ to analyze the incomplete data, where $\theta \in \Theta_C$ is the compatible condition and $\theta = (\theta_{1.[-1]}, \ldots, \theta_{p.[-p]}, \theta \in \Theta_C)$.

Since the joint model has separable marginal parameters, for any pattern k and variable j, we can reparametrize

$$
\theta = (\theta_{[-\mathcal{M}_k(j_+)]},\ \theta_{\mathcal{M}_k(j_+).[-\mathcal{M}_k(j_+)]}) = (\theta_{j.\overline{\mathcal{M}_k(j_+)}},\ \theta_{\overline{\mathcal{M}_k(j_+)}},\ \theta_{\mathcal{M}_k(j_+).[-\mathcal{M}_k(j_+)]}).
$$

We use the expectation-maximization algorithm to obtain the maximum likelihood

estimate θ^*. The expectation step calculates

$$Q(\theta \mid \theta^{(t-1)}) = \sum_{P=0} \log m(y_1, \ldots, y_p \mid \theta) + \sum_{p=1}^{K} \sum_{P=p} \int \log m(y_1, \ldots, y_p \mid \theta) m(y_{[\mathcal{M}_p]} \mid y_{[\mathcal{O}_p]}, \theta^{(t-1)}) \mathrm{d}y_{[\mathcal{M}_p]},$$

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

$$\theta^{(t)} = \arg\max_\theta \, Q(\theta \mid \theta^{(t-1)}).$$

The expected step can be approximated by an asymptotic quantity

$$\tilde{Q}(\theta \mid \theta^{(t-1)}) = n_0 \int \log m(y_1, \ldots, y_p \mid \theta) f(y \mid P = 0) \, \mathrm{d}y + \sum_{p=1}^{K} n_p \int \left\{ \int \log m(y_1, \ldots, y_p \mid \theta) m(y_{[\mathcal{M}_p]} \mid y_{[\mathcal{O}_p]}, \theta^{(t-1)}) \, \mathrm{d}y_{[\mathcal{M}_p]} \right\} f(y_{[\mathcal{O}_p]} \mid P = p) \, \mathrm{d}y_{[\mathcal{O}_p]},$$

and the maximization step maximizes the asymptotic quantity.

Since θ^* is the convergent point for the asymptotic expectation-maximization algorithm, for any parameterization

$$\theta = (\theta_{[-\mathcal{M}_k(j_+)]}, \, \theta_{\mathcal{M}_k(j_+).[-\mathcal{M}_k(j_+)]}) = (\theta_{j.\overline{\mathcal{M}_k(j_+)}}, \, \theta_{\overline{\mathcal{M}_k(j_+)}}, \, \theta_{\mathcal{M}_k(j_+).[-\mathcal{M}_k(j_+)]}),$$

score equations hold at the convergent point because the marginal parameters are separable:

$$\left. \frac{\partial \tilde{Q}(\theta_{j.\overline{\mathcal{M}_k(j_+)}}, \, \theta^*_{\overline{\mathcal{M}_k(j_+)}}, \, \theta^*_{\mathcal{M}_k(j_+).[-\mathcal{M}_k(j_+)]} \mid \theta^*)}{\partial \theta_{j.\overline{\mathcal{M}_k(j_+)}}} \right|_{\theta^*_{j.\overline{\mathcal{M}_k(j_+)}}} = 0.$$

We now show that the maximum likelihood estimate θ^* is also the fixed point of the asymptotic sequential regression imputation algorithm. 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}
&\partial\tilde{Q}(\theta_{j.\overline{\mathcal{M}_k(j_+)}},\ \theta^*_{\overline{\mathcal{M}_k(j_+)}},\ \theta^*_{\mathcal{M}_k(j_+).[-\mathcal{M}_k(j_+)]} \mid \theta^*)\ /\ \partial\theta_{j.\overline{\mathcal{M}_k(j_+)}} \\
&= n_0 \int \frac{\partial \log m(y_1,\ldots,y_p \mid \theta)}{\partial \theta_{j.\overline{\mathcal{M}_k(j_+)}}} f(y \mid P=0)\ \mathrm{d}y \\
&\quad + \sum_{j\notin\mathcal{M}_p} n_p \int \left\{ \int \frac{\partial \log m(y_1,\ldots,y_p \mid \theta)}{\partial \theta_{j.\overline{\mathcal{M}_k(j_+)}}} m(y_{[\mathcal{M}_p]} \mid y_{[\mathcal{O}_p]}, \theta^*)\ \mathrm{d}y_{[\mathcal{M}_p]} \right\} f(y_{[\mathcal{O}_p]} \mid P=p)\ \mathrm{d}y_{[\mathcal{O}_p]} \\
&\quad + \sum_{j\in\mathcal{M}_p} n_p \int \left\{ \int \frac{\partial \log m(y_1,\ldots,y_p \mid \theta)}{\partial \theta_{j.\overline{\mathcal{M}_k(j_+)}}} m(y_{[\mathcal{M}_p]} \mid y_{[\mathcal{O}_p]}, \theta^*)\ \mathrm{d}y_{[\mathcal{M}_p]} \right\} f(y_{[\mathcal{O}_p]} \mid P=p)\ \mathrm{d}y_{[\mathcal{O}_p]} \\
&= n_0 \int \frac{\partial \log m(y_j \mid y_{[\overline{\mathcal{M}_k(j_+)}]}, \theta_{j.\overline{\mathcal{M}_k(j_+)}})}{\partial \theta_{j.\overline{\mathcal{M}_k(j_+)}}} f(y \mid P=0)\ \mathrm{d}y \\
&\quad + \sum_{j\notin\mathcal{M}_p} n_p \int \left\{ \int \frac{\partial \log m(y_j \mid y_{[\overline{\mathcal{M}_k(j_+)}]}, \theta_{j.\overline{\mathcal{M}_k(j_+)}})}{\partial \theta_{j.\overline{\mathcal{M}_k(j_+)}}} m(y_{[\mathcal{M}_p]} \mid y_{[\mathcal{O}_p]}, \theta^*)\ \mathrm{d}y_{[\mathcal{M}_p]} \right\} \\
&\qquad \times f(y_{[\mathcal{O}_p]} \mid P=p)\ \mathrm{d}y_{[\mathcal{O}_p]} \\
&\quad + \sum_{j\in\mathcal{M}_p} n_p \int \left\{ \int \frac{\partial \log m(y_j \mid y_{[\overline{\mathcal{M}_k(j_+)}]}, \theta_{j.\overline{\mathcal{M}_k(j_+)}})}{\partial \theta_{j.\overline{\mathcal{M}_k(j_+)}}} m(y_{[\mathcal{M}_p]} \mid y_{[\mathcal{O}_p]}, \theta^*)\ \mathrm{d}y_{[\mathcal{M}_p]} \right\} \\
&\qquad \times f(y_{[\mathcal{O}_p]} \mid P=p)\ \mathrm{d}y_{[\mathcal{O}_p]} \\
&= n_0 \int \frac{\partial \log m(y_j \mid y_{[\overline{\mathcal{M}_k(j_+)}]}, \theta_{j.\overline{\mathcal{M}_k(j_+)}})}{\partial \theta_{j.\overline{\mathcal{M}_k(j_+)}}} f(y \mid P=0)\ \mathrm{d}y \\
&\quad + \sum_{j\notin\mathcal{M}_p} n_p \int \left\{ \int \frac{\partial \log m(y_j \mid y_{[\overline{\mathcal{M}_k(j_+)}]}, \theta_{j.\overline{\mathcal{M}_k(j_+)}})}{\partial \theta_{j.\overline{\mathcal{M}_k(j_+)}}} m(y_{[\mathcal{M}_p]} \mid y_{[\mathcal{O}_p]}, \theta^*)\ \mathrm{d}y_{[\mathcal{M}_p]} \right\} \\
&\qquad \times f(y_{[\mathcal{O}_p]} \mid P=p)\ \mathrm{d}y_{[\mathcal{O}_p]} \\
&\quad + \sum_{j\in\mathcal{M}_p} n_p \int \left\{ \int \frac{\partial \log m(y_j \mid y_{[\overline{\mathcal{M}_k(j_+)}]}, \theta_{j.\overline{\mathcal{M}_k(j_+)}})}{\partial \theta_{j.\overline{\mathcal{M}_k(j_+)}}} m(y_j \mid y_{[\overline{\mathcal{M}_k(j_+)}]}, \theta^*_{j.\overline{\mathcal{M}_k(j_+)}})\ \mathrm{d}y_j \right\} \\
&\qquad \times m(y_{[\mathcal{M}_k(j_-)]} \mid y_{[-\mathcal{M}_k(j_-)]}) \times m(y_{[\mathcal{M}_k(j_+)]} \mid y_{[-\mathcal{M}_k(j_+)]}) \times f(y_{[\mathcal{O}_p]} \mid P=p)\ \mathrm{d}y_{[-j]} \\
&= \tilde{s}(\theta_{j.\overline{\mathcal{M}_k(j_+)}} \mid \theta^*_{[j-]}, \theta^*_{[-j]}) \\
&\quad + \sum_{j\in\mathcal{M}_p} n_p \int \left\{ \int \frac{\partial \log m(y_j \mid y_{[\overline{\mathcal{M}_k(j_+)}]}, \theta_{j.\overline{\mathcal{M}_k(j_+)}})}{\partial \theta_{j.\overline{\mathcal{M}_k(j_+)}}} m(y_j \mid y_{[\overline{\mathcal{M}_k(j_+)}]}, \theta^*_{j.\overline{\mathcal{M}_k(j_+)}})\ \mathrm{d}y_j \right\} \\
&\qquad \times m(y_{[\mathcal{M}_k(j_-)]} \mid y_{[-\mathcal{M}_k(j_-)]}) \times m(y_{[\mathcal{M}_k(j_+)]} \mid y_{[-\mathcal{M}_k(j_+)]}) \times f(y_{[\mathcal{O}_p]} \mid P=p)\ \mathrm{d}y_{[-j]}.
\end{aligned}
$$

Then the asymptotic score equations holds at θ^*:

$$
\begin{aligned}
& \tilde{s}(\theta^*_{j.\overline{\mathcal{M}_k(j_+)}} \mid \theta^*_{[j-]}, \theta^*_{[-j]}) \\
= \;& \partial\tilde{Q}(\theta_{j.\overline{\mathcal{M}_k(j_+)}},\ \theta^*_{\overline{\mathcal{M}_k(j_+)}},\ \theta^*_{\mathcal{M}_k(j_+).[-\mathcal{M}_k(j_+)]} \mid \theta^*) \ / \ \partial\theta_{j.\overline{\mathcal{M}_k(j_+)}}\Big|_{\theta^*_{j.\overline{\mathcal{M}_k(j_+)}}} \\
& - \sum_{j\in\mathcal{M}_p} n_p \int \left\{ \int \frac{\partial \log m(y_j \mid y_{[\overline{\mathcal{M}_k(j_+)}]}, \theta_{j.\overline{\mathcal{M}_k(j_+)}})}{\partial\theta_{j.\overline{\mathcal{M}_k(j_+)}}} m(y_j \mid y_{[\overline{\mathcal{M}_k(j_+)}]}, \theta^*_{j.\overline{\mathcal{M}_k(j_+)}})\ \mathrm{d}y_j \right\} \Big|_{\theta^*_{j.\overline{\mathcal{M}_k(j_+)}}} \\
& \quad \times m(y_{[\mathcal{M}_k(j_-)]} \mid y_{[-\mathcal{M}_k(j_-)]}) \times m(y_{[\mathcal{M}_k(j_+)]} \mid y_{[-\mathcal{M}_k(j_+)]}) \times f(y_{[\mathcal{O}_p]} \mid P = p)\ \mathrm{d}y_{[-j]} \\
= \;& 0.
\end{aligned}
$$

Supplementary Materials

A Four-Variable Example

We hereby demonstrate how BSRMI is executed for a four-variable case with the most general missing pattern for $t = 2, \ldots, T$. There are 15 combinations of missing indicators of four variables, excluding the combination where all four variables are missing: $1, \ldots, 15$:

Pattern	Y_1	Y_2	Y_3	Y_4
1	✓	✓	✓	✓
2	✓	✓	✓	?
3	✓	✓	?	✓
4	✓	✓	?	?
5	✓	?	✓	✓
6	✓	?	✓	?
7	✓	?	?	✓
8	✓	?	?	?
9	?	✓	✓	✓
10	?	✓	✓	?
11	?	✓	?	✓
12	?	?	✓	✓
13	?	✓	?	?
14	?	?	✓	?
15	?	?	?	✓

For general purposes, we assume the imputation algorithm uses the sequence ordering of Y_1, Y_2, Y_3 and Y_4.

To impute Y_1: The filled-in block B_{10} consisting of patterns $1-8$ is used to fit all the following models:

Block	Pattern	Model
$B_{1.234}$	9	$m_{1.234}(Y_1 \mid Y_2, Y_3, Y_4)$
$B_{1.23}$	10	$m_{1.23}(Y_1 \mid Y_2, Y_3)$
$B_{1.24}$	11	$m_{1.24}(Y_1 \mid Y_2, Y_4)$
$B_{1.34}$	12	$m_{1.34}(Y_1 \mid Y_3, Y_4)$
$B_{1.2}$	13	$m_{1.2}(Y_1 \mid Y_2)$
$B_{1.3}$	14	$m_{1.3}(Y_1 \mid Y_3)$
$B_{1.4}$	15	$m_{1.4}(Y_1 \mid Y_4)$

In each block, the corresponding model is fitted using the filled-in block, and $Y_1^{(t)}$ is drawn from the approximate posterior distribution.

To impute Y_2: The filled-in block B_{20} is used to fit all the following models:

Block	Pattern	Model
$B_{2.134}$	5, 13	$m_{2.134}(Y_2 \mid Y_1, Y_3, Y_4)$
$B_{2.13}$	6, 14	$m_{2.13}(Y_2 \mid Y_1, Y_3)$
$B_{2.14}$	7, 15	$m_{2.14}(Y_2 \mid Y_1, Y_4)$
$B_{2.1}$	8	$m_{2.1}(Y_2 \mid Y_1)$

In each block, the corresponding model is fitted using the filled-in block, and $Y_2^{(t)}$ is drawn from the approximate posterior distribution.

To impute Y_3: The filled-in block B_{30} is used to fit the following models:

Block	Pattern	Model
$B_{3.124}$	3, 7, 11, 15	$m_{3.124}(Y_3 \mid Y_1, Y_2, Y_4)$
$B_{3.1}$	4, 8, 12	$m_{3.1}(Y_3 \mid Y_1)$

In each block, the corresponding model is fitted using the filled-in block, and $Y_3^{(t)}$ is drawn from the approximate posterior distribution.

To impute Y_4: The filled-in block B_{40} is used to fit the following model:

Model $m_{4.123}(Y_4 \mid Y_1, Y_2, Y_3)$ is specified for block $B_{4.123}$.

The model is fitted using the filled-in block, and $Y_4^{(t)}$ is drawn from the approximate posterior distribution.

Additional Tables

Table 3.3: Bias in the parameter estimates based on BSRMI using the poor fitting models under different orders, Replicates=500, Imputations=20, Sample size=10,000

	Order					
	(1,2,3)	(1,3,2)	(2,1,3)	(2,3,1)	(3,2,1)	(3,1,2)
$\alpha_{10} = \log(2)$	0.057	0.031	0.073	0.052	0.043	0.071
$\alpha_{20} = -1$	0.430	0.442	0.404	0.446	0.438	0.393
$\alpha_{21} = 0.3$	-0.197	-0.195	-0.189	-0.192	-0.195	-0.195
$\alpha_{30} = -2$	0.835	0.855	0.888	0.842	0.858	0.858
$\alpha_{31} = 0.5$	-0.270	-0.272	-0.355	-0.404	-0.407	-0.432
$\alpha_{32} = 0.1$	-0.487	-0.494	-0.180	0.043	0.039	0.093

Table 3.4: Bias in the parameter estimates based on BSRMI using the actual conditional models under different orders, Replicates=500, Imputations=20, Sample size=10,000

	Order					
	(1,2,3)	(1,3,2)	(2,1,3)	(2,3,1)	(3,2,1)	(3,1,2)
$\alpha_{10} = \log(2)$	-0.014	-0.013	-0.014	-0.015	-0.014	-0.015
$\alpha_{20} = -1$	0.031	0.042	0.040	0.046	0.038	0.0391
$\alpha_{21} = 0.3$	0.019	0.015	0.013	0.016	0.014	0.013
$\alpha_{30} = -2$	0.045	0.051	0.048	0.044	0.048	0.052
$\alpha_{31} = 0.5$	0.037	0.033	0.035	0.034	0.041	0.032
$\alpha_{32} = 0.1$	0.013	0.014	0.012	0.014	0.013	0.012

CHAPTER IV

Sequential Quasi-Likelihood Regression Multivariate Imputation (SQLRMI)

Abstract

The sequential regression multivariate imputation (SRMI) approach (also known as chained equations, fully conditional specifications, or flexible conditional specifications) imputes missing values using a sequence of conditional univariate regression models. The predictive power of the regression model sequence has been shown to determine its performance. Currently almost all software packages adopting this approach use parametric regression models within the generalized linear regression framework. However, it can be difficult to identify a well-fitting parametric model sequence for some common types of variables. This paper extends the sequential and the block sequential approach by using a quasi-likelihood approach to better capture the structure of the prediction model for the missing values. We examine the performance of the modified approach through simulation studies. We show that quasi-likelihood regression models make it easier to choose better-fitting model sequences to yield desirable repeated sampling properties of the multiple imputation estimates.

Key Words: Missing data; Multiple imputation; Chained equations; Conditional specifications; Quasi-likelihood regression

4.1 Introduction

4.1.1 SRMI Background

The sequential regression imputation approach (Kennickel, 1991, Raghunathan et al., 2001), also called imputation with chained equations or fully conditional specifications (Van Buuren and Oudshoorn, 1999), uses a Gibbs sampling style iterative algorithm. For each missing variable, the algorithm assumes a univariate regression model using all other variables as predictors. Let U be variables with no missing values. Let $Y_1, Y_2, \ldots, Y_p$ be p variables with missing values and let $Pr(Y_j|U, Y_1, Y_2, \ldots, Y_{j-1}, Y_{j+1}, \ldots, Y_p, \theta_j)$ be the conditional distribution of Y_j given $Y_{-j} = \{U, Y_1, \ldots, Y_{j-1}, Y_{j+1}, \ldots, Y_p\}$, and some unknown parameters θ_j. Let $\pi(\theta_j)$ denote the prior density for θ_j. At iteration t, let $Y_{-j}^{(t)} = \{U, Y_1^{(t)}, Y_2^{(t)}, \ldots, Y_{j-1}^{(t)}, Y_{j+1}^{(t-1}, \ldots, Y_p^{(t-1)}\}$ where $Y_k^{(s)}$ is the filled-in data for variable Y_k at iteration s. At iteration t, the imputation of the variable Y_j involves two steps: (1) the observed values Y_j and the corresponding $Y_{[-j]}^{(t)}$ are used to construct the approximate posterior density of θ_j and a value is drawn; and (2) the imputations are drawn from the regression model using the predictors and the drawn value of the parameters.

This approach has two major practical advantages over joint model-based imputation methods. It enables handling of complex data structures by focusing on a set of regression models with a univariate outcome. Compared to a joint model, the flexible selection of regression models enables improved prediction of the missing values based on other variables, and the regression models are more intuitive to analysts. Furthermore, individual regression models can easily account for study designs such as skip patterns, logical constraints, bounds for imputed values and consistency

requirements.

A theoretical weakness of the sequential regression approach, however, is that the specifications of the conditional distributions for a set of variables do not guarantee the existence of a joint distribution, and hence, the algorithm may not converge. Liu et al. (2014), Hughes et al. (2014) and Zhu and Raghunathan (2015) examined the convergence properties through both analytical and simulation investigations. One of the key findings was that if the regression models fit the data well, one obtains valid inferences. This assumption can be challenging for certain types of variables, such as grouped binary or count variables. The generalized linear model usually imposes certain restrictions on the mean and variance of the distributions. A review of current statistical software packages indicates that although features in SRMI applications account for data structures (logic constraints, bounds, censoring, etc.), the packages only use generalized linear regression models. However, the following motivating example will show that it can be difficult to identify a well-fitting generalized linear regression model sequence even for some common variables.

4.1.2 Motivating Example: SRMI for Trivariate Poisson

Consider the trivariate Poisson case discussed in Zhu and Raghunathan(2016), where the complete data are generated from the joint population distribution defined as follows:

$$\begin{aligned} y_1 &\sim \text{Poisson}(\lambda = \exp(\alpha_{10})), \\ y_2 \mid y_1 &\sim \text{Poisson}(\lambda = \exp(\alpha_{20} + \alpha_{21} y_1)), \\ y_3 \mid y_1, y_2 &\sim \text{Poisson}(\lambda = \exp(\alpha_{30} + \alpha_{31} y_1 + \alpha_{32} y_2)), \end{aligned}$$

where $\alpha_{10} = \log(2),\ \alpha_{20} = -1,\ \alpha_{21} = 0.3,\ \alpha_{30} = -2,\ \alpha_{31} = 0.5$ and $\alpha_{32} = 0.1$.

Suppose variables are missing at random with a general pattern of missingness.

The seven possible missing data patterns are listed below, where ✓ denotes observed and ? denotes missing:

P	Y_1	Y_2	Y_3
1	✓	✓	✓
2	✓	✓	?
3	✓	?	✓
4	✓	?	?
5	?	✓	✓
6	?	✓	?
7	?	?	✓

Suppose that an SRMI approach assumes three conditional generalized linear regression models:

$$\begin{aligned} m_1(y_1 \mid y_2, y_3) &\sim \text{Poisson}(\lambda = \exp(\theta_{10} + \theta_{12} y_2 + \theta_{13} y_3)), \\ m_2(y_2 \mid y_1, y_3) &\sim \text{Poisson}(\lambda = \exp(\theta_{20} + \theta_{21} y_1 + \theta_{23} y_3)), \\ m_3(y_3 \mid y_1, y_2) &\sim \text{Poisson}(\lambda = \exp(\theta_{30} + \theta_{31} y_1 + \theta_{32} y_2)). \end{aligned}$$

The algorithm breaks down due to extremely large values imputed within the first few iterations. A quick diagnostic also shows that these models are poor fits for the regression of Y_2 on (Y_1, Y_3) and Y_1 on (Y_2, Y_3). This behavior indicates that, when data are missing simultaneously within a record at the same time that sequential regression imputation models are not good fits, the algorithm may not generate convergent results, let alone valid analysis. In response, Chapter 3 proposes a modified algorithm, Block Sequential Regression Multivariate Imputation (BSRMI), to ensure convergence. However, the modified algorithm still relies on parametric generalized linear models, which may not be well fitting.

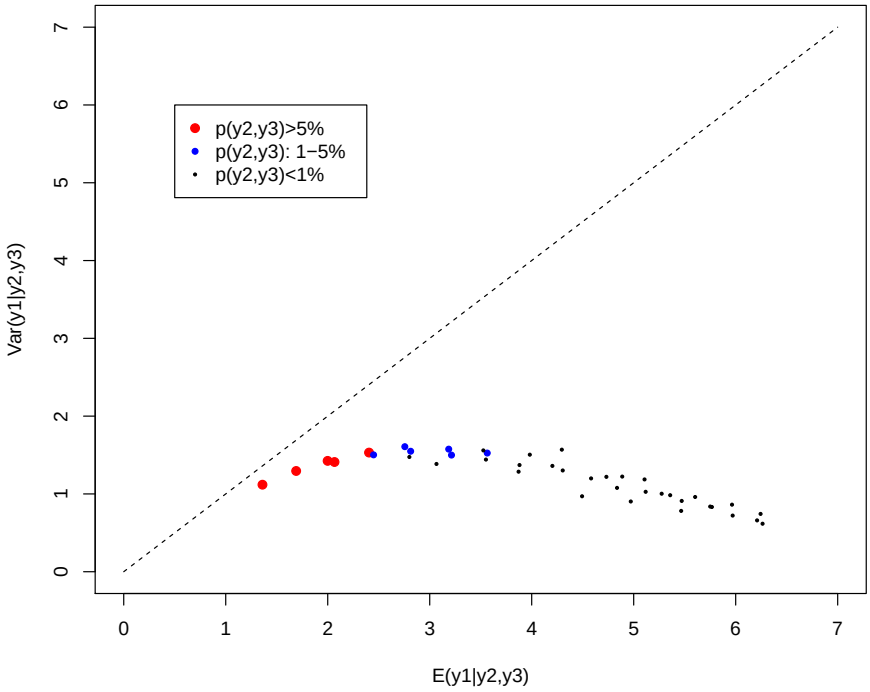

Figure 4.1: The relation between mean and variance of $y_1 \mid y_2, y_3$

We show through model diagnostics that the Poisson regression models can be poorly fitting. Figure 4.1 provides a scatter plot of the variance of the conditional distribution of Y_1 given Y_2 and Y_3 against the mean, distinguished by the joint probability distribution of (Y_2, Y_3). The variance is smaller than the mean and suggests a quadratic relationship between the mean and variance. A similar plot of Y_2 given (Y_1, Y_3) shows that Y_2 is zero-inflated (compared to a Poisson distribution), and the Poisson model is a poor fit. One alternative is to use linear regression models based on transformation methods (such as the Box-Cox transformation), but for this example they do not yield desirable results either.

In this paper, we use quasi-likelihood regression models to accommodate such complexities where the mean and variance function can be modeled to fit the data better and introduce the notion of imputations as draws from a quasi-predictive

distribution. Section 4.2 introduces the proposed algorithm. Section 4.3 revisits the motivating example, and examines the performance of the proposed algorithm through an extensive simulation study. Section 4.4 summarizes the findings, and discusses the advantages and limitations of the proposed sequential regression algorithm.

4.2 Methods

At any given iteration, let Y be the variable to be imputed and X be the remaining variables to be used as predictors. The objective is to fit a regression model,

$$E(Y|X) = \mu = g(X^t\beta, a),$$

where g is a known function, β is a p-dimensional vector of unknown parameters, and a represents known constants-e.g., the "off-set" term in the Poisson regression or the denominator in the grouped binomial outcome. The variance function is modeled as

$$V(Y|X) = V(\phi, \mu, b),$$

where V is a known function, b represents known constants, and ϕ is a vector of unknown parameters.

A particular form of the variance function is $V(Y|X) = \phi^2 V(\mu, b)$, which is useful for handling under- or over-dispersion in the variance relative to the mean. Specification of the mean and variance functions may be enough for many common distributions, but it is not sufficient. We introduce the notion of a pseudo-density,

$$p(y|x, \phi, \beta) = K(\phi, \beta) \exp\left(\int_y^\mu \frac{y-t}{\phi^2 V(t,b)} d\lambda(t)\right),$$

which can be viewed as a saddlepoint approximation of the density based on the first two moments with respect to the appropriate measure, λ (McCullagh and Nelder, 1983, and Butler, 2007).

Treating Y and X as an $n \times 1$ vector and an $n \times p$ matrix, respectively (for n individuals), the parameter β for a given ϕ is estimated by solving the equation,

$$D^t W^{-1}(Y - \mu) = 0,$$

where $D = \partial\mu/\partial\beta$ is an $n \times p$ matrix, and W is a diagonal matrix with variances as the entries. The estimation of ϕ requires an additional objective function,

$$h(\phi, \mu) + (Y - \mu)^t W^{-1}(Y - \mu),$$

to be minimized with respect to ϕ, where h is a known function. One iterates between estimating β and ϕ with these two functions to jointly estimate the unknown parameters (β, ϕ).

If $V(\phi, \mu, b) = \phi^2 V(\mu, b)$, where V is a known function, β can be estimated without knowing ϕ. For a given β, ϕ^2 can be estimated as $\widehat{\phi}^2 = (Y - \widehat{\mu})^t W^{-1}(Y - \widehat{\mu})/(n - p)$, where W is the diagonal matrix with $V(\mu, b)$ as entries.

4.2.1 A Quasi-Predictive Distribution

For a regular SRMI using a parametric model sequence, imputed values are naturally drawn from the posterior predictive distribution. Since quasi-likelihood regression models do not specify the distributions, we need an alternative way to draw values for the missing variables.

It is possible to use the estimated mean function to predict the missing values for imputation. However, this strategy ignores the sampling variability in the estimated parameters and the inherent variability in the actual observations even when the actual or true values of the parameters are known. Therefore, we suggest a modification as follows:

1. Approximate $(n - p)\widehat{\phi}^2/\phi^2$ by a chi-square distribution with $n - p$ degrees of

freedom. This approximation may be viewed as the saddlepoint approximation (Barndorff-Nielsen and Cox, 1979) of the posterior distribution under the pseudo-probability distribution with mean μ and variance $\phi^2 V(\mu, b)$ and a non-informative prior $\pi(\phi) \propto \phi^{-1}$. Define $\phi_*^2 = (n-p)\widehat{\phi}^2/u$, where $u \sim \chi^2_{n-p}$.

2. Define $T = (D^t W^{-1} D)^{-1}$, and C such that $CC^t = T$. Let z be a $p \times 1$ vector of standard normal deviates. Define $\beta_* = \widehat{\beta} + \phi_* C z$. This also can be viewed as a draw from the saddlepoint approximation of the posterior density of β conditional on data and ϕ_* under the pseudo-probability distribution and a non-informative prior $\pi(\beta|\phi) \propto 1$.

3. The missing value for subject i is drawn from a quasi-predictive distribution with the density,

$$f(y_i \mid \phi_*, \beta_*) = K(\phi_*, \beta_*) \exp\left[\int_{y_i}^{\mu_{*i}} \frac{y_i - t}{\phi_*^2 V_*(t, b)} d\lambda(t)\right],$$

where $K(\phi_*, \beta_*)$ is a normalizing constant with respect to the appropriate measure λ. The rejection sampling technique or the inversion of the distribution function can be used to create imputations. Again, this step can be viewed as creating draws from a saddlepoint approximation of the density function with the mean and variance functions.

4. In the sequential or block-sequential regression framework, some or all univariate regression models can be quasi-likelihood regression models.

In the trivariate Poisson example, the following mean

$$E(Y_1 \mid Y_2, Y_3) = \mu_1(\beta_1, Y_2, Y_3) = \beta_{10} + \beta_{11} Y_2 + \beta_{12} Y_3 + \beta_{13} Y_2 Y_3,$$

and variance

$$Var(Y_1 \mid Y_2, Y_3) = \phi^2 \mu_1(\beta, Y_2, Y_3)[1 - 0.16\mu_1(\beta_1, Y_2, Y_3)]$$

fit the data reasonably well, where $\beta = (\beta_{10}, \beta_{11}, \beta_{12}, \beta_{13})$. Without specifying the distribution, the quasi-maximum likelihood estimation of β based on the above mean-variance specification is equivalent to solving the following estimating equations:

$$\sum_s \left(\frac{\partial \mu_{1s}}{\partial \beta_1}\right)^T Var(Y_{1s} \mid Y_{2s}, Y_{3s})^{-1}(Y_{1s} - \mu_{1s}) = 0.$$

For a given β, ϕ^2 can be estimated using the method described above.

Given the estimated parameters β and ϕ, the quasi-predictive distribution of Y_1 given Y_2 and Y_3 is

$$Pr(Y_1 = y \mid Y_2, Y_3) \propto \exp\left(\int_y^{\mu} \frac{y - t}{\phi^2 t(1 - 0.16t)} \mathrm{d}t\right).$$

4.3 Simulation Studies

To evaluate repeated sampling properties of the multiple imputation estimates derived using the quasi-likelihood regression models, we use the trivariate Poisson data as the motivating example. The following steps describe the simulation study:

1. **Data Generation:** 500 data sets of sample size $n = 1000$ are generated.

2. **Missing Data Mechanism:** Missing values are created based on the following missing at random mechanism; data are divided into six random groups:

   (a) In the first group of size 100, y_2 and y_3 are observed and

   $$\text{pr}(y_1 \text{ is set to be missing} \mid y_2, y_3) = [1 + \exp(1.6 - 0.8y_2 - 0.7y_3)]^{-1};$$

   (b) In the second group of size 100, y_1 and y_3 are observed and

   $$\text{pr}(y_2 \text{ is set to be missing} \mid y_1, y_3) = [1 + \exp(2.1 - 0.8y_1 - 0.8y_3)]^{-1};$$

   (c) In the third group of size 100, y_1 and y_2 are observed and

   $$\text{pr}(y_3 \text{ is set to be missing} \mid y_1, y_2) = [1 + \exp(2.3 - 0.9y_1 - 0.8y_2)]^{-1};$$

(d) In the fourth group of size 300, y_3 is observed and

$$\text{pr}(y_1 \text{ and } y_2 \text{ are set to be missing} \mid y_3) = [1 + \exp(1 - 1.2y_3)]^{-1};$$

(e) In the fifth group of size 200, y_1 is observed and

$$\text{pr}(y_2 \text{ and } y_3 \text{ are set to be missing} \mid y_1) = [1 + \exp(1.7 - 1.5y_1)]^{-1};$$

(f) In the sixth group of size 200, y_2 is observed and

$$\text{pr}(y_1 \text{ and } y_3 \text{ are set to be missing} \mid y_2) = [1 + \exp(1.8 - 1.4y_2)]^{-1}.$$

3. **Parameters of Interest:** We choose the true values of population parameters α_{10}, α_{20}, α_{21}, α_{30}, α_{31}, and α_{32}. We consider several multiple imputation algorithms, and compare the MI inference to those based on before-deletion data and complete cases, respectively.

4. **Methods Considered:**

(a) BSRMI with GLM: We consider all six permutations of orders, (1,2,3); however, we only show the details and results for the first imputation order since the order effect is minimal if the models for all imputation orders are reasonably well fitting. The improved algorithm is developed based on the first simulated data set with missing values, just like in any practical application. The following specifications provide well-fitting models:

i. For y_1: Based on $y_{1,obs}$, $y_2^{(t-1)}$ and $y_3^{(t-1)}$, $\theta_{10}^{(t)}$ is estimated from $m_{10}(y_1 \mid y_2, y_3, \theta_{10}) \sim \text{Poisson}(\lambda = \exp(\theta_{100} + \theta_{101}y_2^{1/3} + \theta_{102}1/(1+y_2)^3 + \theta_{103}y_3^{(1/3)} + \theta_{104}1/(1+y_3)^3 + \theta_{105}\log(y3+1) * 1_{(y3<5)} + \theta_{106}y_2^{1/3} * y_3^{1/3}))$; $\theta_{12}^{(t)}$ are estimated from $m_{12}(y_1 \mid y_2, \theta_{12}) \sim \text{Poisson}(\lambda = \exp(\theta_{120} + \theta_{121}y_2^{1/3} + \theta_{122}1/(1+y_2)^3))$ and $\theta_{13}^{(t)}$ is estimated from $m_{13}(y_1 \mid y_3, \theta_{13}) \sim \text{Poisson}(\lambda = \exp(\theta_{130} + \theta_{131}y_3^{(1/3)} + \theta_{132}1/(1+y_3)^3 + \theta_{133}\log(y3+1) * 1_{(y3<5)}))$;

- for records with y_1 missing and y_2 and y_3 observed, values of $y_{1,mis}^{(t)}$ are drawn from the Poisson distribution $m_{10}(y_1 \mid y_{2,obs}, y_{3,obs}, \theta_{10}^{(t)})$;
- for records with y_1 and y_3 missing and y_2 observed, values of $y_{1,mis}^{(t)}$ are drawn from the Poisson distribution $m_{12}(y_1 \mid y_{2,obs}, \theta_{12}^{(t)})$;
- for records with y_1 and y_2 missing and y_3 observed, values of $y_{1,mis}^{(t)}$ are drawn from the Poisson distribution $m_{13}(y_1 \mid y_{3,obs}, \theta_{13}^{(t)})$;

ii. For y_2: Based on $y_{2,obs}$, $y_1^{(t)}$ and $y_3^{(t-1)}$, $\theta_{20}^{(t)}$ is estimated from $m_{20}(y_2 \mid y_1, y_3, \theta_{20}) \sim \text{Poisson}(\lambda = \exp(\theta_{200} + \theta_{201}y_1 + \theta_{203}y_3))$; $\theta_{21}^{(t)}$ is estimated from $m_{21}(y_2 \mid y_1, \theta_{21}) \sim \text{Poisson}(\lambda = \exp(\theta_{210} + \theta_{211}y_1))$;

- for records with y_2 missing and y_3 observed, values of $y_{2,mis}^{(t)}$ are drawn from the Poisson distribution $m_{20}(y_2 \mid y_1^{(t)}, y_{3,obs}, \theta_{20}^{(t)})$;
- for records with y_2 and y_3 missing, values of $y_{2,mis}^{(t)}$ are drawn from the Poisson distribution $m_{21}(y_2 \mid y_1^{(t)}, \theta_{21}^{(t)})$;

iii. For y_3: $\theta_3^{(t)}$ is estimated from $m_3(y_3 \mid y_1, y_2, \theta_3) \sim \text{Poisson}(\lambda = \exp(\theta_{30} + \theta_{31}y_1 + \theta_{32}y_2))$ based on $y_{3,obs}$, $y_1^{(t)}$ and $y_2^{(t)}$. Values of $y_{3,mis}^{(t)}$ are drawn from the Poisson distribution $m_3(y_3 \mid y_1^{(t)}, y_2^{(t)}, \theta_3^{(t)})$.

(b) SRMI using quasi-likelihood regression (SRMI-QLR): Based on model diagnostics, SRMI assumes

i. For Y_1:

$$E(Y_1 \mid Y_2, Y_3) = \mu_1(\beta_1, Y_2, Y_3) = \beta_{10} + \beta_{11}Y_2 + \beta_{12}Y_3 + \beta_{13}Y_2Y_3$$

and

$$Var(Y_1 \mid Y_2, Y_3) = \phi_1\mu_1(\beta_1, Y_2, Y_3)[1 - 0.16\mu_1(\beta_1, Y_2, Y_3)].$$

ii. For Y_2: The model is a two part specification due to the zero inflation in Y_2. First of all, a logistic regression model is assumed to fit whether

Y_2 is 0:

$$\text{logit}[\Pr(Y_2 = 0 \mid Y_1, Y_3)] = \eta_{20} + \eta_{21}Y_1 + \eta_{22}Y_3 + \eta_{23}Y_1Y_3.$$

If Y_2 is predicted to be positive, a second quasi-likelihood regression is assumed:

$$E(Y_2 \mid Y_1, Y_3) = \mu_2(\beta_2, Y_1, Y_3) = \beta_{20} + \beta_{21}Y_1 + \beta_{22}Y_3 + \beta_{23}Y_1Y_3$$

and

$$Var(Y_2 \mid Y_1, Y_3) = \phi_2\mu_2(\beta_2, Y_1, Y_3)[1 - 0.01\mu_2(\beta_2, Y_1, Y_3)].$$

iii. For Y_3:

$$E(Y_3 \mid Y_1, Y_2) = \mu_3(\beta_3, Y_1, Y_2) = \beta_{30} + \beta_{31}Y_1 + \beta_{32}Y_2$$

and

$$Var(Y_3 \mid Y_1, Y_2) = \phi_3\mu_3(\beta_3, Y_1, Y_2).$$

(c) BSRMI using quasi-likelihood regression (BSRMI-QLR): Similarly to BSRMI by GLM, we impute the missing values in the order of $Y1$, Y_2 and Y_3. In addition to the models used in SRMI, three additional models are assumed for the following patterns:

- for records with Y_1 and Y_3 missing and Y_2 observed, the quasi-likelihood regression model assumes

$$E(Y_1 \mid Y_2) = \mu_{12}(\beta_1, Y_2) = \beta_{102} + \beta_{112}Y_2$$

and

$$Var(Y_1 \mid Y_2) = \phi_{12}\mu_{12}(\beta_1, Y_2)[1 - 0.23\mu_{12}(\beta_1, Y_2)].$$

- for records with Y_1 and Y_2 missing and Y_3 observed, the quasi-likelihood regression model assumes

$$E(Y_1 \mid Y_3) = \mu_{13}(\beta_1, Y_3) = \beta_{103} + \beta_{113} Y_3$$

and

$$Var(Y_1 \mid Y_3) = \phi_{13}\mu_{13}(\beta_1, Y_3)[1 - 0.31\mu_{13}(\beta_1, Y_3)].$$

- for records with Y_2 and Y_3 missing, the quasi-likelihood regression model assumes

$$E(Y_2 \mid Y_1) = \mu_{21}(\beta_2, Y_1) = \beta_{201} + \beta_{211} Y_1$$

and

$$Var(Y_1 \mid Y_3) = \phi_{21}\mu_{21}(\beta_2, Y_1).$$

5. **Other factors:** For each imputation algorithm, M=5 completed data sets are generated with $T = 100$ iterations for each data set, and analysis results are combined using Rubin's (1976) formula.

Table 4.1 provides the Monte-Carlo bias for each parameter, which is the difference between the mean of the averaged parameter estimates from 500 data sets and the true value. We use the estimation based on before deletion data as the reference. Table 4.1 also includes the root mean squared error (RMSE) and the coverage rates of the 95% confidence interval for each parameter. The table shows that both SRMI-QLR and BSRMI-QLR perform better than BSRMI-GLM, and BSRMI-QLR is better than SRMI-QLR.

Table 4.1: Biases, RMSE and 95% confidence interval coverage rates of parameter estimates using various approaches, R=500, M=5

	BD			CC			BSRMI GLM			SRMI QLR			BSRMI QLR		
	Bias	RMSE	CR(%)	Bias	RMSE	CR(%)	Bias	RMSE	CR(%)	Bias	RMSE	CR(%)	Bias	RMSE	CR(%)
$\alpha_{10} = \log(2)$	0.008	0.022	95.4	-0.354	0.356	24.2	-0.017	0.034	94.7	-0.013	0.042	95.1	-0.012	0.042	95.2
$\alpha_{20} = -1$	-0.018	0.072	95.5	-0.326	0.351	31.1	0.061	0.117	91.1	0.092	0.151	92.1	0.072	0.131	93.7
$\alpha_{21} = 0.3$	0.008	0.061	94.3	-0.052	0.079	84.3	-0.019	0.063	92.9	0.017	0.064	94.5	0.013	0.066	94.5
$\alpha_{30} = -2$	0.017	0.092	95.2	-0.390	0.443	45.2	0.331	0.359	64.3	0.142	0.180	79.1	0.105	0.175	91.4
$\alpha_{31} = 0.5$	0.011	0.032	94.7	-0.132	0.154	40.8	-0.115	0.125	83.8	0.082	0.096	93.4	0.065	0.088	93.9
$\alpha_{32} = 0.1$	0.003	0.030	95.3	-0.315	0.353	2.1	0.012	0.061	91.5	0.013	0.061	92.8	0.013	0.061	93.6

The Kullback-Leibler divergences between before-deletion and after-imputation data distributions for each algorithm are summarized in Figure 4.2. Each algorithm converges to the same results as the sample size, number of iterations and number of imputations increase. Figure 4.2 further shows that better results are obtained using the improved imputation models.

4.4 Discussion

The sequential regression multivariate imputation algorithm is a popular approach for imputing missing values in a data set with a complex structure, where developing a joint distribution is difficult, if not impossible. Several studies have shown that the properties of the multiple imputation estimates have desirable sampling properties provided the models were good fits for each conditional distribution. However, this approach can break down if the models are not good fits. The trivariate Poisson example is one such instance. Interestingly, it is difficult to find well-fitting models within the exponential family for this example. Thus, we extended the SRMI approach by using the quasi-likelihood regression (QLR) models.

The QLR imputation approach can be used in either regular SRMI or modified BSRMI for data with general missing patterns. Since both SRMI and BSRMI adopt univariate regression models, it is easier to specify and fit the quasi-likelihood models for each variable on a case-by-case basis; more importantly, it is simple to draw imputed values from a quasi-predictive distribution defined by the mean and variance structures. Study features such as bounds and logical constraints can be accounted for similarly, as in parametric SRMI algorithms. The preservation of these features indicates that the extension will only require minimal modification to the current software packages.

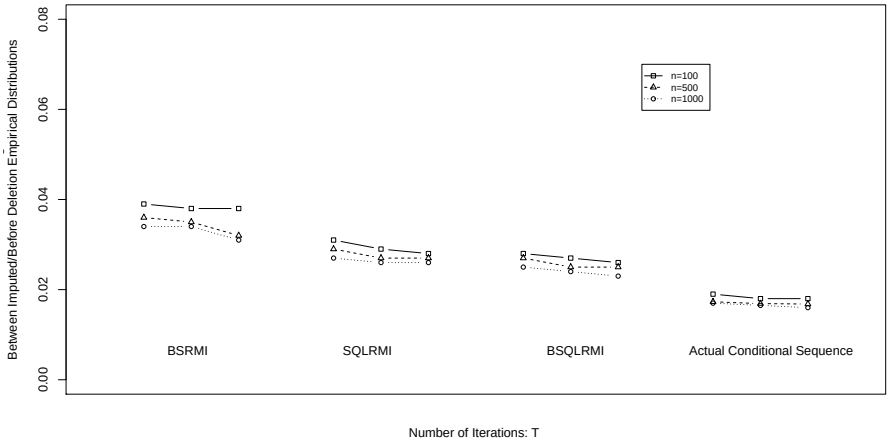

Figure 4.2: Kullback-Leibler divergence between empirical distributions, based on multiply imputed data $\widehat{F}^{n,T}_{MI}(x,y)$ and before deletion data $\widehat{F}^{n}_{BD}(x,y)$ and averaged over 500 replicates of data sets, from three imputation algorithms plotted as a function of sample size n=500, 1000, and 5000 and the number of iterations $T = 50, 100, 1000$ when the data are missing at random. The three imputation algorithms impute the missing values in the order of Y_1, Y_2 and Y_3: the poor sequence assumes linear predictors only, the improved sequence assumes appropriate non-linear terms of the predictors, and the validly specified sequence assumes true conditional models from the data population.

Our simulation study shows that quasi-likelihood models with validly specified mean structures and mean-variance relations yield reasonably good inferences. One limitation is the loss in predictive power using the saddlepoint approximation, when one may be able to draw from the actual predictive distribution whenever the mean and variance structures correspond to an actual predictive distribution. This difference may not be large for large samples. Since the QLR models do not specify the conditional distributions, the proposed approach does not assume compatibility of the models in the sequence, and existing theoretical frameworks do not apply for this method. Further research is needed to develop regularity conditions for convergence of SRMI using QLR for some or all variables.

In order to evaluate the imputation performance, we have assumed the missingness to be ignorable in this paper. Further research is needed to modify this approach

for non-ignorable missing data. One area of further research will be to use the univariate selection or mixture model together with the model for variables subject to non-ignorable missing data mechanisms.

CHAPTER V

Discussion

5.1 Conclusions and Discussions

The sequential regression multivariate imputation algorithm, a Gibbs sampling type iterative approach, is perhaps the most viable method for creating imputations in a complex data set with many types of variables, structural dependencies, skip patterns, and bounds for imputed values, etc. However, one of the theoretical limitations of this approach stems from the fact that specifications of conditional distributions do not guarantee the existence of a joint distribution, and, therefore, the Gibbs style iterative algorithm may not converge. Past simulation studies on SRMI mostly considered linear regression models only, and their conclusions may not extend to general cases. Limited theoretical work on the convergence properties of SRMI required strong assumptions such as the existence of a joint distribution for the algorithm, and tended not to separate compatibility and model fitting issues. In order to avoid these limitations, this thesis work considered many types of regression models for general missing data without assuming the existence of a joint distribution. The effects of model compatibility and model fitting on SRMI convergence and performance were also studied separately. The major findings of this thesis work are summarized as follows.

In Chapter 2, we focused on single variable missingness. We determined several types of model compatibility. Theoretical and simulational results indicate that SRMI algorithms converge under some regulatory conditions. However, eve strongly incompatible specifications tend to converge to a fixed point without achieving any overall compatibility. While it is analytically impossible to determine the fixed point in general, through special cases and simulation studies, we showed that the imputation performance can be evaluated depending on how well the model sequences fit the data. These findings suggest that for this special case, imputers need to focus on improving imputation model fitting instead of enforcing model compatibility.

For SRMI, algorithmic incompatibility may occur and result in algorithm divergence when poorly chosen models are used. This problem was motivated by a trivariate Poisson data set. To reduce the risk of such behavior, we introduced the block sequential multivariate imputation algorithm by modifying the SRMI algorithm through compatible block specification in Chapter 3. A set of regularity conditions for the convergence were developed and repeated sampling properties were evaluated through simulations. Again, after modification, the most important factor is the goodness of fit of the regression models used in the model sequence.

For outcomes such as the count variables, parametric regression models are not rich enough to handle over- or under-dispersion as well as the complex relationship between the mean and variance functions. In Chapter 4, we extended the parametric GLM framework to quasi-likelihood regression models to accommodate complex mean and variance structures without specifying the model distribution. We demonstrated that by choosing a reasonable mean-variance structure in the regression models, the imputation performance can be significantly improved.

Throughout, we assumed that the missing data mechanism is ignorable. Future

work includes extending the sequential regression framework to the nonignorable missing data mechanism. Selection models for variables in the SRMI can be developed to create imputations under the nonignorable missing data mechanism. For the block-specific SRMI, a pattern mixture model can easily adapt the nonignorable missing mechanism into its framework. For quasi-likelihood regression models, the variance can be assumed to be an unknown function of the mean instead of a fixed function.

BIBLIOGRAPHY

[1] P. W. Anderson. Absence of diffusion in certain random lattices. *Phys. Rev.*, 109:1492, 1954.

[2] R.H. Andridge and R. J. A. Little. A review of hot deck imputation for survey nonresponse. *International Statistical Review*, 78(1):40–64, 2010.

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

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

[5] 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.

[6] O. Barndorff-Nielsen and Cox D. R. Edgeworth and saddle-point approximations with statistical applications. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(3):279–312, 1986.

[7] J. Besag. Spatial interaction and the statistical analysis of lattice systems (with discussion). *J. R. Statist. Soc. B*, 36:192–236, 1974.

[8] J. P. L. Brand. Development, implementation and evaluation of multiple imputation strategies for the statistical analysis of incomplete data sets. Technical report, Erasmus University, Rotterdam, 1999.

[9] R. Butler. *Saddlepoint Approximations with Applications*. Cambridge University Press, Cambridge, 2007.

[10] H. Y. Chen, H. Xie, and Y. Qian. Multiple imputation for missing values through conditional semiparametric odds ratio models. *Biometrics*, 67(3):799–809, 2011.

[11] L. M. Collins, J. L. Schafer, and C. M. Kam. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, 6(4):330–51, 2001.

[12] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[13] J. Drechsler and R. Susanne. Does convergence really matter? In *Recent Advances in Linear Models and Related Areas*, pages 341–355. Physica-Verlag HD, 2008.

[14] 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.

[15] S. Geman and D. Geman. Stochastic relaxation, gibbs distributions and the bayesian restoration of images. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.

[16] R. A. Hughes, I. R. White, S. R. Seaman, J. R. Carpenter, K. Tilling, and J. A. C. Sterne. Joint modelling rationale for chained equations. *BMC Medical Research Methodology*, 14(1):1–10, 2014.

[17] 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.

[18] F. Li, M. Baccini, F. Mealli, E.R. Zell, C.E. Frangakis, and D.B. Rubin. Multiple imputation by ordered monotone blocks with application to the anthrax vaccine research program. *Journal of Computational and Graphical Statistics*, 23(3), 2014.

[19] R. J. A. Little and D. B. Rubin. *Statistical Analysis with Missing Data (2nd edition)*. John Wiley & Sons, New York, 2002.

[20] R.J.A. Little and H. An. Robust likelihood-based analysis of multivariate data with missing values. *Statistica Sinica*, 14:949–968, 2004.

[21] J. Liu, A. Gelman, J. Hill, Y.-S. Su, and J. Kropko. On the stationary distribution of iterative imputations. *Biometrika*, 101(1):155–173, 2014.

[22] P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman and Hall, London, 1983.

[23] Royston P. Multiple imputation of missing values -update. *The Stata Journal*, 5:188–201, 2005.

[24] 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.

[25] T.E. Raghunathan and J. Grizzle. A split questionnaire survey design. *Journal of the American Statistical Association*, 90:54–63, 1995.

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

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

[28] D. B. Rubin. Nested multiplei imputation of nmes via partially incompatible mcmc. *Statistica Neerlandica*, 57(1):3–18, 2003.

[29] D.B. Rubin. Formalizing subjective notions about the effect of non-respondents in sample surveys. *The Journal of the American Statistical Association*, 72(359):538–543, 1977.

[30] D.B. Rubin. Basic ideas of multiple imputation. *Survey Methodology*, 12:37–47, 1986.

[31] D.B. Rubin. *Multiple imputation for nonresponse in surveys*. New York: John Wiley & Sons, 1987.

[32] Van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, 16:219–242, 2007.

[33] SAS Institute Inc., Carey, NC. *SAS/STAT User's Guide 9.2*, 2008.

[34] J. L. Schafer and J. W. Graham. Missing data: Our view of the state of the art. *Psychological Methods*, 7(2):147–177, 2002.

[35] SPSS Inc., Chicago, IL. *SPSS Missing Values 17.0*, 2009.

[36] 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*, 45(1):1–31, 2011.

[37] 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.

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

[39] F. Wang, P. X.-K. Song, and L. Wang. Merging multiple longitudinal studies with study-specific missing covariates: A joint estimating function approach. *Biometrics*, 71(4):929–940, 2015.

[40] J. Zhu and T.E. Raghunathan. Convergence properties of a sequential regression multiple imputation algorithm. *Journal of the American Statistical Association*, 110(511):1112–1124, 2015.