Impact of inconsistent imputation models in mediation analysis

Ye, Bo* Yucel, Recai†

October 6, 2020

Abstract

We investigate the impact of possibly incoherent imputation models on mediation analyses. In particular, we discuss commonly used joint modelling approach as well as variable-by-variable approach when the ultimate analytical goal in mediation analysis. Practical advantages of each approach along with the discussion on coherence of the imputation model is our focal point in our manuscript. A comprehensive simulation study is summarized to gauge the performance of widely utilized imputation models.
Key Words: Mediation analysis, missing data, multiple imputation, compatible imputation

1 Introduction

Missing data is one of the most pervasive problem in the public health data. Mediation analysis is no exception to this. Methods by Fritz & MacKinnon (2008) and many others are commonly used techniques in exploring relationships through mediator variables. However, to the best of our knowlegde, limited number of studies have been conducted to investigate how missing values should be handled in mediation analysis. Zhang & Wang (2013) introduce and compare four approaches to deal with missing data in mediation analysis including listwise deletion, pairwise deletion, multiple imputation (MI) by Rubin (1976), and a two-stage maximum likelihood (TS-ML) method under various missingness mechanisms through simulation studies. In this study, they show that MI performed well when missingness mechanism is independent of any of the observed variables (i.e. missing completely at random (MCAR)) or when the missingness mechanism depends only on the observed variables but not on the missing variables (missing at random (MAR)) in the sense defined by Rubin (1976). However, the performance of MI under different imputation models that may or may not be compatible with the mediation analysis were not compared.

There is vast literature on missing data. Since the seminal work of Dempster et al. (1977), there has been a significant development in the missing data methodology. Campion & Rubin (1989) provides principle solutions in multiple imputations for survey data. There are two widely accepted types of solutions towards missing data problems. Researchers either use EM type solutions or multiple imputation type solutions, and multiple imputation is becoming more popular in application. Our focus here is based on multiple imputation to understand the nuances between imputation models leading to inference by MI. There are generally two widely accepted approaches to forming an imputation model: joint modeling (Schafer (1999); Little & Rubin (1988)) and sequential or variable-by-variable modeling (Raghunathan et al. (2001); van Buuren (2007)).

Many of the developments so far on MI have been made available to practitioners. Examples include . van Buuren & Groothuis-Oudshoorn (2011) who developed the R package **mice** which implements a wide variety of algorithms under variable-by-variable imputation

*Department of Epidemiology and Statistics, School of Public Health, State University of New York at Albany

†Department of Epidemiology and Biostatistics, College of Public Health, Temple University

and provides additional options to draw from the potential conditional predictive distributions, such as predictive-mean and propensity-score matching. Joseph L. Schafer and Alvaro A. Novo (2013) developed the R package **norm** for the analysis of multivariate normal datasets with missing values. Raghunathan, Lepkowski, VanHoewyk, and Solenberger (2001) developed a SAS macro, "IveWare", for the application of variable-by-variable imputation in SAS. In STATA, we have the procedure "ICE" by Royston and White (2011) to implement multiple imputation method for missing data problem.

There has been numerous studies investigating performance and robustness of alternative MI models. One example is Yucel et al. (2018) who presented a simulation study assessing the compatibility of sequential approach with the joint data generation mechanism based on a family of hierarchical regression models for correlated data. They found that the sequential method leads to well-calibrated estimates and often performs better than methods that that are currently available to practitioners. Hughes et al. (2014) showed that order effects (systematic differences depending upon the rank of the variables) in categorical variable was small when applying joint modeling method, especially when associations between variables are weak. Order effects are ubiquitous in medical research, but their results recommend that they may be small enough to be negligible. Mistler & Enders (2017) claims that JM draws missing values simultaneously for all incomplete variables using a multivariate distribution, whereas FCS imputes variables one at a time from a series of univariate conditional distributions. The study examined four multilevel multiple imputation approaches and concluded that their analytic work and computer simulations showed that FCS are more restrictive and impose implicit equality constraints on functions of the within- and between-cluster covariance matrices. Akande et al. (2017) used simulation studies to compare repeated sampling properties of three multiple imputation methods for categorical data: chained equations using generalized linear models, chained equations using classification and regression trees, and a fully Bayesian joint distribution based on Dirichlet process mixture models. In the circumstances of this study, the results suggest that default chained equations approaches based on generalized linear models are dominated by the default regression tree and Bayesian mixture model approaches.

In this paper, we discuss and compare joint Model and fully conditional specification imputation methods in mediation analysis with missing data problem. We start by introducing how to estimate mediation effects, followed by how imputation method works in this model. Then, the theoretical comparison of two methods is provided. Finally, we conduct simulation studies with finite samples to assess and compare the performance of the methods under different missing data mechanisms.

2 Methods

2.1 Notation

Suppose K random variables $D = (D_1, ..., D_K)^T$ are intended to be observed on N subjects with missing values. Subscripts i and j are used to index subjects and variables respectively $(i = 1, ..., N; j = 1, ..., K)$. Let d_{ij} denote an i^{th} row and j^{th} column element in an $(N * K)$ matrix. We denote the column j of matrix d by d_j=$(d_{1j}, ..., d_{Nj})^T$. We denote the complete data matrix $U = (u_{ij})$, the missing data indicator matrix $R = (R_{ij})$ and ϕ as the unknown parameters in the section of introducing missingness mechanism. In mediation analysis model, Y is denoted as the response variable, M is denoted as the mediator, and X is denoted as the independent variable. Three variables (X, Y, M) are included in the data matrix U. η is the matrix includes M and Y, δ represents mediation effect, β is the matrix of regression coefficients on mediator, I represents interactive term, and γ is the matrix of regression coefficients on x and I. For the purpose of missing data imputation, we will have to model all variables regardless of covariates and response as variables to be modeled.

2.2 Mediation Analysis Model

In this study, the analysis is based on a specified mediation analysis model. For example, Figure 1 is a typical display of mediation model. This is a single level model with one

mediator only, and there could be multiple independent variables. Additionally, we have

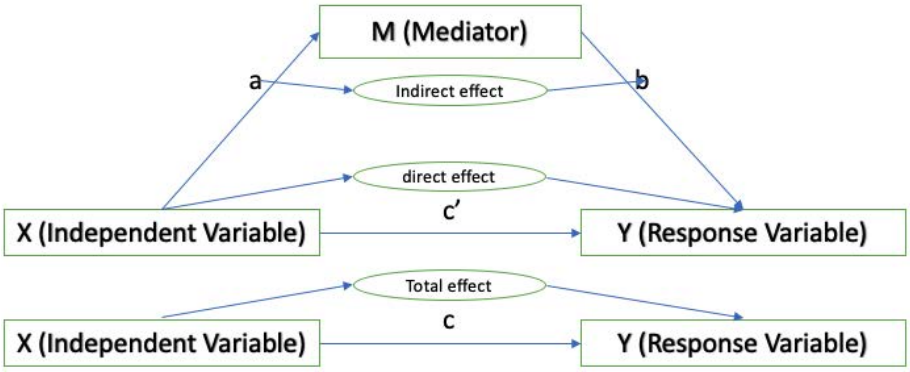

Figure 1: Diagram for mediation analysis

the corresponding relationship expressed in Bentler-weeks form as follows.

$$\eta = \beta * \eta' + \gamma * \epsilon, \tag{1}$$

$$\begin{bmatrix} M \\ Y \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ b & 0 \end{bmatrix} * \begin{bmatrix} M \\ Y \end{bmatrix} + \begin{bmatrix} a & 1 & 1 & 0 \\ c' & 1 & 0 & 1 \end{bmatrix} * \begin{bmatrix} X \\ I \\ e_M \\ e_Y \end{bmatrix}, \tag{2}$$

where M is a representative of the mediator and it could be either discrete or continuous variable. Y is a representative of the response variable and it also could be continuous or discrete variable. for example, the number of hospital visits (disease such as COPD, asthma and etc.), X refers to independent variables, for example, temperature, dewpoint and other covariates. I represents the interactive terms. $e_M \sim N(\mu_M, \sigma_M^2)$ and $e_Y \sim N(\mu_Y, \sigma_Y^2)$. e_M and e_Y represent the error terms in linear regressions for M and Y, respectively.

In this case, we use c' to measure the direct effect of X on Y. And we use a and b to measure the relations between X, M and Y. We use $\delta = ab$, the product of a and b to describe the mediation effect. The total effect is the summation of direct effect and mediation effect, and it could be expressed as $ab + c'$. We use logistic regression and linear regression respectively to establish the model for binary and continuous response variable, respectively.

There are p_1 and q_1 variables in M and Y, respectively. Then, η and η' are (p_1+q_1) *1 vectors, and β is a $(p_1+q_1)*(p_1+q_1)$ vector. Suppose we have p_2 and p_3 variables in X and I, respectively. ξ is a (p_2+p_3+2) *1 vector, and γ is a $(p_1+q_1)*(p_2+p_3+2)$ vector.

We further need to assume an underlying missingness mechanism as defined by Rubin (1976) and have been extensively discussed in missing data literature. We refer to Rubin (1976) and others for in-depth definitions. Here we consider three mechanisms: missing completely at random (MCAR) which means that missingness probabilities are independent of any observed data; missing at random (MAR) which means that missingness probabilities may depend on observed data but not on variables subject to missing data and, finally, missing not at random (MNAR) means that probabilities of missingness depend on the missing values and hence they need to be modeled. We mostly investigate performance under MCAR and MAR.

2.3 Joint and conditional imputation models

There are generally two widely commonly-used approaches in choosing parametric imputation model: joint (Schafer (1999); Little & Rubin (1988)) and conditional (or sequantail or variable-by-variable) (Raghunathan et al. (2001); van Buuren (2007)); Yucel et al. (2018)).

Joint modeling (JM) starts from the assumption that a set of variables is assumed to form a joint model such as multivariate normal distribution (Schafer (1997)). The imputations are basically random samples from the underlying posterior predictive distribution of missing data, which is essentially intractable. Computational implementations such as PROC MI and R package "missing" use data augmentation to bypass the issue of intractability.

JM can be quite impractical when it is applied to multivariate data with large number of categorical variables with missing values and/or with bounds/skip patterns or high level interactions. In such instances, fully conditional specification (FCS) emerged to be practical even though it samples from possibly incoherent imputation model with the underlying joint model Yucel et al. (2018)). In FCS, each variable is modeled at a time conditional on other variables (van Buuren (2007)). One of the major differences between JM and FCS is that FCS models each variable conditional on others and it leads to sampling from the implied posterior predictive distribution for that variable. As a result, these conditional posterior distributions can be inconsistent with the joint distributions. This aspect is the major focal point of our work.

3 Analytic Comparison of Joint versus Variable-by-variable Imputation model

Previous studies suggested two imputation strategies have different characteristics in terms of preserving associations between incomplete and complete variables Mistler & Enders (2017). To investigate inconsistencies between two imputation approaches from a mediation analysis perspective, we consider a simple scenario with three variables one of which is incompletely observed.

3.1 Population Joint Distribution

Consider three variables Y, X_1, and X_2 distributed as multivariate normal distribution:

$$\begin{bmatrix} Y \\ X_1 \\ X_2 \end{bmatrix} \sim N_3(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{3}$$

where the covariace matrices are as follows:

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_Y \\ \mu_{X_1} \\ \mu_{X_2} \end{bmatrix} \tag{4}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma^2_Y & \sigma_{YX_1} & \sigma_{YX_2} \\ \sigma_{X_1Y} & \sigma^2_{X_1} & \sigma_{X_1X_2} \\ \sigma_{X_2Y} & \sigma_{X_2X_1} & \sigma^2_{X_2} \end{bmatrix} \tag{5}$$

3.2 Fully conditional specification

In our particular scenario, parametric FCS would employs a linear models as a base to impute missing values in Y.

We consider the trivariate imputation problem from previous content. FCS imputes the incomplete variables in a sequence. Expressed in scalar notation, the conditional distribution that generates Y imputations is

$$Y^{(t)}_{mis} \sim N(\alpha^{(t-1)}_Y + \beta^{(t-1)}_{Y|X_1} X_{1i} + \beta^{(t-1)}_{Y|X_2} X_{2i},\ \sigma^2_{\epsilon(Y)}) \tag{6}$$

where $Y^{(t)}_{mis}$ represents the t^{th} drawn missing value from its posterior predictive distribution whose parameters are fixed at the previous iteration: $\alpha^{(t-1)}_Y$, $\beta^{(t-1)}_{Y|X_1}$ and $\beta^{(t-1)}_{Y|X_2}$ fixed at the previous iteration.

The marginal distribution of the predictor variables is

$$\begin{bmatrix} X_{1i} \\ X_{2i} \end{bmatrix} \sim N_2 \left(\begin{bmatrix} \bar{X}_1 \\ \bar{X}_2 \end{bmatrix}, \begin{bmatrix} \sigma^2_{X_1} & \sigma^2_{X_1X_2} \\ \sigma^2_{X_2X_1} & \sigma^2_{X_2} \end{bmatrix} \right) \tag{7}$$

and the conditional distribution of Y given X_1 and X_2 is

$$Y|X_1,\ X_2 \sim N\left(\mu_{Y|X_1,\ X_2},\ \sigma^2_{Y|X_1,\ X_2}\right) \tag{8}$$

where

$$\mu_{Y|X_1,\ X_2} = \boldsymbol{\upsilon}_Y + \gamma_{Y|X_1} * (X_{1i} - \bar{X}_1) + \gamma_{Y|X_2} * (X_{2i} - \bar{X}_2) \tag{9}$$

$$\sigma^2_{Y|X_1,\ X_2} = \sigma^2_Y - \gamma^2_{Y|X_1} * \sigma^2_{X_1} - \gamma^2_{Y|X_2} * \sigma^2_{X_2} - 2\gamma_{Y|X_1}\gamma_{Y|X_2}\sigma_{X_1X_2} \tag{10}$$

The procedure to develop FCS was similar with Mistler & Enders (2017). Comparing the number of parameters required by the FCS model (Equation 9 and 10) to JM model (Equation 4 and 5) indicates that JM is more restrictive than FCS. In FCS, it requires a total of 8 parameters (1 intercept, 2 regression coefficients, 2 means, 2 variance and 1 covariance) to govern the model.

From equation 9 to 10, we found that FCS contains different regression coefficients and intercept to preserve the relation between Y, X_1 and X_2. As mentioned previously, mediation analysis is constructed based on several regression models. We have logistic regression of X_1 on X_2, and linear regression of X_1 and X_2 on Y. Thus, in the situation of imputing missing variables in mediation, FCS is more appropriate for models that are embedded in regressions (e.g. mediation model).

3.3 Mediator estimate comparison

We have developed the distribution of dependent and independent variables under different multiple imputation methods in the process. In this section, we theoretically prove the consistency of different multiple imputation methods. We will develop the estimate of mediation effect based on conditional distribution of each variable in the process of imputation.

In general setting of joint modeling, suppose we have

$$\begin{bmatrix} Y \\ X_1 \\ X_2 \end{bmatrix} \sim N_3(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{11}$$

where the mean and covariace matrices are as follows:

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_Y \\ \mu_{X_1} \\ \mu_{X_2} \end{bmatrix} \tag{12}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma^2_Y & \sigma_{YX_1} & \sigma_{YX_2} \\ \sigma_{X_1Y} & \sigma^2_{X_1} & \sigma_{X_1X_2} \\ \sigma_{X_2Y} & \sigma_{X_2X_1} & \sigma^2_{X_2} \end{bmatrix} \tag{13}$$

We developed the conditional distribution of Y given X_1 and X_2, we have

$$\mu_{Y|X_1,X_2} = \mu_Y + \frac{1}{\sigma^2_{X_1}\sigma^2_{X_2} - \sigma^2_{X_1X_2}} \begin{bmatrix} \sigma_{YX_1} & \sigma_{YX_2} \end{bmatrix} \begin{bmatrix} \sigma^2_{X_2} & -\sigma_{X_1X_2} \\ -\sigma_{X_1X_2} & \sigma^2_{X_1} \end{bmatrix} \begin{bmatrix} X_1 - \mu_{X_1} \\ X_2 - \mu_{X_2} \end{bmatrix} \tag{14}$$

After mathematical matrix manipulation, we then have

$$\begin{aligned} \mu_{Y|X_1,X_2} = \mu_Y + \frac{1}{\sigma^2_{X_1}\sigma^2_{X_2} - \sigma^2_{X_1X_2}} \{ & (\sigma^2_{X_2}\sigma_{YX_1} - \sigma_{YX_2}\sigma_{X_1X_2})X_1 + (\sigma^2_{X_1}\sigma_{YX_2} - \\ & \sigma_{YX_1}\sigma_{X_1X_2})X_2 - \mu_{X_1}(\sigma^2_{X_2}\sigma_{YX_1} - \sigma_{yX_2\sigma_{X_1X_2}}) - \mu_{X_2}(\sigma^2_{X_1}\sigma_{YX_2} - \sigma_{yX_1\sigma_{X_1X_2}}) \} \end{aligned} \tag{15}$$

From the equation above, we could regard $\mu_{Y|X_1,X_2}$ as a function of X_1 and X_2. The coefficients in the formula represent the linear relationship between these variables. Our ultimate goal is to estimate the mediation effect based on this formula. The relation of X_1 and X_2 is shown below,

$$\mu_{X_1|X_2} = \mu_{X_1} + \frac{\sigma_{X_1X_2}}{\sigma^2_{X_2}}(X_2 - \mu_{X_2}) \tag{16}$$

$$\hat{\delta}_{conditional} = \hat{a}\hat{b} = \frac{\sigma^2_{X_2}\sigma_{YX_1} - \sigma_{YX_2}\sigma_{X_1X_2}}{\sigma^2_{X_1}\sigma^2_{X_2} - \sigma^2_{X_1X_2}} * \frac{\sigma_{X_1X_2}}{\sigma^2_{X_2}} \tag{17}$$

By mediation analysis, we have the following relationship between Y and X_1 in the formula 18, 19 and 20. As mediation analysis is based on regression, the estimate of mediation effect is the multiplication of regression coefficient. The coefficients in the joint modeling provide a sufficient evidence to compare the accuracy of estimation. We will do the same thing for FCS and estimate the mediation effect based on the parameters from FCS.

$$\hat{Y} = a_1 + b\hat{X_1} + c\hat{X_2} \tag{18}$$

where

$$\hat{b} = \frac{(\sum X^2_{2i})(\sum X_{1i}Y_i) - (\sum X_{1i}X_{2i})(\sum X_{2i}Y_i)}{(\sum X^2_{1i})(\sum X^2_{2i}) - (\sum X_{1i}X_{2i})^2} \tag{19}$$

$$\hat{c} = \frac{(\sum X^2_{1i})(\sum X_{2i}Y_i) - (\sum X_{2i}X_{1i})(\sum X_{1i}Y_i)}{(\sum X^2_{2i})(\sum X^2_{1i}) - (\sum X_{1i}X_{2i})^2} \tag{20}$$

We have $\hat{a}$, the regression coefficient of X_2 on X_1, is estimated as follows.

$$\hat{a} = \frac{\sum(X_{1i} - \bar{X_1})(X_{2i} - \bar{X_2})}{\sum(X_{1i} - \bar{X_1})^2} \tag{21}$$

$$\hat{\delta} = \hat{a}\hat{b} = \frac{\sum(X_{1i} - \bar{X_1})(X_{2i} - \bar{X_2})}{\sum(X_{1i} - \bar{X_1})^2} * \frac{(\sum X^2_{2i})(\sum X_{1i}Y_i) - (\sum X_{1i}X_{2i})(\sum X_{2i}Y_i)}{(\sum X^2_{1i})(\sum X^2_{2i}) - (\sum X_{1i}X_{2i})^2} \tag{22}$$

The conditional distribution of FCS is provided in formula 22. In order to compare the accuracy and compatibility of FCS and JM, we will do the same thing for FCS and estimate the mediation effect under the scenario of FCS. Then we have the following expectation of X_{1i}, X_{2i} and Y_i.

$$E(X_{1i}) = \alpha^{t-1}_{X_1} + \beta^{(t-1)}_{X_1|X_2}X_2 + \beta^{(t-1)}_{X_1|Y}Y \tag{23}$$

$$E(X_{2i}) = \alpha^{t-1}_{X_2} + \beta^{(t-1)}_{X_2|X_1}X_1 + \beta^{(t-1)}_{X_2|Y}Y \tag{24}$$

$$E(Y_i) = \alpha^{t-1}_{Y} + \beta^{(t-1)}_{Y|X_1}X_1 + \beta^{(t-1)}_{Y|X_2}X_2 \tag{25}$$

Under the scenario of JM, we have

$$\mu_{Y_i|X_1,X_2} = (\sigma^2_{X_2}\sigma_{YX_1} - \sigma_{YX_2}\sigma_{X_1X_2})X_{1i} + (\sigma^2_{X_1}\sigma_{YX_2} - \sigma_{YX_1}\sigma_{X_1X_2})X_{2i} - \mu_{X_1}(\sigma^2_{X_2}\sigma_{YX_1} - \sigma_{X_2Y}\sigma_{X_1X_2}) - \mu_{X_2}(\sigma^2_{X_1}\sigma_{YX_2} - \sigma_{X_1X_2}\sigma_{YX_1})$$

$$\mu_{X_{1i}|Y,X_2} = (\sigma^2_{X_2}\sigma_{YX_1} - \sigma_{X_1X_2}\sigma_{YX_2})Y_i + (\sigma^2_{Y}\sigma_{X_1X_2} - \sigma_{X_1Y}\sigma_{YX_2})X_{2i} - \mu_{Y}(\sigma^2_{X_2}\sigma_{X_1Y} - \sigma_{X_2X_1}\sigma_{YX_2}) - \mu_{X_2}(\sigma^2_{Y}\sigma_{X_1X_2} - \sigma_{YX_2}\sigma_{YX_1})$$

$$\mu_{X_{2i}|Y,X_1} = (\sigma^2_{X_1}\sigma_{YX_2} - \sigma_{X_1X_2}\sigma_{YX_1})Y_i + (\sigma^2_{Y}\sigma_{X_1X_2} - \sigma_{X_2Y}\sigma_{YX_1})X_{1i} - \mu_{Y}(\sigma^2_{X_1}\sigma_{X_2Y} - \sigma_{X_2X_1}\sigma_{YX_1}) - \mu_{X_1}(\sigma^2_{Y}\sigma_{X_1X_2} - \sigma_{YX_1}\sigma_{YX_2})$$

Then we used the formula above to make a comparison of the estimate of mediation effect ($\hat{\delta}$) based on the draws from the simulated data in the simulation section. The result and figure will be provided in next section.

4 Simulation Assessment

Simulation studies were conducted in order to test the performance of different Multiple Imputation (MI) methods on the mediation analysis. And the analytical comparison of two methods are displayed in previous section. Before we conducted simulation, we utilized Statewide Planning and Research Cooperative System (SPARCS) data to build up the mediation analysis model and obtain the estimate of parameters in the model. Then we use this established model to form a population with 10 million observations and sample from it. We set up evaluation criterion in 4.2 to compare the accuracy of multiple imputation methods, and we gauge the mediation effect based on section 3.3 by simulation.

4.1 Data Generation

We simulate

$$X = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \tag{26}$$

from normal distribution $x_1 \sim N(\mu_1, \sigma_1^2)$ and $x_2 \sim N(\mu_2, \sigma_2^2)$. Using the established mediation model, we set the variables to be temperature and dewpoint, and μ_1, μ_2, σ_1, σ_2 to the following values to mimic the sampled data closely. M in this simulation study is assumed to be a binary variable, for example, power outage. Based on the logistic regression and linear regression relation between variables, we simulate M and Y by the following form

$$logit(P(M = 1|x_1, x_2)) = a_0 + a_1 * x_1 + a_2 * x_2 + a_3 * x_1 * x_2, \tag{27}$$

$$P(M = 1|x_1, x_2) = \frac{1}{1 + e^{-a_0 - a_1 * x_1 - a_2 * x_2 - a_3 * x_1 * x_2}}, \tag{28}$$

$$y|M, x_1, x_2 = b_0 + b_1 * M + b_2 * x_1 + b_3 * x_2 + b_4 * x_1 * x_2, \tag{29}$$

$$I = x_1 * x_2, \tag{30}$$

all the coefficients $(a_0, a_1, ..., etc.)$ above are estimated based on SPARCS data, and this specification forms our population of size N=10,000,000. Then, we repetitively (1000 times) samples of size n=1000 from the population.

We control the effect size of mediation effect by changing a_1 and b_1 to create no mediation, low mediation, medium mediation and high mediation scenarios by MacKinnon et al. (2004). For example, when there is no mediation effect (i.e. a_1 or $b_1 = 0$, $a_1 * b_1 = 0$, $b_2 = 0.5$), medium mediation effect ($a_1 = b_1 = 0.39$, $a_1 * b_1 = 0.1521$, $b_2 = 0.5$), and large mediation effect ($a_1 = b_1 = 0.59$, $a_1 * b_1 = 0.3481$, $b_2 = 0.5$). Also, the estimate of mediation effect ($a_1 = \alpha_1$, $b_1 = \beta_1$, $a_1 * b_1$=α_1*β_1, b_2=β_2) based on real data is tested in the simulation. The percentage of missing data is set at 10%, 20%, and 30%. We will test the performance based on three different missing data mechanisms: MCAR, MAR, and MNAR (Rubin (1976)).

In order to better gauge the effect of different missing data on the mediation effect, we will only impose the missingness mechanism on Y. Firstly, we generate the data with missing data mechanism, MCAR. Suppose r_Y and r_M denote the missingness indicators for Y and M. Also, r_Y and r_M are simulated from Bernoulli distribution, i.e., r_Y, $r_M \sim Ber(P)$ where P=Pr(r_Y) for r_Y and P=Pr(r_M) for r_M. Under MCAR, these probabilities are independent of any variables whether they are observed or not. In all simulations, we set these probabilities to 0.1, 0.2 and 0.3.

When the missingness mechanism is MAR, then missingness probabilities are specified as:

$$Pr(r_M|x_1, x_2, \xi_M) = \frac{1}{1 + e^{-(a_M + b_M * x_1 + c_M * x_2)}}, \tag{31}$$

$$Pr(r_Y|x_1, x_2, M, \xi_Y) = \frac{1}{1 + e^{-(a_Y + b_Y * x_1 + c_Y * x_2 + d_Y * M)}}, \tag{32}$$

where ξ_M, ξ_Y are true parameters governing the missingness mechanisms. As for the missingness mechanism MNAR, we have the probability of missingness is related to the incompletely observed variable itself, the probability of missingness value probability is calculated as follows:

$$Pr(r_M|x_1, x_2, M, \xi_M) = \frac{1}{1 + e^{-(a_M + b_M * x_1 + c_M * x_2 + d_M * M)}}, \tag{33}$$

$$Pr(r_Y|x_1, x_2, M, Y, \xi_Y) = \frac{1}{1 + e^{-(a_Y + b_Y * x_1 + c_Y * x_2 + d_Y * M + e_Y * Y)}}, \tag{34}$$

4.2 Evaluation criterion

We utilize four different criteria to assess the consistency and accuracy for each method under all conditions. As mention previously, the mediation analysis model and the simulation work were established based on SPARCS data. Thus, the true mediation effect was estimated based on the SPARCS data. We will use the obtained mediation effect as true value to test the accuracy of different imputation methods. $\hat{\delta}$, $\hat{l}_r$, $\hat{u}_r$ are all obtained by simulated data.

The first criterion to assess the performance is the coverage rate (CR):

$$CR = \frac{\sum I_r(\hat{l}_r < \delta < \hat{u}_r)}{1000}, \tag{35}$$

where I_r is an indicator function such that

$$I_r = \begin{cases} 1 & \delta \in (\hat{l}_r, \hat{u}_r) \\ 0 & \delta \notin (\hat{l}_r, \hat{u}_r). \end{cases} \tag{36}$$

$\sum I_r(\hat{I}_r < \delta < \hat{u}_r)$ represents the total number of sample mediation effect confidence interval that cover the true population mediation mediation effect. Let $\hat{l}_r$ and $\hat{u}_r$ denote the lower and upper limits of the 95% confidence interval for the mediation effect at r^{th} sample, and they were estimated by simulation work. When we compare different methods with regard to the coverage rate, the method with a higher value in coverage rate is better than other methods.

Secondly, we estimate the bias under each of the MI approach. Let $\delta = a * b$ denote the true mediation effect, let $\hat{\delta}$=$\hat{a}_r$*$\hat{b}_r$,for $r = 1, \ldots, 1000$, denote the mediation effect estimated in our r^{th} simulation experiment (i.e. r^{th} sample of $n = 2000$ for our problem). The bias is calculated as

$$\hat{Bias} = \frac{1}{1000} * \sum_{r=1}^{1000} |\hat{\delta}_r - \delta|, \tag{37}$$

Bias in this case is estimated as a percentage of how much $\hat{\delta}_r$ is deviated from the δ. When we are comparing bias for different methods, we could make the conclusion that the method that has the smaller bias of the δ is better than other methods.

The third criterion we consider is average width of confidence interval (AW). AW is used to calculate the distance between the average lower and upper confidence interval limits, and it is defined as

$$AW = \frac{\sum_{r=1}^{1000}(\hat{u}_r - \hat{l}_r)}{1000}, \tag{38}$$

when we detect that the method has a lower AW than others, we can conclude that the method has a more consistent result than others.

The fourth criterion is mean square error (MSE). The formula of how to calculate MSE is provided as follows

$$MSE = \frac{1}{1000} * \sum_{r=1}^{1000} (\hat{\delta} - \delta)^2, \tag{39}$$

when we compare different methods with regard to MSE, the method with a lower value in MSE is better than other methods. We should consider these criterions together. A method that has a low bias, small AW, high CR and low MSE has greater accuracy and higher power.

Among four different criterions to assess the consistency and accuracy for each method under all conditions, coverage rate is the most important criterion to show the performance of imputation method under mediation analysis model.

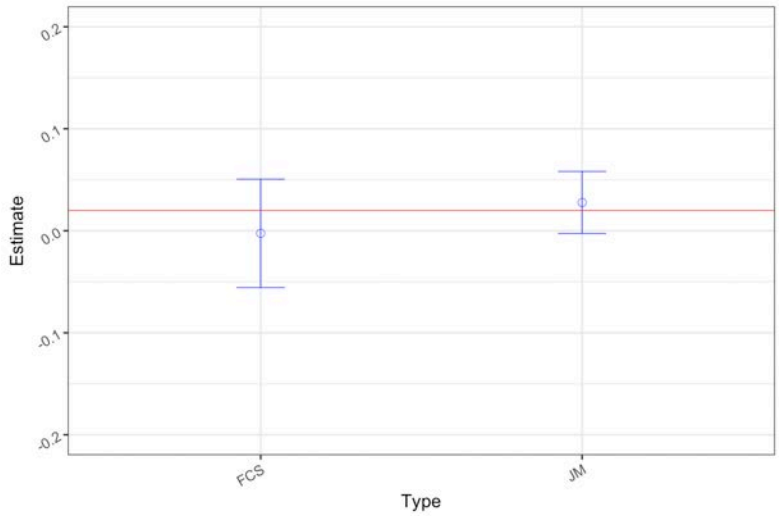

Figure 2: Comparison of FCS and JM

4.3 Results and Summary

The simulation study provides two perspectives of analyzing the consistency of imputation methods in mediation analysis model with missing values. In section 4.3.1, we built up values for each variables based on its conditional distribution and made a theoretical comparison between FCS and JM. In section 4.3.2, the results were acquired through imputation on mediation analysis model by varying mediation effect.

4.3.1 Analytic comparison

We used the simulated data to estimate the mediation effect based on JM and FCS and compared with the true value of the mediation effect according to section 3.3. The whole process was repeated for 1000 times. The confidence interval for mediation effect was estimated. In Figure 2, we have shown the comparison of mediation effect by FCS amd JM with confidence interval. The confidence interval for the mediation effect generated by JM is 0.028 (-0.002, 0.058). The confidence interval for mediation effect generated by FCS imputation is -0.0025 (-0.0557, 0.0506). The true mediation effect we consider in this case is 0.02, which is covered in both scenarios. However, we found that JM imputation does not only have a narrower confidence interval for the estimate, it is also closer to the estimate.

4.3.2 Estimation and performance

In this simulation, we compared the performance of FCS and JM methods in regard of data with missing values under mediation analysis model. From table 1 to table 4, we show the different coverage rate for none mediation effect condition ($a = b = 0$, acme $= ab = 0$), low mediation effect condition ($a = b = 0.1$, acme $= ab = 0.01$), medium mediation condition ($a = b = 0.39$, acme $= ab = 0.1521$) and high mediation ($a = b = 0.59$, acme $= ab = 0.3481$)

condition (MacKinnon et al. (2004)), respectively. The imputation methods applied to the datasets with missing values are fully conditional specification and joint modeling; the missing data rate are set at 10%, 20%, and 30%; the missingness pattern in the analysis are missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). We tested the coverage rate for conditions including before data deletion (BD), complete case (CC) and imputed data (IMP).

Table 1: Coverage rate for none mediation effect (ACME=0)

Pattern	*Rate*	*BD*	*CC*	*IMP*
Fully conditional specification				
mcar	0.1	0.91	0.89	0.95
	0.2	0.92	0.94	0.94
	0.3	0.91	0.95	0.88
mar	0.1	0.93	0.60	0.94
	0.2	0.92	0.53	0.92
	0.3	0.92	0.49	0.92
mnar	0.1	0.94	0.79	0.92
	0.2	0.93	0.63	0.95
	0.3	0.95	0.60	0.96
Joint modeling				
mcar	0.1	0.86	0.89	0.98
	0.2	0.96	0.85	0.95
	0.3	0.97	0.85	0.95
mar	0.1	0.89	0.19	0.94
	0.2	0.87	0.13	0.92
	0.3	0.92	0.05	0.90
mnar	0.1	0.92	0.39	0.99
	0.2	0.89	0.31	0.90
	0.3	0.89	0.25	0.93

note. Pattern: missingness pattern type; Rate: missing data rate; BD: before deletion; CC: complete case; IMP: imputed data.

In Table 1, we found that FCS and JM had similar performance in imputing missing data with none mediation effect condition. We can see that the coverage rate is consistent for both FCS and JM when missingness pattern is MCAR for different missing rate. As for MAR, the coverage rate of acme will decrease dramatically as the missing rate increases. On the contrary, when FSC and JM are applied to impute the missing values, we found that the coverage rate will be close to BD group.

In Table 2, we found that FCS and JM had different performance in imputing missing data with low mediation effect condition. In the condition of low mediation effect, We found that the coverage rate result is similar to none mediation effect for FCS. However, when we used JM to impute missing values, we found that the coverage rate was declining fast with increasing missing rate. Especially, when we have 30% of data missing under MNAR, the coverage rate is 53%.

In Table 3, we found that FCS and JM had similar performance in imputing missing data with medium mediation effect condition. The coverage rate is extremely low (2%) when we had 30% of data missing in the condition of MNAR. However, the imputed data by FCS could provide a coverage rate of 87%, which is much greater than the complete case. Also, the coverage rate also worked well for JM when we had medium mediation effect.

In Table 4, we compared the the performance of FCS and JM in the condition of high mediation effect.We can see that the coverage rate is consistent for both FCS and JM when missingness pattern is mcar for different missing rate. Under MAR and MNAR missingness mechanism scenarios, we observed very poor coverage rate for the mediation effect. The coverage rate for joint and variable-by-variable imputation models are still consistent, and

Table 2: Coverage rate for low mediation effect (ACME=0.01)

Pattern	*Missing Rate*	*BD*	*CC*	*IMP*
Fully conditional specification				
mcar	0.1	0.98	0.92	0.94
	0.2	0.95	0.98	0.94
	0.3	0.97	0.94	0.97
mar	0.1	0.87	0.14	0.97
	0.2	0.95	0.08	0.95
	0.3	0.97	0.02	0.65
mnar	0.1	0.93	0.23	0.95
	0.2	0.95	0.16	0.89
	0.3	0.95	0.10	0.87
Joint modeling				
mcar	0.1	0.97	0.91	0.90
	0.2	0.95	0.92	0.91
	0.3	0.91	0.84	0.89
mar	0.1	0.99	0.38	0.97
	0.2	0.91	0.60	0.83
	0.3	0.95	0.22	0.84
mnar	0.1	0.98	0.14	0.85
	0.2	0.97	0.10	0.59
	0.3	0.95	0.02	0.53

note. Pattern: missingness pattern type; Rate: missing data rate; BD: before deletion; CC: complete case; IMP: imputed data.

they outperform complete case data a lot.

5 Discussion

In the scenario of no mediation effect, low mediation effect, medium mediation effect and high mediation effect, the results are generally consistent in coverage rate for imputation models. Variable-by-variable model outperforms joint modeling when we have high missingness rate. In addition, we use conditional distribution method to estimate the mediation effect, and find the mediation analysis model is not impacted by the specific choice of imputation models.

The study has couples of limitations. The comparison of JM and FCS in this paper is based on a simulation study. We tried to use simulation study to prove and verify our original guess. However, simulation study may potentially introduce some uncertainties to the data. Future researches could examine a complicated model of mediation analysis. Secondly, variable types could be much more complicated than current considered scenarios. In the mediation analysis, response variable, independent variables and mediators could be ordinal, nominal and survival. So, there are still a lot of unknown scenarios. Thirdly, there is multi-level mediation model available for analysis, and researchers may choose multi-level data to develop the model, some future work will be discussed for addressing the missing data problem in multi-level mediation analysis problem.

Table 3: Coverage rate for medium mediation effect (ACME=0.1521)

Pattern	*Rate*	*BD*	*CC*	*IMP*
Fully conditional specification				
mcar	0.1	0.99	0.97	0.99
	0.2	0.91	0.88	0.96
	0.3	0.89	0.86	0.94
mar	0.1	0.89	0.23	0.75
	0.2	0.97	0.14	0.64
	0.3	0.98	0.02	0.63
mnar	0.1	0.98	0.15	0.93
	0.2	0.91	0.80	0.90
	0.3	0.97	0.02	0.87
Joint modeling				
mcar	0.1	0.99	0.84	0.99
	0.2	0.99	0.77	0.99
	0.3	0.99	0.80	0.99
mar	0.1	0.99	0.51	0.97
	0.2	0.89	0.33	0.94
	0.3	0.91	0.08	0.89
mnar	0.1	0.97	0.67	0.97
	0.2	0.95	0.33	0.98
	0.3	0.99	0.13	0.89

note. Pattern: missingness pattern type; Rate: missing data rate; BD: before deletion; CC: complete case; IMP: imputed data.

Table 4: Coverage rate for high mediation effect (ACME=0.3487)

Pattern	*Missing Rate*	*BD*	*CC*	*IMP*
Fully conditional specification				
mcar	0.1	0.91	0.98	0.92
	0.2	0.94	0.95	0.98
	0.3	0.95	0.95	0.97
mar	0.1	0.87	0.69	0.81
	0.2	0.98	0.33	0.69
	0.3	0.87	0.09	0.66
mnar	0.1	0.90	0.56	0.89
	0.2	0.90	0.53	0.74
	0.3	0.93	0.11	0.71
Joint modeling				
mcar	0.1	0.97	0.86	0.99
	0.2	0.98	0.83	0.99
	0.3	0.99	0.88	0.98
mar	0.1	0.98	0.65	0.93
	0.2	0.95	0.56	0.70
	0.3	0.98	0.32	0.62
mnar	0.1	0.95	0.46	0.84
	0.2	0.98	0.26	0.80
	0.3	0.98	0.19	0.76

note. Pattern: missingness pattern type; Rate: missing data rate; BD: before deletion; CC: complete case; IMP: imputed data.

References

Akande, O., Li, F. & Reiter, J. (2017), 'An Empirical Comparison of Multiple Imputation Methods for Categorical Data', *American Statistician* **71**(2), 162–170.
URL: *https://doi.org/10.1080/00031305.2016.1277158*

Campion, W. M. & Rubin, D. B. (1989), 'Multiple Imputation for Nonresponse in Surveys', *Journal of Marketing Research* **26**(4), 485.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977), *Maximum Likelihood from Incomplete Data Via the EM Algorithm* , Vol. 39.

Fritz, M. S. & MacKinnon, D. P. (2008), 'A graphical representation of the mediated effect', *Behavior Research Methods* **40**(1), 55–60.

Hughes, R. A., White, I. R., Seaman, S. R., Carpenter, J. R., Tilling, K. & Sterne, J. A. (2014), 'Joint modelling rationale for chained equations', *BMC Medical Research Methodology* **14**(1).

Little, R. J. A. & Rubin, D. B. (1988), *Statistical Analysis with Missing Data.*, Vol. 151.

MacKinnon, D. P., Lockwood, C. M. & Williams, J. (2004), 'Confidence Limits for the Indirect Effect: Distribution of the Product and Resampling Methods', *Multivariate Behavioral Research* **39**(930578742), 1–24.

Mistler, S. A. & Enders, C. K. (2017), 'A Comparison of Joint Model and Fully Conditional Specification Imputation for Multilevel Missing Data', *Journal of Educational and Behavioral Statistics* **42**(4), 432–466.

Raghunathan, T., Lepkowski, J., Van Hoewyk, J. & Solenberger, P. (2001), 'A multivariate technique for multiply imputing missing values using a sequence of regression models', *Survey methodology* **27**(1), 85–96.

Rubin, D. B. (1976), 'Biometrika Trust Inference and Missing Data Author (s): Donald B . Rubin Published by : Oxford University Press on behalf of Biometrika Trust Stable URL : http://www.jstor.org/stable/2335739 Accessed : 12-06-2016 21 : 34 UTC', *Biometrika* **63**(3), 581–592.

Schafer, J. L. (1999), 'Multiple imputation: a primer', *Statistical Methods in Medical Research 1999;* **2802**(99).

Schafer, J. L. J. L. (1997), *Analysis of incomplete multivariate data*, Chapman & Hall, London.

van Buuren, S. (2007), 'Multiple imputation of discrete and continuous data by fully conditional specification', *Statistical Methods in Medical Research* **16**(3), 219–242.

van Buuren, S. & Groothuis-Oudshoorn, K. (2011), 'mice: Multivariate imputation by chained equations in R', *Journal of Statistical Software* **45**(3), 1–67.

Yucel, R. M., Zhao, E., Schenker, N. & Raghunathan, T. E. (2018), 'Sequential hierarchical regression imputation', *Journal of Survey Statistics and Methodology* **6**(1), 1–22.

Zhang, Z. & Wang, L. (2013), 'Methods for Mediation Analysis with Missing Data', *Psychometrika* **78**(1), 154–184.