Fall 2016

Multiple Imputation of Missing Data in Structural Equation Models with Mediators and Moderators Using Gradient Boosted Machine Learning

Robert J. Milletich II
Old Dominion University

MULTIPLE IMPUTATION OF MISSING DATA IN STRUCTURAL EQUATION MODELS WITH MEDIATORS AND MODERATORS USING GRADIENT BOOSTED MACHINE LEARNING

by

Robert J. Milletich II
B.S. May 2009, Old Dominion University
M.S. May 2012, Old Dominion University
M.S. 2015, Old Dominion University

A Dissertation Submitted to the Faculty of
Old Dominion University in Partial Fulfillment of the
Requirements for the Degree of

DOCTOR OF PHILOSOPHY

APPLIED PSYCHOLOGICAL SCIENCES

OLD DOMINION UNIVERSITY
December 2016

Approved by:

Michelle L. Kelley (Director)

James F. Paulson (Member)

Norou Diawara (Member)

ABSTRACT

MULTIPLE IMPUTATION OF MISSING DATA IN STRUCTURAL EQUATION MODELS WITH MEDIATORS AND MODERATORS USING GRADIENT BOOSTED MACHINE LEARNING

Robert J. Milletich II
Old Dominion University, 2016
Director: Dr. Michelle L. Kelley

Mediation and moderated mediation models are two commonly used models for indirect effects analysis. In practice, missing data is a pervasive problem in structural equation modeling with psychological data. Multiple imputation (MI) is one method used to estimate model parameters in the presence of missing data, while accounting for uncertainty due to the missing data. Unfortunately, commonly used MI methods are not equipped to handle categorical variables or nonlinear variables such as interactions. In this study, we introduce a general MI framework that uses the Bayesian bootstrap (BB) method to generate posterior inferences for indirect effects and gradient boosted machine learning imputation models that can impute missing data in linear and logistic regression models with linear and nonlinear effects.

Two Monte Carlo simulation studies are conducted to examine the empirical performance of a BB procedure for estimation and inference of indirect effects and to examine the performance of the proposed imputation algorithm in indirect effects analysis. Results show that the BB has comparable performance to widely used frequentist methods (e.g., delta methods and nonparametric bootstrap with bias-correction) for indirect effects analysis for a variety of models and conditions. With missing data, in general, results indicate that the proposed MI framework has comparable performance to model-based

estimation and other MI algorithms for indirect effects analysis in mediation models; for indirect effects analysis in moderated mediation models, results indicate that the proposed MI framework outperforms these methods in most conditions. Advantages and limitations of the BB as applied to indirect effects analysis are discussed.

This dissertation is dedicated to my wife, family, and friends.

ACKOWLEDGEMENTS

I would first like to thank my advisor Dr. Michelle Kelley for her continuous support during the completion of this dissertation and throughout my graduate school education. Her immense knowledge and passion for science afforded me the opportunity to work on many challenging and interesting projects, and for that I am forever grateful. I would also like to thank my dissertation committee: Dr. James Paulson and Dr. Norou Diawara for their insightful comments and encouragement during this dissertation. I also owe many thanks to the expertise of Old Dominion University's high performance computing team for helping me run my simulations on the school cluster. Lastly, I would like to thank my family, especially my wonderful wife, Rita, for her patience, support, and encouragement.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table Page

Table Page

LIST OF FIGURES

Figure Page

CHAPTER 1

INTRODUCTION

1. 1. Overview

Mediation and moderation analyses are two commonly used statistical techniques in psychological research. As described by Baron and Kenny (1986), a mediator is a variable that can explain the association between an input variable and an output variable, whereas, a moderator is a variable that affects the direction or magnitude of the relation between an input variable and an output variable. Although statistically distinct, these two techniques have been combined into so-called moderated mediation models (Preacher, Rucker, & Hayes, 2007). Moderated mediation models were developed to test hypotheses about conditional indirect effects, that is, whether a mediation (or indirect) effect between an input variable and outcome variable is influenced (or moderated) by one or more variables. Currently in practice, structural equation modeling is the most common statistical framework used to test for mediation and moderated mediation effects (MacKinnon, Lockwood, & Williams, 2004; Preacher et al., 2007).

Unfortunately, a common problem in structural equation modeling with psychological data is that of missing data. Although the causes of missing data are numerous, missing data can be described based on the pattern and type of missingness. As the name implies, the pattern of missingness refers to the manifested pattern of missing values in a data set (e.g., univariate nonresponse, monotone, general). The type of missingness refers to the underlying functional mechanism that caused the missing values. According to Rubin (1976), the type of missing data can be categorized into three distinct

categories: (1) missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR).

The techniques for handling missing data include deletion methods (e.g., listwise deletion, pairwise deletion), model-based estimation (e.g., full-information maximum likelihood; Archbuckle, 1996), single imputation methods (e.g., unconditional mean imputation, conditional mean imputation), and multiple imputation (MI) methods (e.g., joint model MI, fully-conditional specification MI; Rubin, 1987; van Buuren, 2007). Theoretically, both the pattern and type of missingness determine which missing data techniques are appropriate for obtaining statistically unbiased and valid parameter estimates in the presence of missing data (Little & Rubin, 2002). In practice, other factors such as the sample size and percentage of missingness influence the choice of missing data technique. Research has shown that in large samples with small amounts of missingness (e.g., $< 5\%$) and a MCAR mechanism, the choice of missing data technique is often negligible in parameter estimation (Little & Rubin, 2002). On the contrary, in more practical cases (e.g., when data are MAR and missingness $> 5\%$), extensive research has continually shown the naïve approaches such as listwise deletion or unconditional mean imputation are inferior to model-based estimation and MI approaches for handling missing data (van Buuren, 2012).

The gold standard model-based estimation approach to missing data in structural equation models is model-based estimation (e.g., full-information maximum likelihood [FIML] and weighted-least squares [WLS]). Research has shown FIML (Enders & Bandalos, 2001) and WLS (Asparouhov & Muthén, 2010) to work well in SEMs, however, this method can fail to converge in scenarios in which the likelihood is too complicated to optimize or

there is sparse data (Rubin & Little, 2002). In this case, the more powerful approach is MI. In short, a MI procedure consists of generating M plausible data sets based on the observed data to reflect the uncertainty about the missing data. Complete-data methods are then applied to each pseudo-complete data set and the set of M estimates are pooled together based on Rubin's (1996) rules for pooling to obtain multiply imputed parameter estimates. With the increase in software providing implementations of fully-conditional specification (FCS) algorithms, MI has become increasingly attractive, especially because these algorithms can handle mixed data types, auxiliary variables, and large amounts of missing data.

When using MI techniques, the problem of missing data can be viewed from a supervised machine learning perspective. For instance, in a typical supervised machine learning problem, a training data set is used to estimate or learn the relationship between a set of inputs and labels (e.g., discrete values for classification or continuous values for regression). Then, the trained model is applied on an independent testing data set where the labels are unknown to predict the missing labels. Thus, the goal of supervised machine learning is to build a predictive model based on available data with both inputs and labels, and use this predictive model to make predictions about unknown labels on independent data sets. Note, the term supervised learning is used because labels are provided in the training set. From a missing data perspective, when data are missing on a variable $\mathbf{x}_j$, we can partition the data set into two sets, an observed (training) data set and a missing (testing) data set based on the observed and missing rows for $\mathbf{x}_j$, respectively. The observed (or training) data in $\mathbf{x}_j$ can be used as the labels in a parametric or nonparametric model and the rest of the variables in the data used as inputs. Then, the model can be

trained and applied to predict the missing values in $\mathbf{x}_j$ from the missing (or testing) set. Therefore, we can easily see that the goals of MI and supervised machine learning are similar: Build predictive models based on observed inputs and labels to predict labels on data in which only inputs are available.

The benefit of viewing missing data from a supervised machine learning vantage is that all of the parametric (e.g., linear regression, logistic regression) and nonparametric (tree-based models, nearest neighbors) models widely used in machine learning applications can easily be applied to predict missingness on any variable. This approach is another way of viewing van Buuren's (2012) framework for FCS using the multiple imputation by chained equations (MICE) algorithm. In moderated mediation analyses, however, a notable shortcoming of the MICE algorithm is that it does not readily incorporate nonlinearities such as interactions in the imputation process. Research has shown that ignoring nonlinearities in the imputation model results in biased statistical estimates and inferences for regression coefficients of nonlinear terms (Bartlett, Seaman, White, & Carpenter, 2014; Doove, van Buuren, & Dusseldorp, 2014; Enders, Baraldi, & Cham, 2014; von Hippel, 2009; Seaman, Bartlett, & White, 2012). This study proposes a novel MI algorithm to automatically model linear and nonlinear effects in substantive models using gradient boosted machine learning models. A Bayesian bootstrap resampling strategy is used to generate posterior distributions for Bayesian inference.

1.2. General Structural Equation Model for Mediation-Type Analyses

Structural equation modeling is a flexible statistical framework designed to develop and test conceptual models. A structural equation model (SEM) has two components: (1) a measurement model used to define latent variables using one or more observed indicator

variables, and (2) a structural model that links together the latent variables. For testing mediation and moderated mediation effects, the most common SEM is path analysis. Path analysis can be viewed as a special case of structural equation modeling, one in which only single indicator variables are specified for each of the latent variables in the structural model (Bollen, 1989). As such, the latent variables are actually just the observed variables.

A general SEM with observed variables (or path model) that can be used for mediation-type models is given by

$$\boldsymbol{\nu} = \boldsymbol{\alpha} + \boldsymbol{\beta}\boldsymbol{\nu} + \boldsymbol{\zeta} \tag{1.1}$$

where $\boldsymbol{\nu}$ is a $p \times 1$ vector of observed variables, $\boldsymbol{\alpha}$ is a $p \times 1$ vector of intercepts, $\boldsymbol{\beta}$ is a $p \times p$ matrix of regression coefficients, and $\boldsymbol{\zeta}$ is a $p \times 1$ random vector of errors. It is assumed that $\boldsymbol{\zeta}$ is independent of $\boldsymbol{\nu}$ and $(\mathbf{I} - \boldsymbol{\beta})$ is non-singular. As such, Equation (1.1) can be expressed as

$$\boldsymbol{\nu} - \boldsymbol{\beta}\boldsymbol{\nu} = \boldsymbol{\alpha} + \boldsymbol{\zeta}$$

$$(\mathbf{I} - \boldsymbol{\beta})\boldsymbol{\nu} = \boldsymbol{\alpha} + \boldsymbol{\zeta}$$

$$\boldsymbol{\nu} = (\mathbf{I} - \boldsymbol{\beta})^{-1}(\boldsymbol{\alpha} + \boldsymbol{\zeta}).$$

When all variables in $\boldsymbol{\nu}$ are continuous, the distribution of $\boldsymbol{\zeta}$ is often modeled using a multivariate normal distribution with zero mean vector and covariance matrix $\boldsymbol{\Psi}$. When one or more variables in $\boldsymbol{\nu}$ are discrete, however, the distribution of $\boldsymbol{\zeta}$ becomes more complicated and assuming a multivariate normal distribution may bias statistical inference (Lee, 2007). Fortunately, in models where it is assumed that $\mathrm{Cov}(\zeta_i, \zeta_j) = 0$ for $i, j = 1,2, \ldots, p$ where $i \neq j$, estimating the joint regression model $f(\boldsymbol{\nu}|\boldsymbol{\theta})$ is equivalent to estimating simpler, often univariate regression models. This latter point provides flexibility in models with mixed endogenous variable types. Specific model simplifications will be

discussed later.

The goal in mediation models is to decompose the influence of a variable or set of variables on another variable or set of variables into direct, indirect, and total effects (Fox, 1980). In all decompositions, the total effects are equal to the sum of the direct effects and the indirect effects. Direct effects are those effects unmediated by another variable in the model. Figure (1.1) shows a graphical realization of a

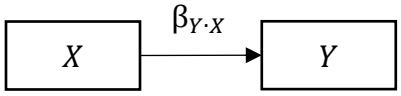


Figure 1.1. Graphical representation of a direct effect X on Y based on the linear model given in (1.2). The intercept and error are omitted for clarity.

direct effect of X on Y based on the simple linear model

$$\begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} \alpha_X \\ \alpha_Y \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ \beta_{Y \cdot X} & 0 \end{bmatrix} \begin{bmatrix} X \\ Y \end{bmatrix} + \begin{bmatrix} \zeta_X \\ \zeta_Y \end{bmatrix} \tag{1.2}$$

where X is the exogenous variable, Y is the endogenous variable, α_X is the intercept for the exogenous variable model, α_Y is the intercept for the endogenous variable model, $\beta_{Y \cdot X}$ is the direct effect of X on Y, and ζ_X and ζ_Y are the error terms for the exogenous variable model and endogenous variable model, respectively. Indirect effects, also referred to as unconditional indirect effects, are those effects mediated by at least one intervening variable. Figure (1.2) shows a graphical realization of an

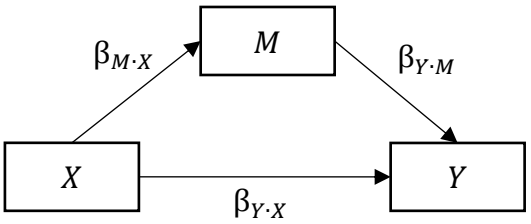


Figure 1.2. Graphical representation of an indirect effect X on Y via a mediating variable M based on the simple mediation model given in (1.3). The intercept and error are omitted for clarity.

indirect effect X on Y via the mediating variable M based on the simple mediation model

$$\begin{bmatrix} X \\ M \\ Y \end{bmatrix} = \begin{bmatrix} \alpha_X \\ \alpha_M \\ \alpha_Y \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ \beta_{M\cdot X} & 0 & 0 \\ \beta_{Y\cdot X} & \beta_{Y\cdot M} & 0 \end{bmatrix} \begin{bmatrix} X \\ M \\ Y \end{bmatrix} + \begin{bmatrix} \zeta_X \\ \zeta_M \\ \zeta_Y \end{bmatrix}, \tag{1.3}$$

where X is the exogenous variable, M is the mediator variable, Y is the endogenous variable, α_X is the intercept for the exogenous variable model, α_M is the intercept for the mediator variable model, α_Y is the intercept for the endogenous variable model, $\beta_{M\cdot X}$ is the direct effect of X on M, $\beta_{Y\cdot X}$ is the direct effect of X on Y, controlling for M, $\beta_{Y\cdot M}$ is the direct effect of M on Y, controlling for X, and ζ_X, ζ_M, ζ_Y are the error terms for the exogenous variable model, mediator variable model, and endogenous variable model, respectively.

More recently, attention has focused on examining how indirect effects vary across levels of another variable or set of variables. These effects are called moderated mediation effects or conditional indirect effects. In a simple case, if Y is an endogenous variable and X and W are two exogenous variables, a moderation (or interaction) effect exists if the regression of Y on X is conditional upon the value of W. In other words, the direct effect of X on Y is not additive across values of W. Figure (1.3) shows a graphical realization of a

moderated effect X on Y conditional

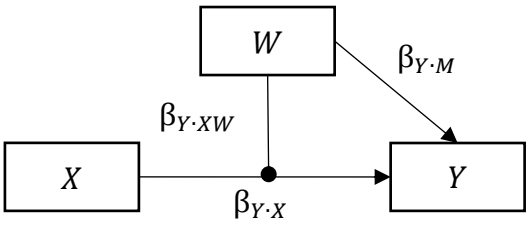


Figure 1.3. Graphical representation of a moderation effect X on Y conditional on W variable M based on the moderation model given in (1.4). The intercept, errors, and covariance between X and W are omitted for clarity.

on the variable W based on the linear model,

$$\begin{bmatrix} X \\ W \\ XW \\ Y \end{bmatrix} = \begin{bmatrix} \alpha_X \\ \alpha_W \\ \alpha_{XW} \\ \alpha_Y \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \beta_{Y\cdot X} & \beta_{Y\cdot W} & \beta_{Y\cdot XW} & 0 \end{bmatrix} \begin{bmatrix} X \\ W \\ XW \\ Y \end{bmatrix} + \begin{bmatrix} \zeta_X \\ \zeta_W \\ \zeta_{XW} \\ \zeta_Y \end{bmatrix} \qquad (1.4)$$

where X is the exogenous variable, W is the moderator variable, XW is the interaction between X and W, Y is the endogenous variable, α_X is the intercept for the exogenous variable model, α_W is the intercept for the moderator variable model, α_{XW} is the intercept for the moderation effect model, α_Y is the intercept for the endogenous variable model, $\beta_{Y\cdot X}$ is the direct effect of X on Y, $\beta_{Y\cdot W}$ is the direct effect of W on Y, $\beta_{Y\cdot XW}$ is the direct effect of XW on Y, and $\zeta_X, \zeta_W, \zeta_{XW}, \zeta_Y$ are the error terms for the exogenous variable model, moderator variable model, moderated effect model, and endogenous variable model, respectively.

In their seminal paper, Preacher et al. (2007) describe five basic moderated mediation models. Let X be the exogenous variable, W and Z be moderating variables, M be the

mediating variable, and Y be the endogenous variable. Then, the five moderated mediation models described in Preacher et al. are as follows:

1. X moderates the direct effect of M to Y
2. W moderates the direct effect of X to M
3. W moderates the direct effect of M to Y
4. W moderates the direct effect of X to M and Z moderates the direct effect of M to Y
5. W moderates both the direct effect of X to M and the direct effect of M to Y

Figure (1.4) presents a graphical representation of these five models. A sixth moderated mediation model is described in Wang and Preacher (2015) based on combining Model 4 and Model 5 from Figure (1.4); however, the conditional indirect effect under this model is hard to interpret. Therefore, we will not discuss their model further.

1.3. Point Estimators for Unconditional and Conditional Indirect Effects

There are three methods for estimating unconditional indirect effects. The first method is the causal steps method. The causal steps method involves testing separate null hypotheses about the direct effects in a SEM. For example, to show that X is indirectly related to Y via M, the Baron and Kenny (1986) method reparametrizes the model in (1.3) using the following three regression models,

1. $Y = \alpha'_Y + \beta'_{Y\cdot X}X + \zeta'_Y$
2. $M = \alpha_M + \beta_{X\cdot M}X + \zeta_M$ (1.5)
3. $Y = \alpha_Y + \beta_{Y\cdot X}X + \beta_{Y\cdot M}M + \zeta_Y$.

From these regression equations, four conditions must hold: (1) X is significantly correlated with Y ($\beta'_{Y\cdot X}$ in Model [1]), (2) X is significantly correlated with M ($\beta_{X\cdot M}$ in Model [2]), (3) M is significantly related to Y after controlling for X ($\beta_{Y\cdot M}$ in Model [3]), and

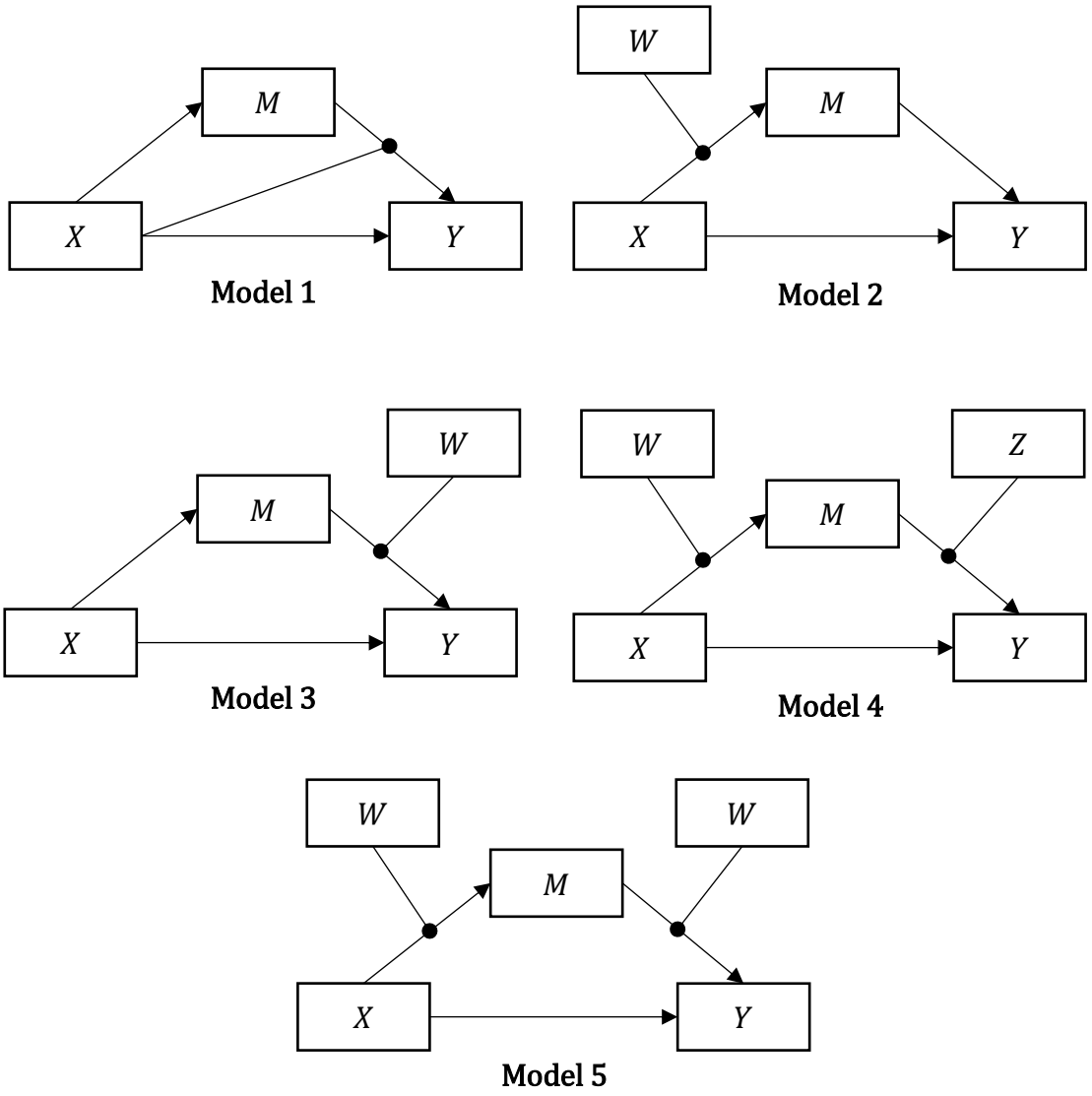


Figure 1.4. Path diagrams representing models described in Preacher et al. (2007). Intercepts, errors, and covariances are omitted for clarity.

(4) after controlling for M, the association between X and Y should be zero ($\beta_{Y \cdot X}$ in Model [3]). If all four of these steps are met, then the data are consistent idea that M completely mediates the relationship between X and Y. However, as is often the case in practice, the first three steps hold but in the fourth step $\beta_{Y \cdot X} \neq 0$; if $\beta_{Y \cdot X}$ is nonzero, but considered small

and negligible in magnitude (small is relative to the data set), then partial mediation is said to occur (MacKinnon, 2008).

In the difference of coefficients method, the indirect effect is conceptualized by the same series of three regression models given in (1.5), In Model 1, the total effect, denoted by $\beta'_{Y\cdot X}$, of X on Y is calculated. In Model 2, the relationship between X and M is established. Lastly, in Model 3, the direct effect of X on Y after controlling for M is calculated, denoted by $\beta_{Y\cdot X}$. The point estimator of the indirect effect given by the difference of coefficients estimator is

$$g(\boldsymbol{\beta}|\mathbf{v}) = \beta'_{Y\cdot X} - \beta_{Y\cdot X}. \tag{1.6}$$

Using the notation in (1.6), $g(\boldsymbol{\beta}|\mathbf{v})$ denotes a function $g(\cdot)$ of the direct effects $\boldsymbol{\beta}$, conditional on the observed data $\mathbf{v}$. The difference in coefficients method estimates the indirect effect as the difference between the total effect $\beta'_{Y\cdot X}$ and the direct effect $\beta_{Y\cdot X}$.

Lastly, the more common method for calculating indirect effects is the product of coefficients. The product of coefficients estimator for indirect effect based on the mediation model in (1.3) is simply $g(\boldsymbol{\beta}|\mathbf{v}) = \beta_{M\cdot X}\beta_{Y\cdot M}$, or the product of direct effect of X on M and the direct effect of M on Y, controlling for X. Note, it can be shown that in the case of continuous variables, the difference of coefficients and product of coefficients estimators lead to the same point estimator for an indirect effect, that is,

$$\beta'_{Y\cdot X} - \beta_{Y\cdot X} = \beta_{M\cdot X}\beta_{Y\cdot M} \tag{1.7}$$

Following the result in MacKinnon, Warsi, and Dwyer (1995), using the model specified in (1.5), the maximum likelihood estimate for $\beta'_{Y\cdot X}$ is

$$\beta'_{Y\cdot X} = \frac{Cov(X,Y)}{Var(\mathrm{X})}. \tag{1.8}$$

In the numerator, the covariance between X and Y is given by

$$\begin{aligned} Cov(X,Y) &= Cov(X, \alpha_Y + \beta_{Y\cdot X}X + \beta_{Y\cdot M}M + \epsilon_Y) \\ &= \beta_{Y\cdot X}Cov(X,X) + \beta_{Y\cdot M}Cov(X,M) \\ &= \beta_{Y\cdot X}Var(\mathrm{X}) + \beta_{Y\cdot M}Cov(X,M). \end{aligned} \tag{1.9}$$

Substituting the covariance (1.9) into (1.8),

$$\beta'_{Y\cdot X} = \frac{\beta_{Y\cdot X}Var(\mathrm{X}) + \beta_{Y\cdot M}Cov(X,M)}{Var(\mathrm{X})}$$

$$\beta'_{Y\cdot X} = \beta_{Y\cdot X} + \beta_{Y\cdot M}\frac{Cov(X,M)}{Var(\mathrm{X})}$$

$$\beta'_{Y\cdot X} = \beta_{Y\cdot X} + \beta_{Y\cdot M}\beta_{M\cdot X}$$

$$\beta'_{Y\cdot X} - \beta_{Y\cdot X} = \beta_{Y\cdot M}\beta_{M\cdot X},$$

where $\beta_{M\cdot X} = \frac{Cov(X,M)}{Var(\mathrm{X})}$ is the maximum likelihood estimate. However, when there are discrete endogenous variables in a model, the equality in (1.7) does not hold. Current research suggests that the product of coefficients estimator is the most flexible because it extends to more complicated models that include multiple mediators and categorical endogenous variables (Enders, Fairchild, & MacKinnon, 2013; MacKinnon, Lockwood, Brown, Wang, & Hoffman, 2007).

Bollen (1989) describes a general method for determining the point estimator for an indirect effect of $\boldsymbol{\nu}$ on $\boldsymbol{\nu}$ based on any SEM. Based on the sum of powers of coefficient matrices, Bollen defines the total effects of $\boldsymbol{\nu}$ on $\boldsymbol{\nu}$, $\mathbf{T}_{\boldsymbol{\nu\nu}}$, as

$$\mathbf{T}_{\boldsymbol{\nu\nu}} = \sum_{k=1}^{\infty} \boldsymbol{\beta}^k, \tag{1.10}$$

where $\boldsymbol{\beta}$ is the matrix of direct effects specified in the general SEM in (1.1). $\mathbf{T}_{\boldsymbol{\nu\nu}}$ is defined only if the infinite series in (1.10) converges to a matrix with finite elements or is stable

(Bentler & Freeman, 1983).

Lemma 1.1. *A square matrix* $\boldsymbol{\beta}$ *is called convergent (Ben-Israel & Greville, 1974) if*

$$\lim_{k\to\infty} \boldsymbol{\beta}^k = \mathbf{0}. \tag{1.11}$$

Proof:

Omitted; see Ben-Israel and Greville (1974).

Theorem 1.1. *A matrix* $\boldsymbol{\beta}$ *is convergent if and only if the absolute value or modulus of the largest eigenvalue is less than one, that is*

$$\rho(\boldsymbol{\beta}) < 1. \tag{1.12}$$

Proof*:*

Let $\boldsymbol{\beta} = \mathbf{UDU}^{-1}$ be a full rank matrix with distinct roots, where $\mathbf{D}$ is a diagonal matrix of eigenvalues. Then, assuming $k > 1$

$$\begin{aligned}\boldsymbol{\beta}^k &= (\mathbf{UDU}^{-1})^k \\ &= (\mathbf{UDU}^{-1}) \cdots (\mathbf{UDU}^{-1}) \quad (k \text{ times}) \\ &= \mathbf{UD}^k\mathbf{U}^{-1}.\end{aligned}$$

Under (1.12), $\rho(\boldsymbol{\beta}) < 1$ so that $\lim_{k\to\infty} \mathbf{D}^k = \mathbf{0}$ and also $\lim_{k\to\infty} \boldsymbol{\beta}^k = \mathbf{0}$. Conversely, (1.11) implies that $\lim_{k\to\infty} \mathbf{D}^k = \mathbf{0}$, therefore requiring that $\rho(\mathbf{D}) < 1$ and hence (1.12). The result also holds more generally when $\mathbf{D} = \mathbf{U}^{-1}\boldsymbol{\beta}\mathbf{U}$ is in Jordan canonical form. ■

To determine to which value $\mathbf{T}_{\boldsymbol{\nu\nu}}$ converges, we can add $\mathbf{I} = \boldsymbol{\beta}^0$ to (1.10) and then premultiply by $\mathbf{I} - \boldsymbol{\beta}$ to get,

$$(\mathbf{I}-\boldsymbol{\beta})(\mathbf{I}+\boldsymbol{\beta}+\boldsymbol{\beta}^2+\cdots+\boldsymbol{\beta}^k)=\mathbf{I}-\boldsymbol{\beta}^{k+1}. \tag{1.13}$$

Since $\boldsymbol{\beta}$ is assumed to be a convergent matrix (i.e., [1.11] holds), in the limit of (1.13),

$$\begin{gathered}\lim_{k\to\infty}[(\mathbf{I}-\boldsymbol{\beta})(\mathbf{I}+\boldsymbol{\beta}+\boldsymbol{\beta}^2+\cdots+\boldsymbol{\beta}^k)]=\lim_{k\to\infty}[\mathbf{I}-\boldsymbol{\beta}^{k+1}] \\ (\mathbf{I}-\boldsymbol{\beta})\lim_{k\to\infty}[\mathbf{I}+\boldsymbol{\beta}+\boldsymbol{\beta}^2+\cdots+\boldsymbol{\beta}^k]=\mathbf{I}.\end{gathered} \tag{1.14}$$

For the product on the left-hand side of (1.14) to equal $\mathbf{I}$, the matrix sum $(\mathbf{I}+\boldsymbol{\beta}+\boldsymbol{\beta}^2+\cdots+\boldsymbol{\beta}^k)$ must converge to $(\mathbf{I}-\boldsymbol{\beta})^{-1}$ as $k\to\infty$ (Bollen, 1987). Subtracting $\mathbf{I}=\boldsymbol{\beta}^0$ from this sum, in a convergent system, the total effects are defined as

$$\begin{aligned}\mathbf{T_{\nu\nu}}&=\sum_{k=1}^{\infty}\boldsymbol{\beta}^k \\ &=(\mathbf{I}-\boldsymbol{\beta})^{-1}-\mathbf{I}.\end{aligned}$$

Given that the $\boldsymbol{\beta}$ is a matrix of direct effects in a SEM, the indirect effects can be calculated as the difference between the total effects and direct effects as

$$g(\boldsymbol{\beta}|\boldsymbol{\nu})=(\mathbf{I}-\boldsymbol{\beta})^{-1}-\mathbf{I}-\boldsymbol{\beta}, \tag{1.15}$$

where $g(\boldsymbol{\beta}|\boldsymbol{\nu})$ denotes the indirect effects of $\boldsymbol{\nu}$ on $\boldsymbol{\nu}$ (Bollen, 1989).

To demonstrate the matrix algebra approach, consider the simple mediation model given in (1.3). First, we need to determine if (1.12) holds, that is, the matrix of direct effects, $\boldsymbol{\beta}$, converges. This condition is easily established by using a basic theorem about lower triangular matrices.

Theorem 1.2. *If* $\boldsymbol{\beta}$ *is an* $n\times n$ *lower triangular matrix, then its eigenvalues,* λ_i*, are the entries on the main diagonal, that is,* $\lambda_i=\boldsymbol{\beta}_{ii}, i=1,\ldots,n.$

Proof:

If $\boldsymbol{\beta}$ is an $n\times n$ lower triangular matrix, where $i=1,\ldots,n$ denotes the rows and $j=$

$1, \ldots, n$ denotes the columns, then $\beta_{ij} = 0$ for all $i > j$. The determinant of $\boldsymbol{\beta}$ is given by

$$\begin{aligned} \det(\boldsymbol{\beta}) &= |\boldsymbol{\beta} - \lambda \mathbf{I}| \\ &= \prod_{i=1}^{n} (\beta_{ii} - \lambda). \end{aligned} \tag{1.16}$$

Since the eigenvalues of $\boldsymbol{\beta}$ are precisely the roots of its characteristic polynomial, the roots of the n^{th} degree polynomial in (1.16) are precisely the eigenvalues. Therefore, the eigenvalues of $\boldsymbol{\beta}$ are $\beta_{ii}, 1 \leq i \leq n$. ■

Using Theorem (1.2), we can easily see that since the direct effects matrix $\boldsymbol{\beta}$ given in (1.3) is lower triangular, all of its eigenvalues are on the main diagonal, which are zero in this case. Since all of the eigenvalues are zero, we can conclude from Theorem (1.2) that $\boldsymbol{\beta}$ is a convergent matrix. Before continuing with the example, we can combine the results of Theorem (1.1) and Theorem (1.2) and establish a corollary about direct effect matrices.

Corollary 1.1. *Any direct effect matrix,* $\boldsymbol{\beta}$*, written in lower triangular form with zeros along the main diagonal, is a convergent matrix.*

From Corollary (1.1), we can directly use Equation (1.15) to calculate the point estimators of indirect effects, if they exist. Continuing with the simple mediation model example, using the formula given in (1.15), the indirect effects of a simple mediation model are given by

$$\begin{aligned} g(\boldsymbol{\beta}|\mathbf{v}) &= (\mathbf{I} - \boldsymbol{\beta})^{-1} - \mathbf{I} - \boldsymbol{\beta} \\ &= \left(\begin{bmatrix} 1 & 0 & 0 \\ -\beta_{M \cdot X} & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} - \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ \beta_{Y \cdot X} & \beta_{Y \cdot M} & 0 \end{bmatrix} \right)^{-1} - \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} - \begin{bmatrix} 0 & 0 & 0 \\ \beta_{M \cdot X} & 0 & 0 \\ \beta_{Y \cdot X} & \beta_{Y \cdot M} & 0 \end{bmatrix} \end{aligned}$$

$$= \begin{bmatrix} 0 & 0 & 0 \\ \beta_{M\cdot X} & 0 & 0 \\ \beta_{Y\cdot X} + \beta_{Y\cdot M}\beta_{M\cdot X} & \beta_{Y\cdot M} & 0 \end{bmatrix} - \begin{bmatrix} 0 & 0 & 0 \\ \beta_{M\cdot X} & 0 & 0 \\ \beta_{Y\cdot X} & \beta_{Y\cdot M} & 0 \end{bmatrix}$$

$$= \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ \beta_{Y\cdot M}\beta_{M\cdot X} & 0 & 0 \end{bmatrix},$$

where $g(\boldsymbol{\beta}|\mathbf{v}) = \beta_{Y\cdot M}\beta_{M\cdot X}$ is the same point estimator of the indirect effect found earlier. The benefit of Bollen's (1987) approach is that the method works for more complicated models and returns equations to calculate point estimators for each indirect effect in a model.

Conveniently, similar matrix approach exists for deriving point estimators for conditional indirect effects. The approach given in Preacher et al. (2007) is based on computing the partial derivatives of the reduced-form equations of endogenous variables to obtain a compact effect matrix, denoted by $\boldsymbol{\beta}^*$. Substituting the compact effect matrix into (1.15) yields the point estimator for conditional indirect effects,

$$g(\boldsymbol{\beta}^*|\mathbf{v}) = (\mathbf{I} - \boldsymbol{\beta}^*)^{-1} - \mathbf{I} - \boldsymbol{\beta}^*. \tag{1.17}$$

To demonstrate this approach, consider the Model 5 in Figure (1.4). This SEM can be expressed as,

$$\begin{bmatrix} X \\ W \\ XW \\ MW \\ M \\ Y \end{bmatrix} = \begin{bmatrix} \alpha_X \\ \alpha_W \\ \alpha_{XW} \\ \alpha_{MW} \\ \alpha_M \\ \alpha_Y \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ \beta_{M\cdot X} & \beta_{M\cdot W} & \beta_{M\cdot XW} & 0 & 0 & 0 \\ \beta_{Y\cdot X} & \beta_{Y\cdot W} & \beta_{Y\cdot XW} & \beta_{Y\cdot MW} & \beta_{Y\cdot M} & 0 \end{bmatrix} \begin{bmatrix} X \\ W \\ XW \\ MW \\ M \\ Y \end{bmatrix} + \begin{bmatrix} \zeta_X \\ \zeta_W \\ \zeta_{XW} \\ \zeta_{MW} \\ \zeta_M \\ \zeta_Y \end{bmatrix}.$$

To calculate the partial derivatives, it is convenient to write out the system of six equations

1. $X = \alpha_X + \zeta_X$
2. $W = \alpha_W + \zeta_W$
3. $XW = \alpha_{XW} + \zeta_{XW}$

4. $MW = \alpha_{MW} + \zeta_{MW}$
5. $M = \alpha_M + \beta_{M \cdot X} X + \beta_{M \cdot W} W + \beta_{M \cdot XW} XW + \zeta_M$
6. $Y = \alpha_Y + \beta_{Y \cdot X} X + \beta_{Y \cdot W} W + \beta_{Y \cdot XW} XW + \beta_{Y \cdot MW} MW + \beta_{Y \cdot M} M + \zeta_Y,$

where X is the exogenous variable, W is the moderator variable, XW is the interaction between the exogenous and moderator variables, M is the mediator variable, MW is the interaction between the mediator and moderator variable, Y is the endogenous variable, and α, β, and ζ represent the intercepts, direct effects, and error terms, respectively for each model.

To obtain the reduced-form equations of endogenous variables, one must first consider the conditional indirect effect(s) of interest. In most cases, interest lies in examining conditional indirect effects of exogenous variables to endogenous variables, rather than of moderator variables to endogenous variables. For the current example, the conditional indirect effect of interest is from the exogenous variable X to the endogenous variable Y. To calculate this effect, using the matrix of direct effects $\boldsymbol{\beta}$ based on the moderated mediation model described above, first drop the rows and columns corresponding to moderator model (since we are not interested in any conditional indirect effect based on this variable) and interaction variable models. This step reduces $\boldsymbol{\beta}$ from a 6×6 matrix to a 3×3 matrix (i.e., since three variables are removed), which is our $\boldsymbol{\beta}^*$. Next, the elements in $\boldsymbol{\beta}^*$ represent the effect (i.e., regression coefficient) of a column on a row, or the partial derivative of a row variable with respect to a column variable.

For the current example,

$$\boldsymbol{\beta}^* = \begin{matrix} \\ X \\ M \\ Y \end{matrix} \begin{matrix} X & M & Y \\ 0 & 0 & 0 \\ \partial M/\partial X & 0 & 0 \\ \partial Y/\partial X & \partial Y/\partial M & 0 \end{matrix},$$

where

$$\frac{\partial M}{\partial X} = \beta_{M \cdot X} + \beta_{M \cdot XW} W$$

$$\frac{\partial Y}{\partial X} = \beta_{Y \cdot X} + \beta_{Y \cdot XW} W$$

$$\frac{\partial Y}{\partial M} = \beta_{Y \cdot M} + \beta_{Y \cdot MW} W.$$

Lastly, using (1.17) we see that

$$g(\boldsymbol{\beta}^* | \boldsymbol{\nu}) = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ (\beta_{M \cdot X} + \beta_{M \cdot XW} W)(\beta_{Y \cdot M} + \beta_{Y \cdot MW} W) & 0 & 0 \end{bmatrix}. \qquad (1.18)$$

Here $g(\boldsymbol{\beta}^* | \boldsymbol{\nu}) = (\beta_{M \cdot X} + \beta_{M \cdot XW} W)(\beta_{Y \cdot M} + \beta_{Y \cdot MW} W)$ is the point estimator of the conditional indirect effect of the exogenous variable X to the endogenous variable Y, via the mediator variable M, conditional upon the moderator variable W.

An alternative approach is to use the chain rule from calculus to calculate specific indirect effects. Using the chain rule, the conditional indirect effect given in (1.18) is obtained by

$$\begin{aligned} g(\beta | \boldsymbol{\nu}) &= \frac{\partial Y}{\partial X} \\ &= \frac{\partial Y}{\partial M} \frac{\partial M}{\partial X} \\ &= (\beta_{Y \cdot M} + \beta_{Y \cdot MW} W)(\beta_{M \cdot X} + \beta_{M \cdot XW} W). \end{aligned}$$

1.4. Estimation of Structural Equation Models

Given that the point estimators for both unconditional and conditional indirect effects are functions of the direct effects **β**, it is necessary to determine how to estimate these parameters. The majority of distribution functions (e.g., normal, Bernoulli, Poisson) used in mediation and moderated mediation models are members of the exponential family. If an endogenous variable Y is a member of the exponential family, then its density function can be expressed in the form

$$f(y|\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}, \tag{1.19}$$

where θ and ϕ are parameters and $a(\cdot), b(\cdot)$, and $c(\cdot)$ are known function. As standard notation, we use Y to denote the random variable and y to denote the realized value of Y. Conveniently, for densities belonging to the exponential family a class of models known as generalized linear models (GLMs) can be used for parameter estimation. GLMs provide a unified modeling framework for handling continuous and discrete random variables that are members of the exponential family (McCullagh & Nelder, 1989). A GLM is characterized by three components, the random component, the systematic component, and the link component.

1.4.1. GLM random component. The random component of a GLM specifies the probability distribution for an endogenous variable Y. The log-likelihood function $\log f(y|\theta,\phi) = \ell(\theta,\phi|y)$, considered as a function of the parameters θ and ϕ with fixed y, of (1.19) is

$$\ell(\theta,\phi|y) = \frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi), \tag{1.20}$$

where log denotes the natural logarithm (Davis, 2002). An important function of a GLM that is used to estimate the mean and variance of Y is known as the score function, denoted U. Mathematically, the score function (Cox & Hinkley, 1974) is defined as the partial derivative of the log-likelihood function with respect to the parameter θ,

$$U = U(\theta) = \frac{\partial}{\partial \theta} l(\theta, \phi | y).$$

The following Proposition gives important identities for the mean and variance of the score function of a GLM.

Proposition 1.1: *Assuming sufficient regularity conditions hold, the mean and variance of the score function, U, are given by*

$$E(U) = 0 \tag{1.21}$$

and

$$Var(U) = -E(U'), \tag{1.22}$$

where U' is the derivative of the score function.

Proof:

See Appendix A.

Using the derived mean and variance of the score function, we can use the log-likelihood given in (1.20) to calculate the score function

$$U = \frac{y - b'(\theta)}{a(\phi)}, \tag{1.23}$$

which can be rewritten in terms of y as

$$y = a(\phi)U + b'(\theta). \tag{1.24}$$

Taking expectations on both sides of (1.24), the expected value of Y is

$$\begin{aligned} E(Y) &= a(\phi)E(U) + b'(\theta) \\ &= b'(\theta) \end{aligned} \tag{1.25}$$

because $E(U) = 0$. To calculate the variance of Y, we first calculate the partial derivative of (1.23) with respect to θ as

$$U' = \frac{-b''(\theta)}{a(\phi)}, \tag{1.26}$$

where $b''(\theta)$ is the second partial derivative of $b(\theta)$. Now, because $E(U^2) = -E(U')$,

$$E\left[\left(\frac{Y - b'(\theta)}{a(\phi)}\right)^2\right] = \frac{b''(\theta)}{a(\phi)}. \tag{1.27}$$

From (1.27), we see that $Var(U) = -E(U') = \frac{b''(\theta)}{a(\phi)}$. Finally, to calculate the variance of Y, using (1.24),

$$\begin{aligned} \text{Var}(Y) &= \text{Var}\big(a(\phi)U + b'(\theta)\big) \\ &= \big(a(\phi)\big)^2 \text{Var}(U) \\ &= a(\phi)b''(\theta). \end{aligned}$$

Let $V(\theta) = b''(\theta)$ be the variance function of a GLM, then the variance of Y can be rewritten as

$$\text{Var}(Y) = a(\phi)V(\theta). \tag{1.28}$$

In order to use the expressions for the mean (1.21) and variance (1.22) of Y, a probability distribution from the exponential family should be written in exponential form given by Equation (1.19). For the purpose of the present study, the normal distribution will be used

to model continuous endogenous variables and the Bernoulli distribution will be used to model discrete binary endogenous variables.

If Y follows a normal distribution with mean $\mu \in \mathcal{R}$ and variance $\sigma^2 > 0$, where $\mathcal{R}$ denotes the set of real numbers, then the density function is given by

$$f(y|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y-\mu)^2}, \tag{1.29}$$

with support $y \in \mathcal{R}$. To show that the normal distribution is a member of the exponential family, we can rewrite (1.29) as

$$f(y|\mu,\sigma^2) = \exp\left\{ \frac{\overset{\theta}{\overbrace{\mu}} y - \overset{b(\theta)}{\overbrace{\mu^2/2}}}{\underbrace{\sigma^2}_{a(\phi)}} - \underbrace{\frac{1}{2}\left[\log(2\pi\sigma^2) + \frac{y}{\sigma^2}\right]}_{c(y,\phi)} \right\}. \tag{1.30}$$

As labeled in (1.30), we see that $\theta = \mu$, $\phi = \sigma^2$, $b(\theta) = \frac{\theta^2}{2}$, $a(\phi) = \phi$. As such, we can see that the mean of Y is

$$E(Y) = b'(\theta)$$

$$= \theta,$$

so $E(Y) = \theta = \mu$. Similarly, for the variance of Y,

$$Var(Y) = a(\phi)b''(\theta)$$

$$= \phi,$$

so $Var(Y) = \phi = \sigma^2$.

If Y follows a Bernoulli distribution with parameter $p \in [0,1]$, then the density function is given by

$$f(y|\,p) = p^y(1-p)^{1-y}, \tag{1.31}$$

where $P(Y=1)=p$ denotes the success probability. Similar to the normal distribution, to show that (1.31) is a member of the exponential family we can rewrite the density as

$$\begin{aligned} f(y|p) &= \exp\{y\log(p)+(1-y)\log(1-p)\} \\ &= \exp\{y\log(p)+\log(1-p)-y\log(1-p)\} \\ &= \exp\left\{y\log\left(\frac{p}{1-p}\right)+\log(1-p)\right\} \end{aligned} \tag{1.32}$$

Looking at the first term, we see that $\theta=\log\left(\frac{p}{1-p}\right)$ and although not as clear, the second term is $b(\theta)=\log(1-p)$. We can rewrite $b(\theta)$ to explicitly have $\log\left(\frac{p}{1-p}\right)$, or the log-odds, in its form as

$$\begin{aligned} \log(1-p) &= -\log\left(\frac{1}{1-p}\right) \\ &= -\log\left(1+\frac{p}{1-p}\right). \end{aligned} \tag{1.33}$$

Substituting (1.33) into (1.32), we see that

$$f(y|p)=\exp\left\{y\underbrace{\log\left(\frac{p}{1-p}\right)}_{\theta}-\underbrace{\log\left(1+\frac{p}{1-p}\right)}_{b(\theta)}\right\}. \tag{1.34}$$

As labeled in (1.34), $\theta=\log\left(\frac{p}{1-p}\right)$, $b(\theta)=\log\left(1+\frac{p}{1-p}\right)$, and implicitly $a(\phi)=1$. By using properties of exponentials and natural logs, $b(\theta)=\log\left(1+e^{\theta}\right)$. The mean of Y is given by

$$\begin{aligned} E(Y) &= b'(\theta) \\ &= \frac{e^{\theta}}{(1+e^{\theta})}. \end{aligned}$$

To simplify, we see that

$$E(Y)=\frac{e^{\theta}}{(1+e^{\theta})}$$

$$= \frac{\frac{p}{1-p}}{1+\frac{p}{1-p}}$$

$$= p.$$

For the variance of Y using the quotient rule,

$$Var(Y) = a(\phi)b''(\theta)$$

$$= \frac{e^{\theta}}{(1+e^{\theta})^2}.$$

Simplifying, we see that

$$Var(Y) = \frac{e^{\theta}}{(1+e^{\theta})^2}$$

$$= \frac{\frac{p}{1-p}}{\left(1+\frac{p}{1-p}\right)^2}$$

$$= \left(\frac{\frac{p}{1-p}}{1+\frac{p}{1-p}}\right)\left(\frac{1}{1+\frac{p}{1-p}}\right)$$

$$= p(1-p).$$

Note, the Bernoulli distribution is equivalent to the Binomial distribution with parameters n, which denotes the number of independent trials, and p, which is also the success probability, but with $n = 1$.

1.4.2. GLM systematic component. After selecting an appropriate distribution function for an endogenous variable Y, a systematic component is specified that relates a covariate vector to Y using a linear predictor η_i as

$$\eta_i = \beta_0 + \sum_{k=1}^{p} x_{ik}\beta_k. \qquad (1.35)$$

Note, p depends on the specified model and each endogenous variable can have its own linear predictor in a SEM. Equivalently, using matrix notation we can rewrite (1.35) as

$$\eta_i = \mathbf{x}_i'\boldsymbol{\beta},$$

where $\mathbf{x}_i' = [1, x_{i1}, \ldots, x_{ip}]$ is a $(p+1)$-dimensional vector of covariates for sample i and $\boldsymbol{\beta} = [\beta_0, \ldots, \beta_p]$ is a $(p+1)$-dimensional vector of unknown regression parameters. The increase in dimensions from p to $p+1$ is a result of augmenting $\mathbf{x}$ and $\boldsymbol{\beta}$ to model the intercept term.

1.4.3. GLM link function. The last component of a GLM relates the linear predictor η_i to the expected value of the random component, or $E(Y_i) = \mu_i$, where in this form, the mean is modeled directly. However, in nonlinear cases (e.g., discrete random variables) the link function is often modeled as a monotone, differentiable function $g(\cdot)$ of μ, where

$$g(\mu_i) = \eta_i.$$

Each probability distribution in the exponential family has one special function of the mean called its natural parameter (Casella & Berger, 2001). For the normal distribution the natural parameter is the mean itself. For the Bernoulli distribution the natural parameter is the logit of the success probability. Link functions that use the natural parameter are called canonical link functions (Cox & Hinkley, 1974), so the canonical link function for the normal distribution is

$$g(\mu_i) = \mu_i,$$

and the canonical link for the Bernoulli distribution is

$$\begin{aligned} g(\mu_i) &= \text{logit}(\mu_i) \\ &= \log\left(\frac{\mu_i}{1-\mu_i}\right). \end{aligned}$$

Here μ_i is equivalent to the probability of success for the ith sample value, that is $P(Y_i = 1) = \mu_i$. The logit link function maps $\mu \in [0,1]$ to the real line and leads to regression parameters with odds-ratio interpretations (Madsen & Thyregod, 2011).

The use of canonical link functions simplifies maximum likelihood estimation routines and leads to inferences for regression parameters based solely on sufficient statistics (Davis, 2002). Briefly, a sufficient statistic for a parameter θ captures all the information about θ contained in the sample (see Casella & Berger, 2001 p. 272 for a more formal definition of sufficient statistic). To show the role that canonical links and sufficient statistics play in GLMs, let $Y_1, \dots, Y_n$ be independent random variables given by (1.19). The log-likelihood function for $y_1, \dots, y_n$ is

$$\ell(\theta_i, \phi | y_i) = \sum_{i=1}^{n} \frac{1}{a(\phi)} y_i \theta_i - \sum_{i=1}^{n} \frac{1}{a(\phi)} b(\theta_i) + \sum_{i=1}^{n} c(y_i, \phi). \qquad (1.36)$$

The use of a canonical link function implies that

$$\theta_i = \eta_i = g(\mu_i) = \mathbf{x}_i' \boldsymbol{\beta},$$

and since the canonical link function is a one-to-one function of μ_i, then the inverse relation $g^{-1}(\mu_i) = \eta_i$ holds. The first term of the (1.36) becomes

$$\sum_{i=1}^{n} \frac{1}{a(\phi)} y_i \mathbf{x}_i' \boldsymbol{\beta}.$$

Let $\mathbf{X} = [\mathbf{1}, \mathbf{x}_1, \dots, \mathbf{x}_p]$ denote the $n \times (p+1)$ matrix of covariates for all n observations, and let $\mathbf{y} = [y_1, \dots, y_n]'$ denote the $n \times 1$ vector of responses for all n observations. Then the $(p+1) \times 1$ vector $\mathbf{X}'\mathbf{y}$ with jth component $\sum_{i=1}^{n} x_{ij} y_i$ is a sufficient statistic for $\boldsymbol{\beta}$, and $\eta = \theta$ is called the canonical link function (McCullagh & Nelder, 1989).

1.4.4. Simplifying the joint SEM model. As mentioned earlier in this chapter, in SEMs that assume $Cov(\zeta_i, \zeta_j) = 0$ for $i, j = 1,2, \dots, p$ where $i \neq j$, then estimating the joint regression model, $f(\boldsymbol{\nu}|\boldsymbol{\theta})$, is equivalent to estimating simpler, often univariate regression models. As such, when distributions belonging to the exponential family are used, these univariate regression models fit into the GLM framework. Before describing methods for estimating SEMs, it is first necessary to show that under certain conditions, the joint SEM can be estimated as a series of independent regression models. Consider an SEM model defined by$f(\boldsymbol{\nu}|\boldsymbol{\theta})$, where $\boldsymbol{\nu}$ is a vector of observed variables and $\boldsymbol{\theta}$ is a vector of unknown parameters. Let $\boldsymbol{\nu}$ denote dependent variables, $\boldsymbol{\eta}$ denote covariates, and $\boldsymbol{\theta}$ denote model parameters:

$\boldsymbol{\nu}_Y$: vector of dependent variables in the endogenous variable model

$\boldsymbol{\nu}_M$: vector of dependent variables in the mediator variable model

$\boldsymbol{\nu}_X$: vector of dependent variables in the exogenous variable model

$\boldsymbol{\eta}_Y$: vector of covariates in the endogenous variable model

$\boldsymbol{\eta}_M$: vector of covariates in the mediator variable model

$\boldsymbol{\eta}_X$: vector of covariates in the exogenous variable model

$\boldsymbol{\theta}_Y$: vector of parameters in the endogenous variable model

$\boldsymbol{\theta}_M$: vector of parameters in the mediator variable model

$\boldsymbol{\theta}_X$: vector of parameters in the exogenous variable model

Using this new notation, if all error terms are assumed to be independent, we can reparametrize the joint SEM model as

$$f(\boldsymbol{\nu}|\boldsymbol{\theta}) = f(\boldsymbol{\nu}_Y, \boldsymbol{\nu}_M, \boldsymbol{\nu}_X|\boldsymbol{\eta}_Y, \boldsymbol{\eta}_M, \boldsymbol{\eta}_X, \boldsymbol{\theta}_Y, \boldsymbol{\theta}_M, \boldsymbol{\theta}_X)$$

$$= f(\boldsymbol{\nu}_Y|\boldsymbol{\nu}_M, \boldsymbol{\nu}_X, \boldsymbol{\eta}_Y, \boldsymbol{\eta}_M, \boldsymbol{\eta}_X, \boldsymbol{\theta}_Y, \boldsymbol{\theta}_M, \boldsymbol{\theta}_X)$$
$$\times f(\boldsymbol{\nu}_M|\boldsymbol{\nu}_X, \boldsymbol{\eta}_Y, \boldsymbol{\eta}_M, \boldsymbol{\eta}_X, \boldsymbol{\theta}_Y, \boldsymbol{\theta}_M, \boldsymbol{\theta}_X)$$
$$\times f(\boldsymbol{\nu}_X|\boldsymbol{\eta}_Y, \boldsymbol{\eta}_M, \boldsymbol{\eta}_X, \boldsymbol{\theta}_Y, \boldsymbol{\theta}_M, \boldsymbol{\theta}_X)$$
$$= f(\boldsymbol{\nu}_Y|\boldsymbol{\eta}_Y, \boldsymbol{\theta}_Y) f(\boldsymbol{\nu}_M|\boldsymbol{\eta}_M, \boldsymbol{\theta}_M) f(\boldsymbol{\nu}_X|\boldsymbol{\eta}_X, \boldsymbol{\theta}_X).$$

The term $f(\boldsymbol{\nu}_Y|\boldsymbol{\nu}_M, \boldsymbol{\nu}_X, \boldsymbol{\eta}_Y, \boldsymbol{\eta}_M, \boldsymbol{\eta}_X, \boldsymbol{\theta}_Y, \boldsymbol{\theta}_M, \boldsymbol{\theta}_X)$ simplifies to $f(\boldsymbol{\nu}_Y|\boldsymbol{\eta}_Y, \boldsymbol{\theta}_Y)$ since $\boldsymbol{\nu}_M, \boldsymbol{\nu}_X, \boldsymbol{\eta}_M, \boldsymbol{\eta}_X, \boldsymbol{\theta}_M$, and $\boldsymbol{\theta}_X$ are independent of the endogenous variable model. Likewise, the terms $f(\boldsymbol{\nu}_M|\boldsymbol{\nu}_X, \boldsymbol{\eta}_Y, \boldsymbol{\eta}_M, \boldsymbol{\eta}_X, \boldsymbol{\theta}_Y, \boldsymbol{\theta}_M, \boldsymbol{\theta}_X)$ simplifies to $f(\boldsymbol{\nu}_M|\boldsymbol{\eta}_M, \boldsymbol{\theta}_M)$ and $f(\boldsymbol{\nu}_X|\boldsymbol{\eta}_Y, \boldsymbol{\eta}_M, \boldsymbol{\eta}_X, \boldsymbol{\theta}_Y, \boldsymbol{\theta}_M, \boldsymbol{\theta}_X)$ simplifies to $f(\boldsymbol{\nu}_X|\boldsymbol{\eta}_X, \boldsymbol{\theta}_X)$ because of the independent regression model specification. Therefore, $f(\boldsymbol{\nu}|\boldsymbol{\theta})$ can be factored as

$$f(\boldsymbol{\nu}|\boldsymbol{\theta}) = \underbrace{f(\boldsymbol{\nu}_Y|\boldsymbol{\eta}_Y, \boldsymbol{\theta}_Y)}_{\text{Model of } Y \in \boldsymbol{\nu}_Y} \underbrace{f(\boldsymbol{\nu}_M|\boldsymbol{\eta}_M, \boldsymbol{\theta}_M)}_{\text{Model of } M \in \boldsymbol{\nu}_M} \underbrace{f(\boldsymbol{\nu}_X|\boldsymbol{\eta}_X, \boldsymbol{\theta}_X)}_{\text{Model of } X \in \boldsymbol{\nu}_X}.$$

When $\boldsymbol{\nu}_i$ is a scalar for $i = X, M$, or Y, $f(\cdot)$ is a univariate regression model, otherwise, $f(\cdot)$ is a multivariate regression model. Note, all models discussed previously in this chapter are univariate regression models for both the endogenous variable model and mediator variable model.

1.4.5. Maximum likelihood estimation. A common method of estimation for SEMs is maximum likelihood (ML). As the name implies, ML estimators are based on the likelihood function. We will now give a more formal definition of the likelihood function.

Definiton 1.1. *Let* $Y_1, \ldots, Y_n$ *be independent random variables with a density functions* $f(y_i|\boldsymbol{\theta})$ *that depend on a vector-valued parameter* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, *where* $\boldsymbol{\Theta}$ *denotes the parameter space of* $\boldsymbol{\theta}$. *The joint density function of n independent observations* $\mathbf{y} = [y_1, \ldots, y_n]'$ *is*

$$f(\mathbf{y}|\boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i|\boldsymbol{\theta})$$

$$= L(\boldsymbol{\theta}|\mathbf{y}).$$

The expression, viewed as a function of the unknown parameters $\boldsymbol{\theta}$ *given the data* $Y_1 = y_1, \ldots, Y_n = y_n$ *are held fixed, is called the likelihood function.*

The method of ML estimates $\boldsymbol{\theta}$ by finding a value $\widehat{\boldsymbol{\theta}}$ that maximizes the likelihood, with the data held fixed (Aldrich, 1997). The maximum likelihood estimator (MLE) of $\boldsymbol{\theta}$ is defined as,

$$\{\widehat{\boldsymbol{\theta}}_{MLE}\} \subseteq \left\{\arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} L(\boldsymbol{\theta}|\mathbf{y})\right\}.$$

Since finding MLEs involves calculating derivatives, it is often easier to work with the logarithm of the likelihood function, called the log-likelihood. The logarithm of the likelihood is a strictly monotonically increasing function, therefore, the MLE estimate is invariant whether we maximize the likelihood or log-likelihood (Lehmann & Casella, 1998).

A mathematical convenience of GLMs is that for all models, ML estimates of the parameter vector $\boldsymbol{\beta}$ can be obtained using the same iteratively reweighted least squares (IRLS) algorithm (McCullagh & Nelder, 1989). The MLEs of the parameter vector $\boldsymbol{\beta}$ are the solutions of the score equations

$$\frac{\partial \ell}{\partial \beta_j} = 0 \qquad (1.37)$$

for $j = 0, \ldots, p$ where $\ell = \sum_{i=1}^{n} \ell_i(\theta_i, \phi|y_i)$. Using the chain rule, the derivative in (1.37) can expressed as

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial \ell_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j} \tag{1.38}$$

for $j = 0, \dots, p$. For general link functions, (1.38) can be calculated by first considering each derivative term separately. For the first derivative,

$$\frac{\partial \ell_i}{\partial \theta_i} = \frac{\partial \ell_i}{\partial \theta_i}\left(\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right)$$

$$= \frac{y_i - b'(\theta_i)}{a(\phi)}$$

$$= \frac{y_i - \mu_i}{a(\phi)}.$$

The second derivative term can be obtained by recognizing that $\frac{\partial \theta_i}{\partial \mu_i} = \left(\frac{\partial \mu_i}{\partial \theta_i}\right)^{-1}$,

$$\frac{\partial \theta_i}{\partial \mu_i} = \left(\frac{\partial \mu_i}{\partial \theta_i}\right)^{-1}$$

$$= \left(\frac{\partial b'(\theta_i)}{\partial \theta_i}\right)^{-1}$$

$$= \frac{1}{b''(\theta_i)}$$

$$= \frac{a(\phi)}{Var(Y_i)}.$$

The third derivative, $\frac{\partial \mu_i}{\partial \eta_i}$, we will keep as is for now until we consider canonical link functions. Lastly, the fourth derivative term can be expressed as

$$\frac{\partial \eta_i}{\partial \beta_j} = \frac{\partial}{\partial \beta_j} \sum_{i=1}^{n} x_{ij}\beta_j$$

$$= x_{ij}.$$

Combining all the partial derivatives, we see that the maximum likelihood estimators of $\boldsymbol{\beta}$ are the solutions to

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{n} \frac{y_i - \mu_i}{a(\phi)} \frac{a(\phi)}{Var(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} x_{ij} = 0$$

$$= \sum_{i=1}^{n} \frac{(y_i - \mu_i) x_{ij}}{Var(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} = 0$$

for $j = 0, \dots, p$.

The IRLS algorithm used in the GLM framework requires the variance-covariance matrix of the score equations. Recall, the information is the variance of the score function and the univariate case was considered above. Now, however, because the maximum likelihood estimators are based on solving a system of $p + 1$ score equations, we must consider the variance-covariance matrix of this set of score equations. Specifically, the variance-covariance matrix of a set of $p + 1$ score equations is referred to as the information matrix, denoted by $\mathbf{I}(\cdot)$. Let $\mathbf{U} = \frac{\partial \ell}{\partial \boldsymbol{\beta}}$ be the $(p + 1) \times 1$ gradient vector defined as

$$\mathbf{U} = \frac{\partial \ell}{\partial \boldsymbol{\beta}}$$

$$= \begin{bmatrix} \frac{\partial \ell}{\partial \beta_0} \\ \vdots \\ \frac{\partial \ell}{\partial \beta_p} \end{bmatrix},$$

then the information $\mathbf{I}(\boldsymbol{\beta})$ is given by

$$\mathbf{I}(\boldsymbol{\beta}) = E\left[\big(\mathbf{U} - E(\mathbf{U})\big)\big(\mathbf{U} - E(\mathbf{U})\big)'\right] = E[\mathbf{U}\mathbf{U}'],$$

since $E(\mathbf{U}) = \mathbf{0}$ (Lehmann, 1999) with (j, k)th element (Davis, 2002)

$$\mathbf{I}_{jk}(\boldsymbol{\beta}) = E\left(\frac{\partial \ell}{\partial \beta_j}\frac{\partial \ell}{\partial \beta_k}\right)$$

$$= E\left(\sum_{i=1}^{n}\frac{(y_i-\mu_i)x_{ij}}{Var(Y_i)}\frac{\partial \mu_i}{\partial \eta_i}\sum_{i=1}^{n}\frac{(y_i-\mu_i)x_{ik}}{Var(Y_i)}\frac{\partial \mu_i}{\partial \eta_i}\right)$$

$$= E\left(\sum_{i=1}^{n}\frac{(y_i-\mu_i)^2 x_{ij}x_{ik}}{Var(Y_i)^2}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2\right)$$

$$= \sum_{i=1}^{n}\frac{x_{ij}x_{ik}}{Var(Y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2.$$

Technically, $\mathbf{I}(\boldsymbol{\beta})$ is the expected information. By properties of the score function under regularity conditions, $\mathbf{I}_{jk}(\boldsymbol{\beta})$ is equivalently

$$E\left(\frac{\partial^2 \ell}{\partial \beta_j \partial \beta_k}\right) = -\sum_{i=1}^{n}\frac{x_{ij}x_{ik}}{Var(Y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2.$$

Using the IRLS algorithm, the kth approximation to $\boldsymbol{\beta}$ can be estimated by an iterative method known as the Fisher Scoring Algorithm,

$$\mathbf{b}^{(k)} = \mathbf{b}^{(k-1)} + \left[\mathbf{I}(\boldsymbol{\beta})^{(k-1)}\right]^{-1}\mathbf{U}^{(k-1)}, \tag{1.39}$$

where $\mathbf{b}^{(k-1)}$ is the estimate of $\boldsymbol{\beta}$ at the $(k-1)$th iteration, $\mathbf{U}^{(k-1)}$ is the gradient vector of the log-likelihood evaluated at $\mathbf{b}^{(k-1)}$, and $\mathbf{I}(\boldsymbol{\beta})^{(k-1)}$ is the information matrix evaluated at $\mathbf{b}^{(k-1)}$.

The logic behind the iterative scheme in (1.39) is based on a Taylor expansion of the score function. In particular, suppose we have a vector of starting values $\boldsymbol{\beta}_0$, then a first-order Taylor expansion of $\mathbf{U}(\boldsymbol{\beta})$ about $\boldsymbol{\beta}_0$ is

$$\mathbf{U}(\boldsymbol{\beta}) = \mathbf{U}(\boldsymbol{\beta}_0) + \boldsymbol{\mathcal{I}}(\boldsymbol{\beta}_0)(\boldsymbol{\beta} - \boldsymbol{\beta}_0), \tag{1.40}$$

where

$$\mathcal{I}(\boldsymbol{\beta}_0) = -\frac{\partial \ell}{\partial \boldsymbol{\beta}_0 \boldsymbol{\beta}_0'}$$

is the observed information at $\boldsymbol{\beta}_0$. Let $\widehat{\boldsymbol{\beta}}$ be the MLE of $\boldsymbol{\beta}$, then substituting $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}$ into (1.40) and observing that $\mathbf{U}(\widehat{\boldsymbol{\beta}}) = \mathbf{0}$ gives

$$\mathcal{I}(\boldsymbol{\beta}_0)(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = \mathbf{U}(\boldsymbol{\beta}_0)$$

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0 + \mathcal{I}(\boldsymbol{\beta}_0)^{-1}\mathbf{U}(\boldsymbol{\beta}_0).$$

By replacing $\mathcal{I}(\boldsymbol{\beta}_0)^{-1}$ with its expectation, we get the iterative routine in (1.40).

Going back to (1.39), we can premultiply both sides by $\mathbf{I}(\boldsymbol{\beta})^{(m-1)}$ to get

$$\mathbf{I}(\boldsymbol{\beta})^{(k-1)}\mathbf{b}^{(k)} = \mathbf{I}(\boldsymbol{\beta})^{(k-1)}\mathbf{b}^{(k-1)} + \mathbf{U}^{(k-1)}. \qquad (1.41)$$

Let $\mathbf{W}$ be an $n \times n$ diagonal matrix with elements

$$w_{ii} = \frac{1}{Var(Y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2,$$

then $\mathbf{I}(\boldsymbol{\beta}) = \mathbf{X}'\mathbf{W}\mathbf{X}$ and the iterative scheme in (1.41) becomes

$$\mathbf{X}'\mathbf{W}\mathbf{X}\mathbf{b}^{(k)} = \mathbf{X}'\mathbf{W}\mathbf{X}\mathbf{b}^{(k-1)} + \mathbf{U}^{(k-1)}.$$

The jth row of the $(p+1) \times n$ matrix $\mathbf{X}'\mathbf{W}$ is

$$[x_{1j}w_{11}, \ldots, x_{nj}w_{nn}] = \left[\frac{x_{1j}}{Var(Y_1)}\left(\frac{\partial \mu_1}{\partial \eta_1}\right)^2, \ldots, \frac{x_{nj}}{Var(Y_n)}\left(\frac{\partial \mu_n}{\partial \eta_n}\right)^2\right]$$

and the jth component of $\mathbf{U}$ is

$$U_j = \sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{ij}}{Var(Y_i)}\frac{\partial \mu_i}{\partial \eta_i}.$$

Define a new $n \times 1$ vector, denoted by $\mathbf{z}$, of 'adjusted' endogenous variables as

$$\mathbf{z} = \begin{bmatrix} \hat{\eta}_1 + (y_1 - \hat{\mu}_1)\dfrac{\partial \eta_1}{\partial \mu_1} \\ \vdots \\ \hat{\eta}_n + (y_n - \hat{\mu}_n)\dfrac{\partial \eta_n}{\partial \mu_n} \end{bmatrix},$$

where $\hat{\eta}_i$ is the linear predictor in (1.35) and $\hat{\eta}_i$, $\hat{\mu}_i$ and $\frac{\partial \mu_i}{\partial \eta_i}$ are all evaluated at the $\mathbf{b}^{(k-1)}$.

The final iterative scheme becomes

$$\mathbf{X}'\mathbf{W}\mathbf{X}\mathbf{b}^{(k)} = \mathbf{X}'\mathbf{W}\mathbf{z}$$

and assuming that $\mathbf{X}'\mathbf{W}\mathbf{X}$ has rank $p+1$,

$$\mathbf{b}^{(k)} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{z}.$$

When the canonical link function is used, $\eta_i = \theta_i = \mathbf{x}'\boldsymbol{\beta}$, then

$$\begin{aligned} \frac{\partial \mu_i}{\partial \eta_i} &= \frac{\partial \mu_i}{\partial \theta_i} \\ &= \frac{\partial \mathrm{b}'(\theta_i)}{\partial \theta_i} \\ &= \mathrm{b}''(\theta_i) \\ &= \frac{Var(Y_i)}{a(\phi)}. \end{aligned}$$

Therefore, the MLEs are the solutions to the simplified system of score equations

$$\begin{aligned} \frac{\partial \ell}{\partial \beta_j} &= \sum_{i=1}^{n} \frac{y_i - \mu_i}{a(\phi)} \frac{a(\phi)}{Var(Y_i)} \frac{Var(Y_i)}{a(\phi)} x_{ij} = 0 \\ &= \sum_{i=1}^{n} \frac{y_i - \mu_i}{a(\phi)} x_{ij} = 0. \end{aligned}$$

Since the (j,k)th component of the observed information matrix is

$$\frac{\partial^2 \ell}{\partial \beta_j \partial \beta_k} = -\sum_{i=1}^{n} \frac{x_{ij}}{a(\phi)} \left(\frac{\partial \mu_i}{\partial \beta_k}\right), \tag{1.42}$$

this implies that $\frac{\partial^2 \ell}{\partial \beta_j \partial \beta_k} = E\left(\frac{\partial^2 \ell}{\partial \beta_j \partial \beta_k}\right)$ because (1.42) does not depend on $\{Y_i\}$. The asymptotic variances of the MLEs are obtained as the negative inverse of the expected information matrix $\mathbf{I}(\boldsymbol{\beta})^{-1}$ (Lehmann & Casella, 1998).

We can use the ML routines described above to estimate the model parameters for the mediation and endogenous variable models. Specifically, for a random sample of n observations, the likelihood of $f(\boldsymbol{\nu}|\boldsymbol{\theta})$ is

$$L(\boldsymbol{\theta}|\boldsymbol{\nu}) = \prod_{i=1}^{n} f(\boldsymbol{\nu}_{Y_i}|\boldsymbol{\eta}_{Y_i}, \boldsymbol{\theta}_Y) f(\boldsymbol{\nu}_{M_i}|\boldsymbol{\eta}_{M_i}, \boldsymbol{\theta}_M) f(\boldsymbol{\nu}_{X_i}|\boldsymbol{\eta}_{X_i}, \boldsymbol{\theta}_X)$$

and the log-likelihood is

$$\ell(\boldsymbol{\theta}|\boldsymbol{\nu}) = \sum_{i=1}^{n} \log f(\boldsymbol{\nu}_{Y_i}|\boldsymbol{\eta}_{Y_i}, \boldsymbol{\theta}_Y) + \sum_{i=1}^{n} \log f(\boldsymbol{\nu}_{M_i}|\boldsymbol{\eta}_{M_i}, \boldsymbol{\theta}_M) + \sum_{i=1}^{n} \log f(\boldsymbol{\nu}_{X_i}|\boldsymbol{\eta}_{X_i}, \boldsymbol{\theta}_X).$$

Here, $f(\boldsymbol{\nu}_{Y_i}|\boldsymbol{\eta}_{Y_i}, \boldsymbol{\theta}_Y)$ and $f(\boldsymbol{\nu}_{M_i}|\boldsymbol{\eta}_{M_i}, \boldsymbol{\theta}_M)$ can be parametrized using a distribution belonging to the exponential family (e.g., normal, Bernoulli) since the models are univariate. Although the parameters of $f(\boldsymbol{\nu}_{X_i}|\boldsymbol{\eta}_{X_i}, \boldsymbol{\theta}_X)$ may be of interest in some applications, they are irrelevant for calculating indirect effects and be safely ignored. Therefore, we can consider the log-likelihood in as proportional to

$$\ell(\boldsymbol{\theta}|\boldsymbol{\nu}) \propto \sum_{i=1}^{n} \log f(\boldsymbol{\nu}_{Y_i}|\boldsymbol{\eta}_{Y_i}, \boldsymbol{\theta}_Y) + \sum_{i=1}^{n} \log f(\boldsymbol{\nu}_{M_i}|\boldsymbol{\eta}_{M_i}, \boldsymbol{\theta}_M).$$

To calculate the regression parameters for each model, we consider the partial derivatives $\frac{\partial}{\partial \boldsymbol{\theta}_Y} \ell(\boldsymbol{\theta}|\boldsymbol{\nu})$ and $\frac{\partial}{\partial \boldsymbol{\theta}_M} \ell(\boldsymbol{\theta}|\boldsymbol{\nu})$,

$$\frac{\partial}{\partial \boldsymbol{\theta}_Y} \ell(\boldsymbol{\theta}|\boldsymbol{\nu}) = \frac{\partial}{\partial \boldsymbol{\theta}_Y} \left[\sum_{i=1}^{n} \log f(\boldsymbol{\nu}_{Y_i}|\boldsymbol{\eta}_{Y_i}, \boldsymbol{\theta}_Y) + \sum_{i=1}^{n} \log f(\boldsymbol{\nu}_{M_i}|\boldsymbol{\eta}_{M_i}, \boldsymbol{\theta}_M) \right]$$

$$= \sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}_Y} \log f\left(\boldsymbol{\nu}_{Y_i} | \boldsymbol{\eta}_{Y_i}, \boldsymbol{\theta}_Y\right)$$

and

$$\frac{\partial}{\partial \boldsymbol{\theta}_M} \ell(\boldsymbol{\theta}|\boldsymbol{\nu}) = \frac{\partial}{\partial \boldsymbol{\theta}_M} \left[\sum_{i=1}^{n} \log f\left(\boldsymbol{\nu}_{Y_i} | \boldsymbol{\eta}_{Y_i}, \boldsymbol{\theta}_Y\right) + \sum_{i=1}^{n} \log f\left(\boldsymbol{\nu}_{M_i} | \boldsymbol{\eta}_{M_i}, \boldsymbol{\theta}_M\right) \right]$$

$$= \sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}_M} \log f\left(\boldsymbol{\nu}_{M_i} | \boldsymbol{\eta}_{M_i}, \boldsymbol{\theta}_M\right).$$

These calculations reinforce the notion that assuming independent error terms, the regression parameters for the mediator and endogenous variable models can be estimated separately.

1.5. Confidence Interval Estimation for Indirect-Type Effects

1.5.1. Multivariate delta method. Confidence intervals are the most widely used method to test the significance of an indirect effect, that is testing the null hypothesis

$$H_0: g(\boldsymbol{\beta}|\boldsymbol{\nu}) = 0.$$

A popular approximation method for obtaining confidence intervals for indirect effects in SEMs is based on the (first-order and second-order) multivariate delta method. The multivariate delta method is a general technique for deriving the asymptotic distribution of a differentiable vector function of a multivariate normally distributed vector (Sobel, 1986).

Theorem 1.3. *Let g be a function that is differentiable in a neighborhood of a k-dimensional vector* $\boldsymbol{\theta}$. *Let* $\widehat{\boldsymbol{\theta}}_n$ be a sequence of random vectors that satisfies

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\right) \xrightarrow{\mathcal{D}} N\left(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\theta})\right),$$

where $\xrightarrow{\mathcal{D}}$ *denotes convergence in distribution, then*

$$\sqrt{n}\left(g(\widehat{\boldsymbol{\theta}}_n) - g(\boldsymbol{\theta})\right) \xrightarrow{\mathcal{D}} N\left(\mathbf{0}, \left(\frac{\partial g}{\partial \boldsymbol{\theta}}\right)' \boldsymbol{\Sigma}(\boldsymbol{\theta}) \left(\frac{\partial g}{\partial \boldsymbol{\theta}}\right)\right),$$

provided that the quadratic form $\left(\frac{\partial g}{\partial \boldsymbol{\theta}}\right)' \boldsymbol{\Sigma}(\boldsymbol{\theta}) \left(\frac{\partial g}{\partial \boldsymbol{\theta}}\right)$ *does not vanish, where*

$$\left(\frac{\partial g}{\partial \boldsymbol{\theta}}\right)' = \left[\frac{\partial g}{\partial \theta_1}, \ldots, \frac{\partial g}{\partial \theta_k}\right]$$

is the gradient vector of g.

Proof:

A formal proof necessitates introducing technical convergence concepts of multivariate normal random variables, therefore, only a general outline of the proof will be discussed. For a detailed proof, see Lehman and Casella (1998). Suppose g has a first-order Taylor series expansion of $\widehat{\boldsymbol{\theta}}_n$ around $\boldsymbol{\theta}$ given by

$$\begin{aligned} g(\widehat{\boldsymbol{\theta}}_n) &\cong g(\boldsymbol{\theta}) + (\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})' \frac{\partial g}{\partial \boldsymbol{\theta}} \\ g(\widehat{\boldsymbol{\theta}}_n) - g(\boldsymbol{\theta}) &\cong (\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})' \frac{\partial g}{\partial \boldsymbol{\theta}} \end{aligned} \qquad (1.42)$$

The form given in (1.42) is a linear combination of a multivariate normal random vector. An important property of the multivariate normal distribution is that the distribution of a linear combination of a multivariate normal random vector also has a multivariate normal distribution (Johnson & Wichern, 2002; Rao, 1976). Therefore, the distribution of $g(\widehat{\boldsymbol{\theta}}_n)$ is multivariate normal with mean

$$\begin{aligned} E\left[g(\widehat{\boldsymbol{\theta}}_n) - g(\boldsymbol{\theta})\right] &= E\left[(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})' \frac{\partial g}{\partial \boldsymbol{\theta}}\right] \\ &= \mathbf{0} \end{aligned}$$

since $g(\boldsymbol{\theta})$ is constant and $E(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) = \mathbf{0}$ and variance-covariance matrix

$$Var\left[g\left(\widehat{\boldsymbol{\theta}}_n\right) - g(\boldsymbol{\theta})\right] = Var\left[\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\right)' \frac{\partial g}{\partial \boldsymbol{\theta}}\right]$$

$$= \left[\left(\frac{\partial g}{\partial \boldsymbol{\theta}}\right)' Var\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\right)' \left(\frac{\partial g}{\partial \boldsymbol{\theta}}\right)\right]$$

$$= n^{-1}\left[\left(\frac{\partial g}{\partial \boldsymbol{\theta}}\right)' \boldsymbol{\Sigma}(\boldsymbol{\theta}) \left(\frac{\partial g}{\partial \boldsymbol{\theta}}\right)\right],$$

where $Var\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\right)' = n^{-1}\boldsymbol{\Sigma}(\boldsymbol{\theta})$. Therefore, the distribution of $g\left(\widehat{\boldsymbol{\theta}}_n\right) - g(\boldsymbol{\theta})$ is asymptotically normal with mean given by the zero mean vector and variance-covariance matrix given by the quadratic form $n^{-1}\left[\left(\frac{\partial g}{\partial \boldsymbol{\theta}}\right)' \boldsymbol{\Sigma}(\boldsymbol{\theta}) \left(\frac{\partial g}{\partial \boldsymbol{\theta}}\right)\right]$. ■

A formal theorem for the second-order multivariate delta method will not be given, however, we will consider its variance approximation. Although the second-order delta method is direct extension of the first-order method, it requires introducing properties about the trace of a matrix and the distribution of quadratic forms involving multivariate normal random vectors.

Lemma 1.1. *The trace of an* $n \times n$ *square matrix* $\mathbf{A}$, *denote by* $\text{tr}(\mathbf{A})$, *is the sum of elements along the main diagonal of* $\mathbf{A}$,

$$\text{tr}(\mathbf{A}) = \sum_{i=1}^{n} a_{ii}.$$

Assuming all matrices conform to addition and multiplication operations, the trace operator is characterized by the following properties:

i. $\text{tr}(\mathbf{A} + \mathbf{B}) = \text{tr}(\mathbf{A}) + \text{tr}(\mathbf{B})$

ii. $\text{tr}(c\mathbf{A}) = c\text{tr}(\mathbf{A})$

iii. $\text{tr}(\mathbf{ABC}) = \text{tr}(\mathbf{BCA}) = \text{tr}(\mathbf{CAB})$

where c is a constant.

Proof:

Omitted; see Searle (1982) and Harville (1997).

Using the properties stated in Lemma 1.1, we can state an important theorem about quadratic forms involving normal random vectors.

Theorem 1.4. *Let* $\boldsymbol{\epsilon}$ *be a random multivariate normally distributed vector with mean* $\boldsymbol{\mu}$ *and covariance matrix* $\boldsymbol{\Sigma}$*. Then for symmetric, positive definite matrix* $\mathbf{A}$*, the mean of the quadratic form* $\boldsymbol{\epsilon}'\mathbf{A}\boldsymbol{\epsilon}$ *is*

$$E(\boldsymbol{\epsilon}'\mathbf{A}\boldsymbol{\epsilon}) = \boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu} + \text{tr}(\mathbf{A}\boldsymbol{\Sigma}) \tag{1.43}$$

and variance is

$$Var(\boldsymbol{\epsilon}'\mathbf{A}\boldsymbol{\epsilon}) = 2\text{tr}[(\mathbf{A}\boldsymbol{\Sigma})^2] + 4\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}\boldsymbol{\mu}. \tag{1.44}$$

Proof:

The mean of the quadratic form $\boldsymbol{\epsilon}'\mathbf{A}\boldsymbol{\epsilon}$ can be derived by using properties of the trace operator in Lemma (1.1) and of moments of multivariate normal random vectors,

$$\begin{aligned} E(\boldsymbol{\epsilon}'\mathbf{A}\boldsymbol{\epsilon}) &= E[\text{tr}(\boldsymbol{\epsilon}'\mathbf{A}\boldsymbol{\epsilon})] \\ &= \text{tr}[\mathbf{A}E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}')] \\ &= \text{tr}[\mathbf{A}(\text{Cov}(\boldsymbol{\epsilon}) + \boldsymbol{\mu}\boldsymbol{\mu}')] \\ &= \text{tr}(\mathbf{A}\boldsymbol{\Sigma}) + \boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}. \end{aligned}$$

Here, $E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}') = (\text{Cov}(\boldsymbol{\epsilon}) + \boldsymbol{\mu}\boldsymbol{\mu}')$ because $\text{Cov}(\boldsymbol{\epsilon}) = E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}') - \boldsymbol{\mu}\boldsymbol{\mu}'$. Using the result of the mean of the quadratic form, we can start with the definition of a variance to derive the variance of $\boldsymbol{\epsilon}'\mathbf{A}\boldsymbol{\epsilon}$,

$$\begin{aligned} Var(\boldsymbol{\epsilon}'\mathbf{A}\boldsymbol{\epsilon}) &= \boldsymbol{E}\left[\left(\boldsymbol{\epsilon}'\mathbf{A}\boldsymbol{\epsilon} - \mathbf{E}(\boldsymbol{\epsilon}'\mathbf{A}\boldsymbol{\epsilon})\right)\left(\boldsymbol{\epsilon}'\mathbf{A}\boldsymbol{\epsilon} - \mathbf{E}(\boldsymbol{\epsilon}'\mathbf{A}\boldsymbol{\epsilon})\right)'\right] \\ &= \boldsymbol{E}\left[\left(\boldsymbol{\epsilon}'\mathbf{A}\boldsymbol{\epsilon} - \mathbf{E}(\boldsymbol{\epsilon}'\mathbf{A}\boldsymbol{\epsilon})\right)\left(\boldsymbol{\epsilon}'\mathbf{A}\boldsymbol{\epsilon} - \mathbf{E}(\boldsymbol{\epsilon}'\mathbf{A}\boldsymbol{\epsilon})\right)\right] \\ &= \boldsymbol{E}\begin{bmatrix}\boldsymbol{\epsilon}'\mathbf{A}\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\mathbf{A}\boldsymbol{\epsilon} - \boldsymbol{\epsilon}'\mathbf{A}\boldsymbol{\epsilon}\mathbf{E}(\boldsymbol{\epsilon}'\mathbf{A}\boldsymbol{\epsilon}) - \boldsymbol{\epsilon}'\mathbf{A}\boldsymbol{\epsilon}\mathbf{E}(\boldsymbol{\epsilon}'\mathbf{A}\boldsymbol{\epsilon}) \\ +\left(\mathbf{E}(\boldsymbol{\epsilon}'\mathbf{A}\boldsymbol{\epsilon})\right)^2\end{bmatrix} \\ &= \boldsymbol{E}(\boldsymbol{\epsilon}'\mathbf{A}\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\mathbf{A}\boldsymbol{\epsilon}) - \boldsymbol{E}(\boldsymbol{\epsilon}'\mathbf{A}\boldsymbol{\epsilon})\boldsymbol{E}(\boldsymbol{\epsilon}'\mathbf{A}\boldsymbol{\epsilon}) \\ &\quad - \boldsymbol{E}(\boldsymbol{\epsilon}'\mathbf{A}\boldsymbol{\epsilon})\boldsymbol{E}(\boldsymbol{\epsilon}'\mathbf{A}\boldsymbol{\epsilon}) + \left(\mathbf{E}(\boldsymbol{\epsilon}'\mathbf{A}\boldsymbol{\epsilon})\right)^2 \\ &= \boldsymbol{E}(\boldsymbol{\epsilon}'\mathbf{A}\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\mathbf{A}\boldsymbol{\epsilon}) - \left(E(\boldsymbol{\epsilon}'\mathbf{A}\boldsymbol{\epsilon})\right)^2 \\ &= \boldsymbol{E}(\boldsymbol{\epsilon}'\mathbf{A}\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\mathbf{A}\boldsymbol{\epsilon}) - \left(\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu} + \text{tr}(\mathbf{A}\boldsymbol{\Sigma})\right)^2 \end{aligned} \tag{1.45}$$

Now, the first term on the right hand-side of (1.45) is the expectation of the product of two Gaussian quadratic forms, or a quartic form. Using (8.2.4) from (Petersen & Pedersen, 2012),

$$\begin{aligned} \boldsymbol{E}(\boldsymbol{\epsilon}'\mathbf{A}\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\mathbf{A}\boldsymbol{\epsilon}) &= 2\text{tr}[\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}\boldsymbol{\Sigma}] + 4\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}\boldsymbol{\mu} \\ &\quad +\left(\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu} + \text{tr}(\mathbf{A}\boldsymbol{\Sigma})\right)^2. \end{aligned} \tag{1.46}$$

Substituting (1.46) into (1.45),

$$\begin{aligned} Var(\boldsymbol{\epsilon}'\mathbf{A}\boldsymbol{\epsilon}) &= 2\text{tr}[\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}\boldsymbol{\Sigma}] + 4\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}\boldsymbol{\mu} + (\text{tr}(\mathbf{A}\boldsymbol{\Sigma}) + \boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu})^2 - \left(\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu} + \text{tr}(\mathbf{A}\boldsymbol{\Sigma})\right)^2 \\ &= 2\text{tr}[\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}\boldsymbol{\Sigma}] + 4\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}\boldsymbol{\mu} \\ &= 2\text{tr}[(\mathbf{A}\boldsymbol{\Sigma})^2] + 4\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}\boldsymbol{\mu} \end{aligned}$$

which completes the proof. ■

Based on results in Lemma (1.1) and Theorem (1.4), we can extend the first-order multivariate delta method to a second-order approximation. Often, the second-order multivariate delta method is used as an alternative approximation to the first-order because in some cases, $\left(\frac{\partial g}{\partial \boldsymbol{\theta}}\right)' \boldsymbol{\Sigma}(\boldsymbol{\theta}) \left(\frac{\partial g}{\partial \boldsymbol{\theta}}\right)$ goes to zero due to a vanishing gradient problem (Sobel, 1982) in the variance approximation. However, for purposes here we will assume that $\left(\frac{\partial g}{\partial \boldsymbol{\theta}}\right)' \boldsymbol{\Sigma}(\boldsymbol{\theta}) \left(\frac{\partial g}{\partial \boldsymbol{\theta}}\right)$ exists and is non-null.

The second-order multivariate delta method uses a second-order Taylor expansion of $g(\widehat{\boldsymbol{\theta}}_n)$ around $\boldsymbol{\theta}$ as

$$g(\widehat{\boldsymbol{\theta}}_n) \approx g(\boldsymbol{\theta}) + \left(\frac{\partial g}{\partial \boldsymbol{\theta}}\right)' (\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) + \frac{1}{2}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})' \frac{\partial^2 g}{\partial \boldsymbol{\theta}\boldsymbol{\theta}'} (\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}), \tag{1.47}$$

where $\frac{\partial^2 g}{\partial \boldsymbol{\theta}\boldsymbol{\theta}'}$ is the matrix of second partial derivatives of g, or more commonly referred to as the Hessian of g. To simplify notation, let $\mathbf{d} = \frac{\partial g}{\partial \boldsymbol{\theta}}$ denote the gradient vector of g and $\mathbf{H} = \frac{\partial^2 g}{\partial \boldsymbol{\theta}\boldsymbol{\theta}'}$ denote the Hessian matrix of g, then (1.47) can be expressed as

$$\begin{aligned} g(\widehat{\boldsymbol{\theta}}_n) &\approx g(\boldsymbol{\theta}) + \mathbf{d}'(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) + \frac{1}{2}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})' \mathbf{H} (\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \\ g(\widehat{\boldsymbol{\theta}}_n) - g(\boldsymbol{\theta}) &\approx \mathbf{d}'(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) + \frac{1}{2}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})' \mathbf{H} (\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}). \end{aligned} \tag{1.48}$$

The variance of $g(\widehat{\boldsymbol{\theta}}_n) - g(\boldsymbol{\theta})$ can be calculated using the general variance formula

$$Var[g(\widehat{\boldsymbol{\theta}}_n) - g(\boldsymbol{\theta})] = E\left[\left(g(\widehat{\boldsymbol{\theta}}_n) - g(\boldsymbol{\theta})\right)^2\right] - E[g(\widehat{\boldsymbol{\theta}}_n) - g(\boldsymbol{\theta})]^2.$$

The term $E\left[\left(g(\widehat{\boldsymbol{\theta}}_n) - g(\boldsymbol{\theta})\right)^2\right]$ can be calculated by first expanding the square of (1.48) as,

$$\left(g(\widehat{\boldsymbol{\theta}}_n) - g(\boldsymbol{\theta})\right)^2 \approx \left(\mathbf{d}'(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) + \frac{1}{2}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})' \mathbf{H} (\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})\right)^2$$

$$= \begin{pmatrix} \mathbf{d}'(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})\mathbf{d}'(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) + \frac{1}{2}\mathbf{d}'(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})'\mathbf{H}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \\ + \frac{1}{2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})'\mathbf{H}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})\mathbf{d}'(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \\ + \frac{1}{4}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})'\mathbf{H}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})'\mathbf{H}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \end{pmatrix}$$

$$= \begin{pmatrix} \mathbf{d}'(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})'\mathbf{d} + \mathbf{d}'(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})'\mathbf{H}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \\ + \frac{1}{4}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})'\mathbf{H}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})'\mathbf{H}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \end{pmatrix}.$$

Taking the expected value of $\left(g(\hat{\boldsymbol{\theta}}_n) - g(\boldsymbol{\theta})\right)^2$,

$$E\left[\left(g(\hat{\boldsymbol{\theta}}_n) - g(\boldsymbol{\theta})\right)^2\right] \approx E\begin{bmatrix} \mathbf{d}'(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})'\mathbf{d} + \mathbf{d}'(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})'\mathbf{H}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \\ + \frac{1}{4}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})'\mathbf{H}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})'\mathbf{H}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \end{bmatrix}$$

$$= \mathbf{d}'\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\theta}})\mathbf{d} + \frac{1}{4}E\left[(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})'\mathbf{H}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})'\mathbf{H}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})\right].$$

The term $\frac{1}{4}E\left[(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})'\mathbf{H}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})'\mathbf{H}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})\right]$ is the expected value of a quartic form of Gaussians. Let $\mathbf{Q} = (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})'\mathbf{H}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})$, then this expected value can be derived by rearranging the variance formula as

$$E(\mathbf{QQ}') = Var(\mathbf{Q}) + E(\mathbf{Q})E(\mathbf{Q}').$$

Using the results from Theorem (1.4), we can see that $E(\mathbf{QQ}')$ is

$$E(\mathbf{QQ}') = 2\text{tr}\left[\left(\mathbf{H}\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\theta}})\right)^2\right] + 4\boldsymbol{\mu}'\mathbf{H}\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\theta}})\mathbf{H}\boldsymbol{\mu} + \left(\boldsymbol{\mu}'\mathbf{H}\boldsymbol{\mu} + \text{tr}\left(\mathbf{H}\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\theta}})\right)\right)\left(\boldsymbol{\mu}'\mathbf{H}\boldsymbol{\mu} + \text{tr}\left(\mathbf{H}\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\theta}})\right)\right). \quad (1.49)$$

We know from Theorem (1.3) that the asymptotic distribution of $\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}$ is normal with zero mean vector, or $\boldsymbol{\mu} = \mathbf{0}$ so (1.49) simplifies to

$$E(\mathbf{QQ}') = 2\text{tr}\left[\left(\mathbf{H}\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\theta}})\right)^2\right] + \left(\text{tr}\left(\mathbf{H}\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\theta}})\right)\right)\left(\text{tr}\left(\mathbf{H}\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\theta}})\right)\right)$$

$$= 2\text{tr}\left[\left(\mathbf{H}\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\theta}})\right)^2\right] + \left(\text{tr}\left(\mathbf{H}\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\theta}})\right)\right)^2. \tag{1.50}$$

Following identities given in Petersen and Pedersen (2012), the expected value of the quartic form in (1.50) becomes

$$\begin{aligned}\frac{1}{4}E(\mathbf{QQ}') &= \frac{1}{4}\left(2\text{tr}\left[\left(\mathbf{H}\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\theta}})\right)^2\right] + \left(\text{tr}\left(\mathbf{H}\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\theta}})\right)\right)^2\right) \\ &= \frac{1}{2}\text{tr}\left[\left(\mathbf{H}\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\theta}})\right)^2\right] + \frac{1}{4}\left(\text{tr}\left(\mathbf{H}\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\theta}})\right)\right)^2.\end{aligned} \tag{1.51}$$

Therefore,

$$E\left[\left(g(\hat{\boldsymbol{\theta}}_n) - g(\boldsymbol{\theta})\right)^2\right] = \mathbf{d}'\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\theta}})\mathbf{d} + \frac{1}{2}\text{tr}\left[\left(\mathbf{H}\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\theta}})\right)^2\right] + \frac{1}{4}\left(\text{tr}[\mathbf{H}\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\theta}})]\right)^2.$$

Next, $E[g(\hat{\boldsymbol{\theta}}_n) - g(\boldsymbol{\theta})]$ is

$$\begin{aligned}E[g(\hat{\boldsymbol{\theta}}_n) - g(\boldsymbol{\theta})] &\approx E\left[\mathbf{d}'(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) + \frac{1}{2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})'\mathbf{H}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})\right] \\ &= \frac{1}{2}\text{tr}[\mathbf{H}\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\theta}})]\end{aligned}$$

which implies that

$$E[g(\hat{\boldsymbol{\theta}}_n) - g(\boldsymbol{\theta})]^2 = \frac{1}{4}\left(\text{tr}[\mathbf{H}\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\theta}})]\right)^2. \tag{1.52}$$

Combining (1.51) and (1.52), the variance of $g(\hat{\boldsymbol{\theta}}_n) - g(\boldsymbol{\theta})$ is

$$\begin{aligned}&= \mathbf{d}'\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\theta}})\mathbf{d} + \frac{1}{2}\text{tr}\left[\left(\mathbf{H}\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\theta}})\right)^2\right] + \frac{1}{4}\left(\text{tr}[\mathbf{H}\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\theta}})]\right)^2 - \frac{1}{4}\left(\text{tr}[\mathbf{H}\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\theta}})]\right)^2 \\ &= \mathbf{d}'\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\theta}})\mathbf{d} + \frac{1}{2}\text{tr}\left[\left(\mathbf{H}\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\theta}})\right)^2\right],\end{aligned}$$

which matches the result in Preacher et al. (2007). The second-order multivariate delta method is similar to the first-order approximation, except that a second term is added to the first-order variance

$$Var\left[g\left(\widehat{\boldsymbol{\theta}}_n\right) - g(\boldsymbol{\theta})\right] = \underbrace{\mathbf{d}'\widehat{\boldsymbol{\Sigma}}\left(\widehat{\boldsymbol{\theta}}\right)\mathbf{d}}_{\text{first-order}} + \underbrace{\frac{1}{2}\text{tr}\left[\left(\mathbf{H}\widehat{\boldsymbol{\Sigma}}\left(\widehat{\boldsymbol{\theta}}\right)\right)^2\right]}_{\text{second-order}}. \quad (1.53)$$

Having derived the necessary variances, $100(1-\alpha)\%$ confidence intervals for indirect effects using the second-order delta method are of the form

$$g(\boldsymbol{\beta}|\mathbf{v}) \pm z_{\alpha/2} Var[g(\boldsymbol{\beta}|\mathbf{v})]^{1/2}, \quad (1.54)$$

where $z_{\alpha/2}$ is the $\alpha/2$ quantile of the standard normal distribution and $Var[g(\boldsymbol{\beta}|\mathbf{v})]^{1/2}$ is the standard error. The confidence interval in (1.54) is estimated by

$$g\left(\widehat{\boldsymbol{\beta}}|\mathbf{v}\right) \pm z_{\alpha/2} Var\left[g\left(\widehat{\boldsymbol{\beta}}|\mathbf{v}\right)\right]^{1/2}, \quad (1.55)$$

where $g\left(\widehat{\boldsymbol{\beta}}|\mathbf{v}\right)$ is the estimated indirect effect and $Var\left[g\left(\widehat{\boldsymbol{\beta}}|\mathbf{v}\right)\right]^{1/2}$ is the estimated standard error using the first- or second-order multivariate delta method. Confidence intervals for two of the models discussed earlier in this Chapter (Figure [1.4]) are presented in Appendix B.

From a mathematical point of view, a quadratic approximation to a nonlinear function results in better accuracy than a linear approximation. Current research, however, has been inconclusive in determining which approximation method results in better statistical performance for indirect effect testing. For simple mediation models, MacKinnon (1992) found that standard errors for the second-order approximation of standard errors were less biased than first-order standard errors; however, Mackinnon et al. (1995) found the opposite. Preacher et al. (2007) found that in general, for the five moderated mediation models presented in Figure (1.4), the second-order delta method led to equal or slightly lower rejection rates compared to the first-order delta method. Given the mixed results, in practice, researchers use both the first-order and the second-order multivariate delta method to construct confidence intervals for indirect effects. However, the second-order

variance term is often negligible (MacKinnon et al., 2002; Preacher & Hayes, 2004) and thus can be safely ignored, which is why the default in popular SEM software packages such as Mplus and Lavaan in R use the first-order approximation.

1.5.2. Nonparametric bootstrap. The bootstrap is the most popular method of hypothesis testing for indirect effects (Cheung, 2007; MacKinnon et al., 1995; MacKinnon et al., 2004; MacKinnon, Lockwood, Hoffman, West, & Virgil, 2002; Preacher et al., 2007; Preacher & Hayes, 2008). Figure (1.5) presents a schematic diagram of the bootstrap applied to mediation-type data structures. On the left-hand side of Figure (1.5) is the real world. In the real world an unknown probability mechanism P yields an observed data set $\boldsymbol{\mathcal{D}}$ by random sampling (Efron & Tibshirani, 1993). Note, the variables in $\boldsymbol{\mathcal{D}}$ correspond to the same variables in the vector $\mathbf{v}$ described earlier in this chapter. The use of notation here is to emphasize that $\boldsymbol{\mathcal{D}}$ is a complete data set with n samples. Using the observed data, we estimate the unknown regression parameters $\boldsymbol{\beta}$ using ML (as discussed in the previous section) and use our estimates $\widehat{\boldsymbol{\beta}}$ to calculate indirect effects, denoted by $\widehat{\boldsymbol{\theta}}$. Furthermore, we often wish to know something about $\widehat{\boldsymbol{\theta}}$'s statistical behavior (e.g., bias, standard error, or confidence interval). Here, we use a vector notation for the indirect effects because using the matrix algebra approaches described earlier in this chapter to derive point estimators, $k \geq 1$ indirect effects can be calculated simultaneously.

On the right-hand side of Figure (1.5) is the bootstrap setting. In the bootstrap world, the empirical probability mechanism $\hat{P}$ is used to generate B bootstrap samples $\boldsymbol{\mathcal{D}}^* = \{\mathbf{v}_i^*\}_{i=1}^n = \{x_i^*, m_i^*, \ldots, y_i^*\}_1^n$ by randomly sampling uniformly from the rows of $\boldsymbol{\mathcal{D}}$ with replacement. That is, if $\mathbf{v}_i$ is the ith row of the observed data, then

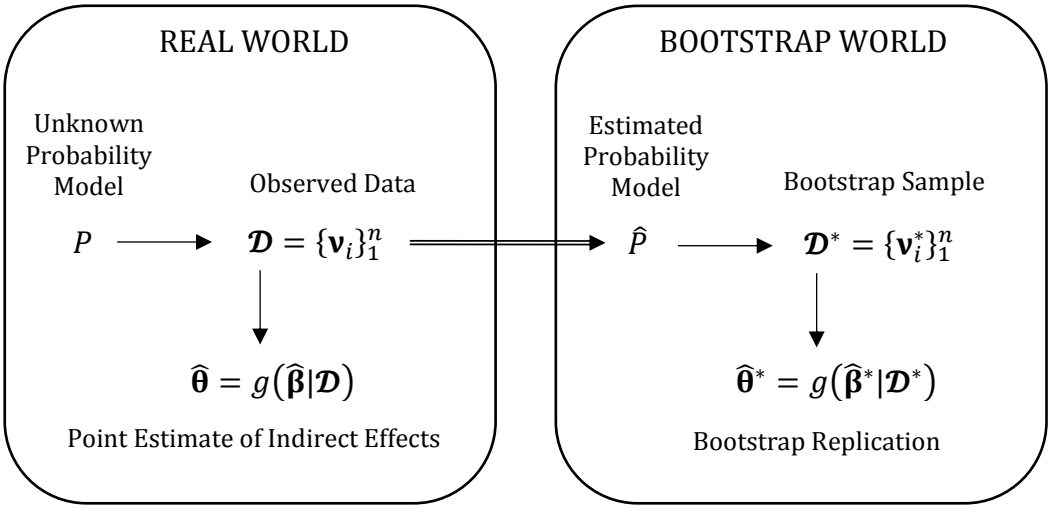


Figure 1.5. Diagram of the bootstrap applied to mediation-type data structures; adapted from Efron and Tibshirani (1993). $\boldsymbol{\mathcal{D}}$ denotes an observed data set with n samples and p variables. The dimensionality of p reflects the variables in $\mathbf{v}$ from previous sections.

$$P(\mathbf{v}_i \text{ sampled}) = \frac{1}{n}.$$

In each of the B bootstrap samples, we calculate a bootstrap replication of the unknown regression parameters and use our estimates $\widehat{\boldsymbol{\beta}}^*$ to calculate indirect effects, denoted by $\widehat{\boldsymbol{\theta}}^*$. Together, the B bootstrap replications form the bootstrap distribution of the indirect effects of interest. The bootstrap distribution is an estimate of the indirect effect's sampling distribution. Therefore, with an empirical sampling distribution at hand, biases, standard errors, and confidence intervals for indirect effects can be calculated directly.

The bootstrap approximation of the sampling distribution is not without error. The bootstrap approximates with sampling distribution with three sources of approximation error. First, simulation error arises from using finitely many replications to generate a statistic's sampling distribution (Efron, 1987). Often simulation error can be attenuated by

drawing a large number of bootstrap replications (e.g., $B \geq 1{,}000$). Second, statistical error arises because the empirical sampling distribution generated by the bootstrap replications under the estimated model $\hat{P}$ is not equivalent to the theoretical sampling distribution of the statistic under the true data-generating process P (Bickel & Freedman, 1981). Often statistical error can be attenuated by using bias-correction techniques (described below; Efron, 1987) for the final bootstrap estimates. Lastly, specification error arises because the data does not exactly follow our specified model. As a result, simulating the model never quite matches the actual sampling distribution. Specification error is attenuated by resampling from the data as opposed to simulating the probability model; this is the key idea of nonparametric bootstrapping (Efron & Tibshirani, 1993).

Before performing a bootstrap analysis based on Figure (1.5), we need to consider two problems. First, we have to determine how to estimate the probability mechanism P from the observed data $\boldsymbol{\mathcal{D}}$, denoted by the double arrow going from $\boldsymbol{\mathcal{D}}$ to $\hat{P}$. There are several ways to conceptualize the probability model $P \longrightarrow \boldsymbol{\mathcal{D}}$, however, we will only consider the nonparametric or row-resampling approach. In the nonparametric approach, P is the distribution function that generated $\boldsymbol{\mathcal{D}}$. As such, $P \longrightarrow \boldsymbol{\mathcal{D}}$ means that $\boldsymbol{\mathcal{D}}$ is a random sample from P. If we treat $\boldsymbol{\mathcal{D}}$ as the complete population of data, we can estimate P by the empirical distribution function $\hat{P}$ based on $\boldsymbol{\mathcal{D}}$. The empirical distribution function is defined to be the discrete distribution that puts probability $1/n$ on each $\mathbf{v}_i, i = 1, \ldots, n$ (Efron & Tibshirani, 1993). The second problem is determining how to simulate bootstrap data from $\hat{P}$ according to the relevant data structure. As mentioned above, to reduce specification error of bootstrapping, instead of simulating the probability model $\hat{P}$, we randomly resample uniformly with replacement n rows of $\boldsymbol{\mathcal{D}}$ to obtain a bootstrap sample $\boldsymbol{\mathcal{D}}^*$.

After we have generated B bootstrap samples and calculated B bootstrap replicates for the indirect effects of interest, the overall bootstrap estimate $\bar{\hat{\boldsymbol{\theta}}}^*$ of $\boldsymbol{\theta}$ is given by the average of the B replicates,

$$\bar{\hat{\boldsymbol{\theta}}}^* = \frac{1}{B}\sum_{b=1}^{B} \hat{\boldsymbol{\theta}}^{*b} .$$

Empirically, the indirect effect is either estimated from the sample data $\hat{\boldsymbol{\theta}}$ (i.e., without bootstrapping; MacKinnon et al., 2004) or as the mean of the bootstrap distribution of the indirect effect $\bar{\hat{\boldsymbol{\theta}}}^*$(Preacher & Hayes, 2004). In the present study we use the former method for estimating indirect effects since this method is most commonly implemented in popular SEM software packages (Rosseel, 2012). In practice, hypothesis testing of indirect effects is generally done using bootstrapped confidence intervals (CIs). Many methods for CIs from $\{\hat{\boldsymbol{\theta}}^{*b}\}_1^B$ have been proposed for indirect effects such as the percentile interval, bias-corrected (BC) interval, and the bias-corrected and accelerated (BCa) interval (Efron, 1987; MacKinnon et al., 2004; Preacher et al., 2007). For the purpose of the present study, however, we only consider the BC confidence interval because research has empirically shown the BC confidence interval to perform similar to the BCa interval in terms of Type I Error and power for conditional indirect effects, with smaller computational costs (Preacher et al., 2007), and better than BCa intervals and other resampling approaches (e.g., jackknife interval, bootstrap-t interval, Monte Carlo interval) in terms of Type I Error, power, and coverage probabilities for unconditional indirect effects (MacKinnon et al., 2004).

The $100(1-\alpha)\%$ BC confidence interval for the kth element of $\boldsymbol{\theta}$ can be constructed using the percentiles α_{LL} and α_{UL} of $\{\hat{\boldsymbol{\theta}}_k^{*b}\}_{b=1}^{B}$. Here

$$\alpha_{LL} = \Phi(2z_0 + z^{\alpha})$$

and

$$\alpha_{UL} = \Phi(2z_0 + z^{1-\alpha}),$$

where Φ is the standard cumulative normal distribution function and z^{α} is the α percentile of the standard normal distribution and

$$z_0 = \Phi^{-1}\left(\frac{\sum_{b=1}^{B} I\left(\hat{\boldsymbol{\theta}}_k^{*b} < \hat{\boldsymbol{\theta}}_k\right)}{B}\right),$$

where $I\left(\hat{\boldsymbol{\theta}}_k^{*b} < \hat{\boldsymbol{\theta}}_k\right)$ is an indicator function defined as

$$I\left(\hat{\boldsymbol{\theta}}_k^{*b} < \hat{\boldsymbol{\theta}}_k\right) = \begin{cases} 1, & \hat{\boldsymbol{\theta}}_k^{*b} < \hat{\boldsymbol{\theta}}_k \\ 0, & \hat{\boldsymbol{\theta}}_k^{*b} \geq \hat{\boldsymbol{\theta}}_k \end{cases}.$$

Here, z_0 is defined to measure the median bias of the bootstrap distribution of the indirect effect. In the case of no bias, the bootstrap distribution is symmetric and $\sum_{b=1}^{B} I\left(\hat{\boldsymbol{\theta}}_k^{*b} < \hat{\boldsymbol{\theta}}_k\right) / B = 0.5$, which implies that $\Phi^{-1}(0.5) = 0$ and the CI limits reduce to $\alpha_{LL} = \Phi(z^{\alpha})$ and $\alpha_{UL} = \Phi(z^{1-\alpha})$.

1.5.3. Bayesian bootstrap. The Bayesian analog of the nonparametric bootstrap (i.e., Efron's bootstrap) was proposed by Rubin (1981) called the Bayesian bootstrap (BB). To understand the BB, suppose that the observed data vector $\mathbf{x}$ can assume at most K distinct values given by the vector $\mathbf{d} = [d_1, d_2, \ldots, d_K]$. Let $\mathbf{w} = [w_1, w_2, \ldots, w_K]'$ be the vector of probabilities defined as

$$Pr(x_i = d_k|\mathbf{w}) = w_k, \qquad \sum_{k=1}^{K} w_k = 1.$$

Assuming $x_1, x_2, \ldots, x_n$ given $\mathbf{w}$ are independent, the BB applies the improper prior distribution on $\mathbf{w}$

$$\pi(\mathbf{w}) = \begin{cases} \prod_{k=1}^{K} w_k^{-1}, & \text{if } \sum_{k=1}^{K} w_k = 1 \\ 0, & \text{otherwise} \end{cases} \tag{1.56}$$

Under the nonparametric bootstrap, $\mathbf{w}$ is also a vector of sampling weights, but now $w_j = \frac{n_j}{n}$ is the probability that a sample point falls in category j, for K distinct categories. The nonparametric bootstrap distribution, obtained by sampling with replacement from the data, is realized as sampling the category proportions from a multinomial distribution

$$n\mathbf{w} \sim \text{Multinomial}(n, \hat{\mathbf{w}}), \tag{1.57}$$

where $\hat{\mathbf{w}} = [\hat{w}_1, \ldots, \hat{w}_K]'$ are the observed probabilities. The multinomial distribution of the nonparametric bootstrap sampling weights in (1.57) is proportional to

$$P(n\mathbf{w}) \propto \prod_{k=1}^{K} \hat{w}_k^{nw_k} = \prod_{k=1}^{K} \hat{w}_k^{n_k} \tag{1.58}$$

The mean and variance of the jth nonparametric bootstrap weight from (1.58) are given by

$$E(\mathbf{w}_j) = E\left(\frac{n_j}{n}\right)$$

$$= \frac{1}{n}$$

and

$$Var(\mathbf{w}_j) = Var\left(\frac{n_j}{n}\right)$$

$$= \frac{1}{n^2}\left(1 - \frac{1}{n}\right)$$

$$= \frac{n-1}{n^3}$$

since each $n_j \sim$ Binomial $\left(n, \frac{1}{n}\right)$.

Applying the improper prior in (1.56) for $\mathbf{w}$, we obtain the posterior distribution of $\mathbf{w}$ as proportional to

$$P(\mathbf{w}|\mathbf{x}) \propto \prod_{k=1}^{K} w_k^{n_k - 1},$$

which is the kernel of a $(K-1)$-dimensonal Dirichlet distribution (Cohen, 1997). Under this posterior distribution, the mean and variance of the jth BB sampling weight are given by

$$E(\mathbf{w}|\mathbf{x}) = \frac{1}{\sum_{k=1}^{K} n_k}$$

$$= \frac{1}{n}$$

and

$$Var(\mathbf{w}|\mathbf{x}) = \frac{n-1}{n^2(n+1)}.$$

Comparing Efron's nonparametric bootstrap to the Rubin's BB bootstrap, we see the following relations,

$$E(\mathbf{w}_{\text{Efron}}) = E(\mathbf{w}_{\text{Rubin}}) = \frac{1}{n}$$

$$Var(\mathbf{w}_{\text{Efron}}) = Var(\mathbf{w}_{\text{Rubin}}) \left(\frac{n+1}{n}\right).$$

If we denote any estimator by $h(\cdot)$ (e.g., indirect effect), the nonparametric bootstrap distribution of our estimator $h(\mathbf{w}_{\text{Efron}})$ will closely approximate the posterior distribution of $h(\mathbf{w}_{\text{Rubin}})$ (Hastie et al., 2009).

Although the BB is slightly more complicated than the nonparametric bootstrap, BB can be conducted using a two-step procedure as described in Rubin (1981) and Cohen (1997): (1) independently draw n uniform random variables between 0 and 1 and order them such that $u_1 < u_2 < \cdots < u_n$, where $u_1 = 0$ and $u_n = 1$, (2) define a vector of differences $\mathbf{w} = [u_1 - u_0, u_2 - u_1, \ldots, u_{n-1} - u_n]$. Draw each of the n values in $\mathbf{x}^{*b}$ by drawing from $x_1, x_2, \ldots, x_n$ with associated probabilities defined by $\mathbf{w}$. In practice, the BB can be sampled in two different stages: (1) the number of bootstrap samples, and (2) the number of resamples to draw at each bootstrap sample. The idea behind a two-stage sampling scheme for BB is used to obtain bootstrap samples that reflect the correct frequency as the weights defined in $\mathbf{w}$. In other words, since the sampling weights are not uniform, more samples can be drawn (e.g., 1,000) for each Bayesian bootstrap sample so that the frequency of rows in a particular sample more accurately reflect the sampling weight associated with the row. Importantly, despite the difference in implementation, Lo (1987) showed that the BB has the same desirable asymptotic convergence properties as the nonparametric bootstrap.

In missing data applications, the most common application of the BB is in multiple imputation of missing data, specifically using the approximate Bayesian bootstrap (ABB; Rubin, 1987). Rubin (1996) recommends using the BB to incorporate a systematic between-imputation component of variability when drawing missing values from observed values. In other imputation scheme, Rubin (1996) notes that the BB can be used to automatically incorporate parameter uncertainty in the estimation of population parameters. In non-missing data applications, research has shown that the BB often leads to similar and sometimes narrower confidence intervals than the nonparametric bootstrap

(Taddy, Chen, Yu, & Wyle, 2015). Theoretically, the nonparametric bootstrap can be viewed as an approximation to the BB (Hastie, Tibshirani, & Friedman 2009), which would explain the similarities in confidence interval lengths (Taddy et al., 2015). In practice, the BB distribution tends to be smoother than the nonparametric bootstrap distribution (Rubin, 1981), which is due to its smoother choices of sampling weights.

Recent research has demonstrated the benefits of Bayesian estimation for testing indirect effects in mediation (Enders et al., 2013; Yuan & MacKinnon, 2009) and moderated mediation models (Wang & Preacher, 2015). A drawback with proper Bayesian estimation, however, is that model specification requires additional steps, that is, by specifying prior distributions for all model parameters (Muthén & Asparouhov, 2012). Although some SEM software (e.g., Mplus) sets default prior distributions for all model parameters automatically, other non-SEM software (e.g., WinBUGs, PyMCMC, Stan) that can estimate Bayesian SEMs, require the user to explicitly specify priors for all model parameters. As an SEM model becomes increasingly complex (e.g., additional parameters and mixed variable types), Bayesian estimation can also become increasingly complex (Lee, 2007). In some cases, to exactly sample from a parameter's posterior distribution, models may need to be reparametrized to use the same sampler across all parameters or hybrid MCMC samplers (e.g., Gibbs and sequential Monte-Carlo) may need to be used (Gelman, Carlin, Stern, Dunson, Vehtari, & Rubin, 2013).

For example, consider the simple mediation model in (1.3) (ignoring the exogenous variable model) where the errors of $f(M|X, \boldsymbol{\theta}_M)$ follow a normal distribution and the errors of $f(Y|X, M, \boldsymbol{\theta}_Y)$ follow a Bernoulli distribution. The regression analogs of this setup are to use a linear regression model to estimate $\boldsymbol{\theta}_M$ for $f(M|X, \boldsymbol{\theta}_M)$ and logistic regression model

to estimate $\boldsymbol{\theta}_Y$ for $f(Y|X, M, \boldsymbol{\theta}_Y)$. For $\boldsymbol{\theta}_M$, common priors in practice include using a non-informative prior such as Jeffrey's prior or conjugate priors for $\boldsymbol{\theta}_M$, which include an inverse-gamma distribution for the variance and normal distributions for the regression parameters (Gelman et al., 2013). In both cases, Jeffrey's prior and conjugate priors lead to a tractable normal-inverse Wishart posterior distribution for $P(\boldsymbol{\theta}_M|M, X)$ that can be exactly sampled from using a standard Gibbs sampler (Press, 1972).

With regards to $\boldsymbol{\theta}_Y$, assume we follow common practice and use normal priors for the regression parameters in $\boldsymbol{\theta}_Y$ (Groenewald & Mokgatlhe, 2005). Now, since the normal distribution is not the conjugate prior of the likelihood function in logistic regression, the posterior distribution for $P(\boldsymbol{\theta}_Y|Y, M, X)$ is difficult to calculate. To exactly sample from $P(\boldsymbol{\theta}_Y|Y, M, X)$ using a Gibbs sampler, since the full conditionals are intractable, we could reparametrize the logistic model using the latent variable approach described in (Groenewald & Mokgatlhe, 2005) and then apply a Gibbs sampler. Alternatively, we could use a sequential Monte-Carlo sampler as described in Geweke, Durham, and Hu (2013). Although other methods exist, such as using approximate MCMC samplers, these methods are not without their drawbacks. For instance, if approximate MCMC methods are used, careful diagnostic checking is required for each parameter's posterior distribution to determine if the approximate MCMC sampler converged to the stationary multivariate posterior distribution.

The complexities of proper Bayesian SEM estimation and limited SEM software to conduct proper Bayesian inference may contribute to the widespread use of alternative methods to testing indirect effects, such as the bootstrap. Empirically, the bootstrap has been found to work well with unconditional and conditional indirect effects testing, despite

some minor drawbacks with small sample sizes (Koopman, Howe, Hollenbeck, & Sin, 2015). Importantly, by simply changing the resampling scheme of the classic bootstrap, the Bayesian bootstrap simulates a parameter's posterior distribution without having to specify priors explicitly on the model parameters. However, with regards to estimation and inference in mediation models, despite its similarity to the nonparametric bootstrap and ease of implementation, the performance of the BB in estimating indirect type effects is unknown.

CHAPTER 2

MISSING DATA IN STRUCTURAL EQUATION MODELS

2.1. Overview

Missing data is a pervasive problem that affects nearly all research domains. In psychological research, the most common data used in structural equation modeling is that of survey data, which is inherently susceptible to missing data problems (Rubin, 1987). Fortunately, several techniques are available to estimate SEMs with missing data. This chapter introduces two methods to handle missing data in SEMs, namely model-based estimation techniques, and imputation techniques.

Missing data can be categorized based on the pattern of missing data and the missing data mechanism. The pattern of missing data describes which values are observed and which values are missing. Although there are many distinct patterns of missingness (e.g., see Little & Rubin, 2002 for an overview), we will only consider the so-called 'general case' of missingness. Figure 2.1 presents the case of a complete multivariate data set; an example of a general missingness pattern for a multivariate data set is given in Figure 2.2. More importantly, the second category for missing data classification regards the type or mechanism of missing data. Here, the type of missing data describes the relationship between the missingness and the values of variables in the data matrix.

2.1.1. Missing data mechanisms. Rubin (1976) describes three distinct types of missing data: (1) missing completely at random (MCAR), (2) missing at random

	1	2	3	⋯	p
1					
2					
3					
4					
5					
6					
.					
.					
.					
n					

Figure 2.1. Complete multivariate data set.

	1	2	3	⋯	p
1	?				
2			?		
3					
4	?				
5			?		
6					?
.		?			
.				?	
.					
n					

Figure 2.2. General missingness pattern for incomplete multivariate data set, where ? denotes a missing value.

(MAR), and (3) missing not at random (MNAR). Let $\mathbf{R}$ be an $n \times p$ matrix of indicator values defined by

$$\mathbf{R} = \begin{cases} 1, & \text{if } x_{ij} \text{ is observed} \\ 0, & \text{if } x_{ij} \text{ is missing} \end{cases}, \tag{2.1}$$

and let $\mathbf{X} = (\mathbf{X}_{\text{obs}}, \mathbf{X}_{\text{mis}})$, where $\mathbf{X}_{\text{obs}}$ denote the n_1 rows of fully observed data and $\mathbf{X}_{\text{mis}}$

denote the $n_2 = n - n_1$ rows containing missing values on $p \geq 1$ variables (Schafer, 1997). The matrix $\mathbf{R}$ stores the (i, j) locations of the missing data in $\mathbf{X}$. To formalize the mechanisms of missing data, a probability model is posited between $\mathbf{R}$ and $\mathbf{X}$, $P(\mathbf{R}|\mathbf{X}_{\text{obs}}, \mathbf{X}_{\text{mis}}, \boldsymbol{\psi})$, where $\boldsymbol{\psi}$ is the parameter vector corresponding to the parameters of the missing data model (van Buuren, 2012). The distribution of $\mathbf{R}$ may depend on $\mathbf{X} = (\mathbf{X}_{\text{obs}}, \mathbf{X}_{\text{mis}})$, and this dependence is what characterizes the missing data mechanism; this relation is what is referred to as the missing data model (Rubin, 1987).

In the simplest missing data model, the data are said to be MCAR if the distribution of $\mathbf{R}$ does not depend on $\mathbf{X}_{\text{obs}}$ or $\mathbf{X}_{\text{mis}}$

$$P(\mathbf{R} = 0|\mathbf{X}_{obs}, \mathbf{X}_{\text{mis}}, \boldsymbol{\psi}) = P(\mathbf{R} = 0|\boldsymbol{\psi}), \tag{2.2}$$

for all $\mathbf{X}$ and $\boldsymbol{\psi}$. Under model (2.2), the probability of missing data depends only on some parameter $\boldsymbol{\psi}$. MCAR is the ideal missing data mechanism in which the missing data values are a simple random sample of all data values (Graham, 2009). Compared to MCAR, MAR is a more realistic assumption of the missing data mechanism in practice. Specifically, MAR assumes that the missing values behave like a random sample of all values within subclasses defined by observed data (Schafer & Graham, 2002). The data are said to be MAR if the distribution of $\mathbf{R}$ does not depend on $\mathbf{X}_{\text{mis}}$

$$P(\mathbf{R} = 0|\mathbf{X}_{\text{obs}}, \mathbf{X}_{\text{mis}}, \boldsymbol{\psi}) = P(\mathbf{R} = 0|\mathbf{X}_{\text{obs}}, \boldsymbol{\psi}), \tag{2.3}$$

for all $\mathbf{X}_{\text{mis}}$ and $\boldsymbol{\psi}$. Under MAR, the missingness probability may depend on observed information. Lastly, the most restrictive missing data mechanism is MNAR. The data are said to be MNAR if the distribution of $\mathbf{R}$ depends on the missing data, $\mathbf{X}_{\text{mis}}$, or equivalently that

$$p(\mathbf{R} = 0|\mathbf{X}_{\text{obs}}, \mathbf{X}_{\text{mis}}, \boldsymbol{\psi})$$

does not simplify. Under MNAR, the probability of missingness depends on observed information and also on unobserved information, including $\mathbf{X}_{\text{mis}}$ (Enders, 2010).

2.1.2. Ignorability. In practical applications, the focus is on making inferences about the model parameters $\boldsymbol{\theta}$ as opposed to the missing data parameters $\boldsymbol{\psi}$. In this case, $\boldsymbol{\psi}$ is treated as a nuisance parameter (Basu, 1977; Casella & Berger, 2001). Importantly, as Schafer (1997) notes, $\boldsymbol{\theta}$ refers to the parameters of the complete data $\mathbf{X}$ and not the parameters for the distribution of $\mathbf{X}_{\text{obs}}$ alone. As such, the end goal of the analysis is to make inferences about the parameters of the complete-data model as opposed to the parameters of the marginal distribution of only the observed data. In terms of estimation and inference of $\boldsymbol{\theta}$, the analysis would be simplified if the nuisance parameter $\boldsymbol{\psi}$ could be ignored. This latter simplification underscores the importance of distinguishing between MCAR, MAR, and MNAR because these mechanisms clarify the conditions under which we can estimate and make inferences about $\boldsymbol{\theta}$ without having to know $\boldsymbol{\psi}$ (Rubin, 1976).

With missing data, the observed data consist not only of $\mathbf{X}_{\text{obs}}$ but also of $\mathbf{R}$, which implies that the joint probability distribution of the observed data is given by

$$P(\mathbf{X}_{\text{obs}}, \mathbf{R}|\boldsymbol{\theta}, \boldsymbol{\psi}) = \int P(\mathbf{X}, \mathbf{R}|\boldsymbol{\theta}, \boldsymbol{\psi}) d\mathbf{X}_{\text{mis}}$$

$$= \int P(\mathbf{R}|\mathbf{X}, \boldsymbol{\psi}) P(\mathbf{X}|\boldsymbol{\theta}) d\mathbf{X}_{\text{mis}},$$

where the integral is replaced by a summation in the discrete case. The condition required to estimate and make inferences about $\boldsymbol{\theta}$ without knowing the $\boldsymbol{\psi}$ is called ignorability. Ignorability has two conditions: (1) the missing data mechanism is MAR and (2) $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ are distinct (Rubin, 1976, 1987; Schafer, 1997). For the first condition, recall that under MAR the missing data does not depend on the missing values themselves (see Equation

[2.3]). For the second condition, we can describe the distinctiveness condition from two perspectives. From a frequentist perspective the distinctiveness of $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ implies that the joint parameter space of $(\boldsymbol{\theta}, \boldsymbol{\psi})$ is the Cartesian cross-product of the individual parameter spaces for $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ (Schafer, 1997). From the joint probability model of the observed data, under the MAR assumption,

$$\begin{aligned} P(\mathbf{X}_{\text{obs}}, \mathbf{R}|\boldsymbol{\theta}, \boldsymbol{\psi}) &= P(\mathbf{R}|\mathbf{X}_{\text{obs}}, \boldsymbol{\psi}) \int P(\mathbf{X}|\boldsymbol{\theta}) d\mathbf{X}_{\text{mis}} \\ &= P(\mathbf{R}|\mathbf{X}_{\text{obs}}, \boldsymbol{\psi}) P(\mathbf{X}_{\text{obs}}|\boldsymbol{\theta}). \end{aligned} \tag{2.4}$$

From (2.4), the likelihood of the observed data under MAR can thus be factored into two pieces, one pertaining to the nuisance parameter $\boldsymbol{\psi}$, $P(\mathbf{R}|\mathbf{X}_{\text{obs}}, \boldsymbol{\psi})$, and the other pertaining to the parameter of interest $\boldsymbol{\theta}$, $P(\mathbf{X}_{\text{obs}}|\boldsymbol{\theta})$. When the two parameters are distinct, then likelihood-based inferences about $\boldsymbol{\theta}$ will be unaffected by $P(\mathbf{R}|\mathbf{X}_{\text{obs}}, \boldsymbol{\psi})$ (Little & Rubin, 2002). Under these conditions, the missing data mechanism can be safely ignored (Little & Rubin, 1987; Rubin, 1976), that is, we do not need to consider the model for $\mathbf{R}$ nor the nuisance parameters $\boldsymbol{\psi}$ when making inferences about $\boldsymbol{\theta}$. Little and Rubin (1987) refer to the factor in (2.4) as the likelihood ignoring the missing data mechanism as the observed-data likelihood

$$L(\mathbf{X}_{\text{obs}}|\boldsymbol{\theta}) \propto P(\mathbf{X}_{\text{obs}}|\boldsymbol{\theta}).$$

From a Bayesian perspective, the distinctiveness assumption implies that a joint distribution applied to $(\boldsymbol{\theta}, \boldsymbol{\psi})$ factors into independent marginal priors for $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ (Schafer, 1997). In Bayesian analysis, all inferences are based on a posterior probability distribution for unknown parameters that conditions on the quantities that are observed (Gelman et al., 2013). For instance, let $\boldsymbol{\theta}$ denote the parameter vector for a given model and $\mathbf{X}$ denote the observed data. Bayesian estimation involves three general steps: (1)

specifying a joint probability model for all observable and unobservable quantities in a problem, $P(\mathbf{X}, \boldsymbol{\theta})$, (2) deriving the posterior distribution, that is, the distribution of the model parameters conditioning on the observed data, $P(\boldsymbol{\theta}|\mathbf{X})$, and (3) evaluating the fit of the model and statistical estimates from the posterior distribution. Using basic rules about conditional probabilities, the joint probability model $P(\mathbf{X}, \boldsymbol{\theta})$ can be written as the product of two densities, the prior distribution for $\boldsymbol{\theta}$ and data distribution for $\mathbf{X}|\boldsymbol{\theta}$,

$$\underbrace{P(\mathbf{X}, \boldsymbol{\theta})}_{\text{Joint}} = \underbrace{P(\mathbf{X}|\boldsymbol{\theta})}_{\text{Data}} \underbrace{P(\boldsymbol{\theta})}_{\text{Prior}}.$$

In order to make probability statements about $\boldsymbol{\theta}$ given $\mathbf{X}$, the posterior distribution is needed. After specifying the joint probability model, by conditioning on the observed data and applying Bayes' rule, the posterior distribution can be derived

$$\begin{aligned} P(\boldsymbol{\theta}|\mathbf{X}) &= \frac{P(\mathbf{X}, \boldsymbol{\theta})}{P(\mathbf{X})} \\ &= \frac{P(\mathbf{X}|\boldsymbol{\theta})P(\boldsymbol{\theta})}{P(\mathbf{X})}, \end{aligned} \tag{2.5}$$

where $P(\mathbf{X}) = \int P(\mathbf{X}|\boldsymbol{\theta})P(\boldsymbol{\theta})d\boldsymbol{\theta}$ for continuous $\boldsymbol{\theta}$ or $P(\mathbf{X}) = \sum_{\boldsymbol{\theta}} P(\mathbf{X}|\boldsymbol{\theta})P(\boldsymbol{\theta})$ for discrete $\boldsymbol{\theta}$. Given that $P(\mathbf{X})$ does not depend on $\boldsymbol{\theta}$ and with fixed $\mathbf{X}$, $P(\mathbf{X})$ can be viewed as a normalizing constant. Thus, a proportional form of (2.5) omits this normalizing constant

$$P(\boldsymbol{\theta}|\mathbf{X}) \propto P(\mathbf{X}|\boldsymbol{\theta})P(\boldsymbol{\theta}).$$

The term, $P(\mathbf{X}|\boldsymbol{\theta})$, when regarded as a function of $\boldsymbol{\theta}$ for fixed $\mathbf{X}$ is the familiar likelihood function used in ML estimation. As such, the data only affect the posterior distribution through the likelihood function.

By Bayes' Theorem, the posterior distribution of $(\boldsymbol{\theta}, \boldsymbol{\psi})$ may be written as

$$P(\boldsymbol{\theta}, \boldsymbol{\psi}|\mathbf{X}_{\text{obs}}, \mathbf{R}) = \frac{P(\mathbf{X}_{\text{obs}}, \mathbf{R}|\boldsymbol{\theta}, \boldsymbol{\psi})\pi(\boldsymbol{\theta}, \boldsymbol{\psi})}{P(\mathbf{X}_{\text{obs}}, \mathbf{R})}$$
$$\propto P(\mathbf{X}_{\text{obs}}, \mathbf{R}|\boldsymbol{\theta}, \boldsymbol{\psi})\pi(\boldsymbol{\theta}, \boldsymbol{\psi}) \quad (2.6)$$

where $\pi(\cdot)$ denotes a joint prior distribution applied to $(\boldsymbol{\theta}, \boldsymbol{\psi})$ and $P(\mathbf{X}_{\text{obs}}, \mathbf{R})$ is given by

$$P(\mathbf{X}_{\text{obs}}, \mathbf{R}) = \int\int P(\mathbf{X}_{\text{obs}}, \mathbf{R}|\boldsymbol{\theta}, \boldsymbol{\psi})\pi(\boldsymbol{\theta}, \boldsymbol{\psi})d\boldsymbol{\theta}d\boldsymbol{\psi}. \quad (2.7)$$

Under the MAR assumption, (2.6) becomes

$$P(\boldsymbol{\theta}, \boldsymbol{\psi}|\mathbf{X}_{\text{obs}}, \mathbf{R}) \propto P(\mathbf{R}|\mathbf{X}_{\text{obs}}, \boldsymbol{\psi})P(\mathbf{X}_{\text{obs}}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}, \boldsymbol{\psi}). \quad (2.8)$$

Moreover, $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ are distinct, the prior distribution factors as

$$\pi(\boldsymbol{\theta}, \boldsymbol{\psi}) = \pi(\boldsymbol{\theta})\,\pi(\boldsymbol{\psi})$$

and now (2.6) simplifies to

$$P(\boldsymbol{\theta}, \boldsymbol{\psi}|\mathbf{X}_{\text{obs}}, \mathbf{R}) \propto P(\mathbf{R}|\mathbf{X}_{\text{obs}}, \boldsymbol{\psi})P(\mathbf{X}_{\text{obs}}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\pi(\boldsymbol{\psi}).$$

Bayesian inferences about $\boldsymbol{\theta}$ are based on the marginal posterior obtained by integrating the function over the nuisance parameter $\boldsymbol{\psi}$ (Lee, 2007). Hence, under ignorability, the marginal posterior distribution for $\boldsymbol{\theta}$ is

$$P(\boldsymbol{\theta}|\mathbf{X}_{\text{obs}}, \mathbf{R}) = \int P(\boldsymbol{\theta}, \boldsymbol{\psi}|\mathbf{X}_{\text{obs}}, \mathbf{R})d\boldsymbol{\psi}$$
$$\propto \int P(\mathbf{R}|\mathbf{X}_{\text{obs}}, \boldsymbol{\psi})P(\mathbf{X}_{\text{obs}}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\pi(\boldsymbol{\psi})d\boldsymbol{\psi}$$
$$\propto P(\mathbf{X}_{\text{obs}}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \int P(\mathbf{R}|\mathbf{X}_{\text{obs}}, \boldsymbol{\psi})\pi(\boldsymbol{\psi})d\boldsymbol{\psi}$$
$$\propto L(\boldsymbol{\theta}|\mathbf{X}_{\text{obs}})\pi(\boldsymbol{\theta}),$$

where the proportionality is up to a multiplicative factor that does not involve $\boldsymbol{\theta}$ (Schafer, 1997). This result implies that $P(\boldsymbol{\theta}|\mathbf{X}_{\text{obs}}, \mathbf{R}) = P(\boldsymbol{\theta}|\mathbf{X}_{\text{obs}})$. Therefore, under the ignorability condition all information about $\boldsymbol{\theta}$ is summarized in the posterior that ignores the missing-data mechanism,

$$P(\boldsymbol{\theta}|\mathbf{X}_{\text{obs}}) \propto L(\boldsymbol{\theta}|\mathbf{X}_{\text{obs}})\pi(\boldsymbol{\theta}). \quad (2.9)$$

The form of (2.9) is called the observed-data posterior (Little & Rubin, 2002).

The concept of ignorability plays an important role in the construction of imputation models. Briefly, in imputing missing data we want to draw synthetic observations from the posterior distribution of the missing data, given the observed data and process that generated the missing data, that is, $P(\mathbf{X}_{\text{mis}}|\mathbf{X}_{\text{obs}}, \mathbf{R})$. Under ignorability, the posterior distribution of the missing data does not depend on $\mathbf{R}$,

$$P(\mathbf{X}_{\text{mis}}|\mathbf{X}_{\text{obs}}, \mathbf{R}) = P(\mathbf{X}_{\text{mis}}|\mathbf{X}_{\text{obs}}). \quad (2.10)$$

Assuming (2.10) holds, the implication is that

$$P(\mathbf{X}|\mathbf{X}_{\text{obs}}, \mathbf{R} = 1) = P(\mathbf{X}|\mathbf{X}_{\text{obs}}, \mathbf{R} = 0),$$

so the distribution of the data $\mathbf{X}$ is the same in the response and nonresponse groups (van Buuren & Groothuis-Oudshoorn, 2011). Most importantly, if the missing data model is ignorable we can model the posterior distribution $P(\mathbf{X}|\mathbf{X}_{\text{obs}}, \mathbf{R} = 1)$ from the observed data, and use this model to create imputations for the missing data. In other words, ignorability is the assumption that the available data $\mathbf{X}_{\text{obs}}$ are sufficient to correct for the effects of missing data $\mathbf{X}_{\text{mis}}$ (van Buuren, 2007), however, this assumption is theoretical and cannot be tested on the data itself. In many situations, however, the MAR assumption is tenable and the distinctiveness assumption is intuitive, as knowing $\boldsymbol{\theta}$ will provide little information about $\boldsymbol{\psi}$ and vice-versa (Little & Rubin, 2002; Schafer, 1997, van Buuren, 2012).

2.2. Model-Based Methods for Missing Data

The 'gold standard' model-based method for handling missing data in SEMs is full-information maximum likelihood (FIML) when all response variables are continuous and

weighted-least squares (WLS) when at least one response variable is categorical. FIML assumes the joint distribution of observed and missing variables is multivariate normal. Rather than impute missing values, FIML routines use all available data when estimating model parameters (Arbuckle, 1996; Enders, 2001). For mediation-type models, the maximum likelihood estimators with incomplete data are still obtained by maximizing the log-likelihood of the data with complete data,

$$\ell(\boldsymbol{\theta}|\boldsymbol{\nu}) = \sum_{i=1}^{n} \log f\left(\boldsymbol{\nu}_{Y_i}|\boldsymbol{\eta}_{Y_i}, \boldsymbol{\theta}_Y\right) + \sum_{i=1}^{n} \log f\left(\boldsymbol{\nu}_{M_i}|\boldsymbol{\eta}_{M_i}, \boldsymbol{\theta}_M\right) + \sum_{i=1}^{n} \log f\left(\boldsymbol{\nu}_{X_i}|\boldsymbol{\eta}_{X_i}, \boldsymbol{\theta}_X\right), \tag{2.11}$$

but now the density functions f are based on available data. Conceptually, FIML is similar to pairwise-deletion (Enders, 2001). Under the model parametrization given in (2.11), if an endogenous variable, say Y, has a missing value, then the entire observation is removed (including other variables in the log-likelihood) and the observation contributes nothing to the log-likelihood. This case is analogous to listwise deletion.

On the contrary, consider the case in which there are no missing values for an endogenous variable, Y, no missing values for a covariate X, but r missing values for a covariate M. If we want to use FIML to estimate the regression parameters of the linear model $Y = \alpha_Y + \beta_{Y \cdot X} X + \beta_{Y \cdot M} M + \zeta_Y$ based on a random sample of size n, then the log-likelihood function we would need to maximize would be of the form

$$\ell(\boldsymbol{\theta}|\boldsymbol{\nu}) = \underbrace{\sum_{i=1}^{n} \log f(y_i|x_i, \alpha_Y, \beta_{Y \cdot X})}_{\{Y,X\} \text{ observed}} + \underbrace{\sum_{i=1}^{n-r} \log f(y_i|x_i, m_i, \alpha_Y, \beta_{Y \cdot X}, \beta_{Y \cdot M})}_{\{Y,X,M\} \text{ observed}}.$$

The first term on the right-hand side of this equation is based only on the set of n

observations where $\{Y, X\}$ are observed; the second term on the right-hand side of this equation is based only on the set of $n - r$ observations where $\{Y, X, M\}$ are observed. Importantly, the regression coefficients α_Y and $\beta_{Y \cdot X}$ are estimated data from both $\{Y, X\}$ and $\{Y, X, M\}$, whereas, the regression coefficient $\beta_{Y \cdot M}$ is estimated using only data from $\{Y, X, M\}$. Hence, the MLEs are obtained using all available information. It is important to note that most implementations of FIML do not consider covariates as an explicit part of the model, therefore, if missing data are present for any covariates, the entire row is excluded from further analysis.

The WLS approach to handling missing data is similar to FIML, but does not make the assumption of multivariate normality. A common implementation of WLS to estimate parameters in the presence of missing data relies on a three-stage approach described in Muthén (1984): (1) using univariate probit regression models, estimate the thresholds of categorical response variables using ML estimation, (2) holding estimates in (1) fixed, estimate the bivariate correlations (e.g., tetrachoric, biserial, pearson, etc.) in the data using ML estimation, and (3) estimate the model parameters by minimizing a WLS objective function given by,

$$F_{WLS} = \sum_{g=1}^{G} \left(\mathbf{s}^{(g)} - \hat{\boldsymbol{\sigma}}^{(g)}\right)' \mathbf{W}^{(g)^{-1}} \left(\mathbf{s}^{(g)} - \hat{\boldsymbol{\sigma}}^{(g)}\right),$$

where g denotes the group with specific missing data pattern, $\mathbf{s}^{(g)}$ contains all relevant sample-based statistics (i.e., level 1 and level 2 estimates) for the gth group, $\hat{\boldsymbol{\sigma}}^{(g)}$ contains all relevant model-based statistics for the gth group (i.e., statistics implied by the structural equation model), and $\mathbf{W}^{(g)^{-1}}$ is a positive definite weight matrix as described in Muthén (1984). At the first two stages of estimation, only pairwise information is used. Moreover,

similar to FIML, most implementations of WLS for missing data do not consider covariates as an explicit part of the model; therefore, observations with missing data on the covariates are excluded from further analysis.

FIML is an attractive option for missing data problems in SEMs because when the observed information matrix is used, the MLEs remain consistent under the MAR mechanism (Little & Rubin, 2002) for both normal data and non-normal data (Yuan, 2009; Yuan & Bentler, 2000). However, given the asymptotic conditions needed to obtain the desirable properties of MLEs, research has shown that in smaller sample sizes (i.e., $100 \leq N \leq 200$) with a large proportion of missing data (e.g., 20%), FIML estimators have lower than nominal coverage and rejection rates (Savalei, 2010), even using robust corrections for non-normality (i.e., the sandwich estimator). To help combat the potential limitations of FIML estimators in small samples, Enders et al. (2013) showed that a nonparametric bootstrap can be used. Specifically, Enders et al. (2013) found that for under a MAR mechanism, combining FIML with a BC bootstrap routine for estimating indirect effects (based on a simple mediation model) resulted in coverage rates and empirical power estimates that reached nominal levels with sample sizes as small of $N = 100$ under specific conditions (e.g., large effects simulated). These findings highlight the advantages that bootstrapping can have in small data sets, but are limited in scope since all simulations were conducted using multivariate normal data.

Although less research has systematically examined the empirical performance of WLS for estimating parameters in the presence of missing data, the extant research shows that Muthén's (1984) three-stage WLS implementation produces estimates that are consistent and efficient when the missing data mechanism is MCAR or MAR (Asparouhov &

Muthén, 2010).

2.3. Imputation-Based Methods for Missing Data

A limitation of using model-based estimation methods to handle missing data is that as both the number of variables and missingness patterns increase in a data set, the likelihood function becomes intractable to optimize (Enders, 2010); in SEMs sparse data often result in nonpositive definite covariance matrices and inadmissible solutions (Kline, 2010; Rosseel, 2012). Again, although bootstrap may remedy situations where sparseness applies, a flexible alternative is to consider imputation-based methods to estimate parameters with missing data. The general idea behind imputation methods for missing data is that missing values are 'filled-in' with plausible values. For single imputation methods, only one complete data set is generated, for multiple imputation methods, *M* complete data sets are generated. Complete data estimation techniques (e.g., maximum likelihood) can be used on filled-in data sets to obtain parameter estimates pooled for multiple imputation (MI) methods (Enders, 2010).

2.3.1. Mean imputation. A naïve, but often used single imputation method for handling missing values is mean imputation. As shown in Algorithm (2.1), under this method for imputation the missing values for a continuous-valued variable are filled-in using the arithmetic mean of the observed values, whereas, the missing values for discrete-valued variables are filled-in using the modal value of the observed values. Although mean imputation is easy to understand and implement in practice, it can severely distort estimates and inferences, especially when data are not MCAR (Collins, Schafer, & Kam, 2001; Cook, Zeng, & Yi, 2004; Jansen, Beunckens, Molenberghs, Verbeke, & Mallinckrodt, 2006). In fact, even if one can avoid bias in parameter estimates, the mean imputation

Algorithm 2.1: Mean Imputation for Missing Data

Require: $\mathbf{X}_{n\times p} = (\mathbf{X}_{\text{obs}}, \mathbf{X}_{\text{mis}})$ matrix sorted by missingness

for $j \leftarrow 1$ to p **do**
  if any($\mathbf{x}_j$) missing
    $i_{\text{obs}} \leftarrow \{\text{indices observed data in } \mathbf{x}_j\}$
    $i_{\text{mis}} \leftarrow \{\text{indices missing data in } \mathbf{x}_j\}$
    if $\mathbf{x}_j$ is continuous
      $\mathbf{x}_{i_{\text{mis}}(j)} \leftarrow \text{mean}(\mathbf{x}_{i_{\text{obs}}(j)})$
    Else
      $\mathbf{x}_{i_{\text{mis}}(j)} \leftarrow \text{mode}(\mathbf{x}_{i_{\text{obs}}(j)})$
    end if
  end if
end for

method can still severely distort a variable's distribution by artificially reducing its variability (Allison, 2003). Given that methods such as ML estimation assume that all data are real, if some data are imputed, the imputation process introduces additional sampling variability that is not adequately accounted for. Importantly, artificially reducing variability can lead to underestimated standard errors and higher rates of Type I error (Horton & Kleinman, 2007). Therefore, unless a very small percentage of values are missing, best practice is to avoid mean imputation altogether (and other single imputation methods) in favor of more flexible (MI) methods (Schafer & Graham, 2002).

2.3.2. Basic theory of multiple imputation. The goal of MI is to find an estimate $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ that is both unbiased and confidence valid in the presence in missing data (Rubin, 1996). The unbiased condition implies that

$$E(\widehat{\boldsymbol{\theta}}|\mathbf{X}) = \boldsymbol{\theta}. \tag{2.12}$$

Let $\mathbf{U}$ be the estimated variance-covariance matrix of $\widehat{\boldsymbol{\theta}}$, then the confidence valid condition implies that

$$E(\mathbf{U}|\mathbf{X}) \geq Var(\widehat{\boldsymbol{\theta}}|\mathbf{X}), \qquad (2.13)$$

where $Var(\widehat{\boldsymbol{\theta}}|\mathbf{X})$ is the variance attributable to sampling error. The confidence valid condition states that the average of $\mathbf{U}$ over all possible samples is at least as large as the sampling variance of $\widehat{\boldsymbol{\theta}}$. As applied to hypothesis testing, the goal of multiple imputation is to obtain unbiased estimates of $\boldsymbol{\theta}$ (2.12) with associated confidence intervals and hypothesis tests that should achieve at least the nominal rejection rates (2.13) (Rubin, 1987, 1996).

As compared to single imputation, MI is a more robust and flexible approach for the analysis of incomplete data sets. A statistical analysis using MI involves three general steps: (1) generating M plausible imputations for missing values (based implicitly or explicitly from a joint model), (2) analyzing each of the M pseudo-complete data sets as if the data sets were complete (i.e., contained no missing values), and (3) pooling (or combining) the M set of parameter estimates together to obtain aggregate parameter estimates based on the multiply imputed data. The two widely used frameworks for generating imputations are using an explicit joint model (JM) or an implicit joint model estimating through a series of conditionally specified models, called fully conditionally specification (FCS).

The former joint modeling framework follows the theoretical framework pioneered by Rubin (1976). Let $\mathbf{X}$ denote a data matrix with partially observed variables and $\mathbf{R}$ be the missing data indicator matrix of $\mathbf{X}$ as defined in (2.1). In the JM approach, imputations are generated using three steps: (1) modeling; specify joint distribution of all the data, $P(\mathbf{X}, \mathbf{R})$, (2) imputation; derive the posterior predictive distribution of the missing values given the observed data, $P(\mathbf{X}_{\text{mis}}|\mathbf{X}_{\text{obs}}, \mathbf{R})$, and (3) estimation; calculate the posterior distribution of the parameters $\boldsymbol{\theta}$ so that random draws can be made from it (van Buuren, 2007).

In generating imputations under the JM, a repeated imputation procedure is used (Li, Raghunathan, & Rubin, 1991). Repeated imputations are draws from the posterior predictive distribution of missing values under a Bayesian model that accounts for both the observed data and missing data mechanism (Little & Rubin, 2002; Rubin, 1996). After M set of imputations are generated under an imputation model (often a Bayesian model), the set of M complete data sets are analyzed using frequentist-based methods (e.g., ML estimation in SEMs) as if they were complete. According to Rubin (1996), MI was designed to use Bayesian methods to create imputations and frequentist methods to evaluate procedures. The M set of k parameter estimates from the complete data sets $\widehat{\boldsymbol{\theta}}_1, \dots, \widehat{\boldsymbol{\theta}}_m$, where $\widehat{\boldsymbol{\theta}}_i = \left[\hat{\theta}^{i1}, \dots, \hat{\theta}^{ik}\right]$ and $i = 1, \dots, M$ are combined to form one repeated-imputation inference that approximately adjusts for the missing data mechanism under the model used to create the imputations (Zhang, 2003). For each $\widehat{\boldsymbol{\theta}}_i$, there is an estimated variance-covariance matrix $\mathbf{U}_i$, resulting in a set of M estimated variance-covariance matrices.

The key Bayesian motivation for MI is based on the posterior distribution of $\boldsymbol{\theta}$ and its first two moments. The posterior distribution of $\boldsymbol{\theta}$ is average complete-data posterior distribution of $\boldsymbol{\theta}$,

$$P(\boldsymbol{\theta}|\mathbf{X}_{\text{obs}}) = \int P(\boldsymbol{\theta}|\mathbf{X}_{\text{obs}}, \mathbf{X}_{\text{mis}}) P(\mathbf{X}_{\text{mis}}|\mathbf{X}_{\text{obs}}) d\mathbf{X}_{\text{mis}}. \qquad (2.14)$$

In (2.14), the average is over the repeated imputations, which are draws from the posterior predictive distribution of the missing data given the observed data, $P(\mathbf{X}_{\text{mis}}|\mathbf{X}_{\text{obs}})$. A more intuitive interpretation of (2.14) is done from right to left, which provides the intuition about the JM approach to MI (van Buuren, 2012). Specifically, $P(\mathbf{X}_{\text{mis}}|\mathbf{X}_{\text{obs}})$ is used to draw imputations of $\mathbf{X}_{\text{mis}}$, denoted as $\dot{\mathbf{X}}_{\text{mis}}$ (the imputation step, which uses an imputation

model). Then, $P(\boldsymbol{\theta}|\mathbf{X}_{\text{obs}}, \dot{\mathbf{X}}_{\text{mis}})$ can be used to calculate the parameters of interest from the hypothetically complete data $(\mathbf{X}_{\text{obs}}, \dot{\mathbf{X}}_{\text{mis}})$ and make random draws (the posterior step). This two-step procedure is iterated until specified convergence criteria are met. Thus, Equation (2.14) says that the actual posterior distribution of $\boldsymbol{\theta}$ is equal to the average over the repeated draws of $\boldsymbol{\theta}$.

From (2.14), the mean of the posterior distribution of $\boldsymbol{\theta}$ is

$$E(\boldsymbol{\theta}|\mathbf{X}_{\text{obs}}) = E[E(\boldsymbol{\theta}|\mathbf{X}_{\text{obs}}, \mathbf{X}_{\text{mis}})|\mathbf{X}_{\text{obs}}], \tag{2.15}$$

which is the average of the repeated complete-data posterior means of $\boldsymbol{\theta}$. Similarly, the variance of the posterior distribution of $\boldsymbol{\theta}$ is

$$\begin{aligned} Var(\boldsymbol{\theta}|\mathbf{X}_{\text{obs}}) = {} & E[Var(\boldsymbol{\theta}|\mathbf{X}_{\text{obs}}, \mathbf{X}_{\text{mis}})|\mathbf{X}_{\text{obs}}] \\ & + Var[E(\boldsymbol{\theta}|\mathbf{X}_{\text{obs}}, \mathbf{X}_{\text{mis}})|\mathbf{X}_{\text{obs}}], \end{aligned} \tag{2.16}$$

which is the sum of the average repeated complete-data variances of $\boldsymbol{\theta}$ and the variance of the repeated complete-data posterior means of $\boldsymbol{\theta}$. Equations (2.14) – (2.16) lead to Rubin's (1996) so-called 'rules' for combining repeated imputations. The formulas are based on the empirical estimates of the mean and variance of the posterior distribution. Specifically, with M multiple imputations the empirical estimate of the posterior mean of $\boldsymbol{\theta}$ is given by

$$\bar{\bar{\boldsymbol{\theta}}}_M = \frac{1}{M}\sum_{j=1}^{M} \hat{\boldsymbol{\theta}}_j. \tag{2.17}$$

The empirical variance estimate of the posterior variance of $\boldsymbol{\theta}$ is given by

$$\begin{aligned} Var\left(\bar{\bar{\boldsymbol{\theta}}}_M\right) &= \bar{\mathbf{U}}_M + \mathbf{B}_M + \mathbf{B}_M/M \\ &= \bar{\mathbf{U}}_M + \frac{M+1}{M}\mathbf{B}_M \end{aligned} \tag{2.18}$$

where

$$\bar{\mathbf{U}}_M = \frac{1}{M}\sum_{j=1}^{M}\mathbf{U}_j \tag{2.19}$$

is the within imputation variability and

$$\mathbf{B}_M = \frac{1}{M-1}\sum_{j=1}^{M}\left(\hat{\boldsymbol{\theta}}_j - \bar{\hat{\boldsymbol{\theta}}}_M\right)\left(\hat{\boldsymbol{\theta}}_j - \bar{\hat{\boldsymbol{\theta}}}_M\right)' \tag{2.20}$$

is the between imputation variability. As can be seen in (2.18), the total variability of $\bar{\hat{\boldsymbol{\theta}}}_M$ is the sum of three components: (1) $\bar{\mathbf{U}}_M$, sampling error (2.19), (2) $\mathbf{B}_M$, extra variance due to the missing data (2.20), and (3) $\mathbf{B}_M/M$, extra simulation variance due to $\bar{\hat{\boldsymbol{\theta}}}_M$ being estimated by finite M (van Buuren, 2012). In practice, the effect of the latter term can be attenuated with larger values of M. Often, $Var\left(\bar{\hat{\boldsymbol{\theta}}}_M\right)$ denoted as $\mathbf{T}_M$ to denote the total posterior variance of $\boldsymbol{\theta}$.

In terms of inference using MI, for large M, the distribution of the multiply imputed estimate $\left(\boldsymbol{\theta} - \bar{\hat{\boldsymbol{\theta}}}_M\right)$ can be approximated as a normal random variable with covariance matrix $\mathbf{T}_M$, which is based off the asymptotic distribution letting $M \to \infty$ as

$$\left(\boldsymbol{\theta} - \bar{\hat{\boldsymbol{\theta}}}_\infty\right) \sim N(\mathbf{0}, \mathbf{T}_\infty),$$

where $\mathbf{T}_\infty = \bar{\mathbf{U}}_\infty + \mathbf{B}_\infty$ (Meng & Rubin, 1993; Wang & Robins, 1998), since the term $\mathbf{B}_M/M$ goes to zero as $M \to \infty$. In practice, $\mathbf{T}$ is not known a priori so it is estimated using the multiply imputed estimator $\mathbf{T}_M$, which is a function of the number of imputations, M. Rubin (1987, p. 114) showed that the asymptotic variance $\mathbf{T}_\infty$ and the finite variance $\mathbf{T}_M$ are related by

$$\mathbf{T}_M = \left(1 + \frac{\gamma}{M}\right)\mathbf{T}_\infty,$$

where γ is the fraction of missing information (FMI), which is a variance ratio measure of

the proportion of variation that can be attributed to the missing data. FMI is defined as

$$\gamma = \frac{\left(\left(\frac{\mathbf{B}_M + \mathbf{B}_M/M}{\bar{\mathbf{U}}_M}\right) + 2\right)/(df + 3)}{1 + \left(\frac{\mathbf{B}_M + \mathbf{B}_M/M}{\bar{\mathbf{U}}_M}\right)}, \tag{2.21}$$

where df are the finite, sample-adjusted degrees of freedom given by

$$df = \frac{\vartheta\varphi}{\vartheta + \varphi}, \tag{2.22}$$

where

$$\vartheta = (M-1)\left(1 + \left(\frac{\mathbf{B}_M + \mathbf{B}_M/M}{\bar{\mathbf{U}}_M}\right)^{-2}\right)$$

and

$$\varphi = \frac{n-k+1}{n-k+3}(n-k)\left(1 - \frac{\mathbf{B}_M + \mathbf{B}_M/M}{\mathbf{T}_M}\right)$$

and k is the number of parameters fit to the data.

For finite M, confidence intervals for the jth element of $\bar{\bar{\boldsymbol{\theta}}}_M$ can be constructed using a t-distribution as

$$\bar{\bar{\boldsymbol{\theta}}}_{M,j} \pm t_{\nu,1-\alpha/2}\sqrt{\mathbf{T}_{M,j}},$$

where $t_{df,1-\alpha/2}$ is the quantile of the t-distribution with df degrees of freedom (defined in [2.22]) corresponding to probability $1-\alpha/2$ and $\mathbf{T}_{M,j} = Var\left(\bar{\bar{\boldsymbol{\theta}}}_M\right)_{jj}$ is the jth diagonal component of the estimated covariance matrix of $\bar{\bar{\boldsymbol{\theta}}}_M$.

2.3.3. Proper imputations. Imputations procedures that lead to valid statistical inferences are said to be proper. Although there are varying degrees of proper in terms of nominal confidence interval coverage, we use a less stringent type referred to as

confidence proper. In a missing data context, an imputation procedure is said to be confidence proper if for large M, three conditions hold. First, $\bar{\bar{\boldsymbol{\theta}}}_M$ is an unbiased estimate of $\hat{\boldsymbol{\theta}}$,

$$E\left(\bar{\bar{\boldsymbol{\theta}}}_M|\mathbf{X}\right) = \hat{\boldsymbol{\theta}}. \tag{2.23}$$

Second, $\bar{\mathbf{U}}_M$ is an unbiased estimate of $\mathbf{U}$,

$$E(\bar{\mathbf{U}}_M|\mathbf{X}) = \mathbf{U}. \tag{2.24}$$

Third, similar to (2.13), the average estimate $\mathbf{B}$ of the variance due to missing data (i.e., the extra inferential uncertainty about $\bar{\bar{\boldsymbol{\theta}}}_M$ due to missing data; van Buuren, 2012) should be at least as large as the variance observed in the MI estimator $\bar{\bar{\boldsymbol{\theta}}}_M$,

$$\frac{M+1}{M}E(\mathbf{B}_M|\mathbf{X}) \geq Var\left(\bar{\bar{\boldsymbol{\theta}}}_M\right). \tag{2.25}$$

Empirically, the most common way to determine if an imputation procedure is confidence proper is through statistical simulation. Briefly, population parameters and models are chosen, samples are generated based on these models, incomplete data are generated, the imputation procedure is implemented, estimates of the population parameters are calculated, and then results are averaged over iterations. Under such a simulation design, confidence proper imputation routines provide unbiased estimates (i.e., satisfy [2.23] and [2.24]) and provide nominal confidence interval coverage (i.e., satisfy [2.25]).

2.3.4. Joint model multiple imputation. A widely used JM MI algorithm was developed by Schafer (1997) and follows from Rubin's theory of MI, specifically using a Bayesian imputation model. As opposed to the FCS approach to MI, the JM approach explicitly models the joint distribution of $P(\mathbf{X}, \mathbf{R})$ using a multivariate normal distribution. Compared to frequentist modeling, Bayesian modeling explicitly allows parameter uncertainty (or the

lack thereof) to be incorporated by the choice of prior distribution for $\boldsymbol{\theta}$ (Song & Lee, 2012). Often, little or no prior information is known about $\boldsymbol{\theta}$ and a non-informative prior can be used for $\boldsymbol{\theta}$. Two common choices of non-informative priors are those that are proportional to a constant and Jeffrey's prior. In the former case, the prior for $\boldsymbol{\theta}$ is of the form

$$P(\boldsymbol{\theta}) \propto C,$$

where C is a constant that does not depend on $\boldsymbol{\theta}$. Under a non-informative prior of this type, we 'let the data speak for itself' and the posterior distribution takes the form

$$P(\boldsymbol{\theta}|\mathbf{X}) \propto P(\mathbf{X}|\boldsymbol{\theta}) \times C \tag{2.26}$$

In the case of (2.26), the posterior distribution is proportional to the likelihood function and Bayesian estimates coincide with MLEs (Gelman et al., 2013). A more theoretically sound non-informative prior is Jeffrey's prior. Jeffrey's prior uses the prior density of the form,

$$P(\boldsymbol{\theta}) \propto |\mathbf{I}(\boldsymbol{\theta})|^{1/2}, \tag{2.27}$$

where $\mathbf{I}(\boldsymbol{\theta})$ is Fisher's (expected) information matrix (see Chapter 1 for a review). Jeffreys (1961) argued that in the case of multiparameter distributions, the prior in (2.27) should be applied to each parameter separately, assuming the other parameters are known constants. With respect to MI, Schafer's JM MI algorithm uses Jeffrey's prior.

The underlying algorithm of the JM approach to MI uses Tanner and Wong's (1987) data augmentation (DA) algorithm to impute missing values based on an underlying multivariate normal distribution for the data. The multivariate normal density is presented in Definition (2.1).

Definition 2.1. *Let* $\mathbf{x}$ be a $p \times 1$ *random vector in* $\mathcal{R}^p$ *with mean vector* $\boldsymbol{\mu}$ *in* $\mathcal{R}^p$ *and covariance matrix* $\boldsymbol{\Sigma}$ *in* $\mathcal{R}^{p \times p}$. *Then* $\mathbf{x}$ *is said to follow a multivariate normal distribution if for* $|\boldsymbol{\Sigma}| > 0$,

$$f(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right).$$

A convenient shorthand notation for Definition (2.1) is $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ or simply $\mathbf{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Note, when $p = 1$, the multivariate normal distribution reduces to the univariate normal distribution. Before introducing the JM MI algorithm in detail, we first describe some important results regarding partitioned matrices and properties associated with the multivariate normal distribution.

Lemma 2.1. *Let* $\boldsymbol{\Sigma}$ *be a* $p \times p$ *symmetric matrix partitioned as*

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix},$$

where $\boldsymbol{\Sigma}_{11}$ *is a* $p_1 \times p_1$ *matrix,* $\boldsymbol{\Sigma}_{22}$ *is a* $p_2 \times p_2$ *matrix,* $\boldsymbol{\Sigma}_{12}$ *is a* $p_1 \times p_2$ *matrix,* $\boldsymbol{\Sigma}_{21} = \boldsymbol{\Sigma}_{12}'$, $|\boldsymbol{\Sigma}_{11}| > 0$, *and* $|\boldsymbol{\Sigma}_{22}| > 0$. *Then, if* $|\boldsymbol{\Sigma}_{22}| \neq 0$, *the determinant of* $\boldsymbol{\Sigma}$ *can be expressed as*

$$|\boldsymbol{\Sigma}| = \left|\begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}\right| = |\boldsymbol{\Sigma}_{22}||\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}|.$$

Equivalently, if $|\boldsymbol{\Sigma}_{11}| \neq 0$, *then the determinant of* $\boldsymbol{\Sigma}$ *can be expressed as*

$$|\boldsymbol{\Sigma}| = \left|\begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}\right| = |\boldsymbol{\Sigma}_{11}||\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}|.$$

In addition, the inverse of $\boldsymbol{\Sigma}$ *is*

$$\boldsymbol{\Sigma}^{-1} = \begin{bmatrix} \boldsymbol{\Sigma}^{11} & \boldsymbol{\Sigma}^{12} \\ \boldsymbol{\Sigma}^{21} & \boldsymbol{\Sigma}^{22} \end{bmatrix},$$

where

$$\boldsymbol{\Sigma}^{11} = (\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21})^{-1}$$

$$\boldsymbol{\Sigma}_{22} = (\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})^{-1}$$

$$\boldsymbol{\Sigma}_{21} = -\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}(\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21})^{-1}$$

$$\boldsymbol{\Sigma}_{12} = -\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}(\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})^{-1}$$

and $\boldsymbol{\Sigma}^{11}$ *and* $\boldsymbol{\Sigma}^{22}$ *are the Schur complements.*

Proof:

See Appendix A.

Lemma (2.2) shows another useful form of the inverse expression along the main diagonal of a symmetric partitioned matrix.

Lemma 2.2. *Let* $\mathbf{A}, \mathbf{B}, \mathbf{C}$, *and* $\mathbf{D}$ *be matrices that conform in addition and multiplication where both* $\mathbf{A}$ *and* $\mathbf{B}$ *are square matrices with* $|\mathbf{A}| \neq 0$ *and* $|\mathbf{B}| \neq 0$, *then the inverse of* $(\mathbf{A} - \mathbf{CBD})^{-1}$ *is*

$$(\mathbf{A} - \mathbf{CBD})^{-1} = \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{C}(\mathbf{B}^{-1} - \mathbf{DA}^{-1}\mathbf{C})^{-1}\mathbf{DA}^{-1}.$$

Proof:

To prove equality, the product $(\mathbf{A} - \mathbf{CBD})(\mathbf{A} - \mathbf{CBD})^{-1}$ should be the identity matrix. To show this,

$$\begin{aligned}(\mathbf{A} - \mathbf{CBD})(\mathbf{A} - \mathbf{CBD})^{-1} &= (\mathbf{A} - \mathbf{CBD})(\mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{C}(\mathbf{B}^{-1} - \mathbf{DA}^{-1}\mathbf{C})^{-1}\mathbf{DA}^{-1}) \\ &= (\mathbf{A} - \mathbf{CBD})\mathbf{A}^{-1} + (\mathbf{A} - \mathbf{CBD})\mathbf{A}^{-1}\mathbf{C}(\mathbf{B}^{-1} - \mathbf{DA}^{-1}\mathbf{C})^{-1}\mathbf{DA}^{-1} \\ &= \mathbf{I} - \mathbf{CBDA}^{-1} + (\mathbf{C} - \mathbf{CBDA}^{-1}\mathbf{C})(\mathbf{B}^{-1} - \mathbf{DA}^{-1}\mathbf{C})^{-1}\mathbf{DA}^{-1} \\ &= \mathbf{I} - \mathbf{CBDA}^{-1} + \mathbf{CB}(\mathbf{B}^{-1} - \mathbf{DA}^{-1}\mathbf{C})(\mathbf{B}^{-1} - \mathbf{DA}^{-1}\mathbf{C})^{-1}\mathbf{DA}^{-1} \\ &= \mathbf{I} - \mathbf{CBDA}^{-1} + \mathbf{CBDA}^{-1}\end{aligned}$$

$$= \mathbf{I},$$

which completes the proof. ■

The next Lemmas provides import results for derivations used in Schafer's JM MI algorithm. Let $\mathbf{x}$ be a $p \times 1$ normally distributed random vector with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Let $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2]'$ be a partition with subvector lengths p_1 and p_2 for $\mathbf{x}_1$ and $\mathbf{x}_2$, respectively, where $p_1 + p_2 = p$. According to the partition described, the mean vector $\boldsymbol{\mu} = E(\mathbf{x})$ and covariance matrix $\boldsymbol{\Sigma} = Cov(\mathbf{x}_i, \mathbf{x}_j), i, j = 1,2$ can be expressed as

$$\boldsymbol{\mu} = \begin{bmatrix} E(\mathbf{x}_1) \\ E(\mathbf{x}_2) \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}$$

and for $|\boldsymbol{\Sigma}_{11}| > \mathbf{0}$ and $|\boldsymbol{\Sigma}_{22}| > 0$,

$$\boldsymbol{\Sigma} = \begin{bmatrix} Cov(\mathbf{x}_1, \mathbf{x}_1) & Cov(\mathbf{x}_1, \mathbf{x}_2) \\ Cov(\mathbf{x}_2, \mathbf{x}_1) & Cov(\mathbf{x}_2, \mathbf{x}_2) \end{bmatrix}$$

$$= \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix},$$

where $\boldsymbol{\mu}_1$ is a $p_1 \times 1$ vector, $\boldsymbol{\mu}_2$ is a $p_2 \times 1$ vector, $\boldsymbol{\Sigma}_{11}$ is a $p_1 \times p_1$ matrix, $\boldsymbol{\Sigma}_{22}$ is a $p_2 \times p_2$ matrix, $\boldsymbol{\Sigma}_{12}$ is a $p_1 \times p_2$ matrix, and $\boldsymbol{\Sigma}_{21} = \boldsymbol{\Sigma}_{12}'$. The first Lemma gives results of the marginal distributions of the partitioned multivariate normal random vector $\mathbf{x}$.

Lemma 2.3. *The marginal distributions of* $\mathbf{x}_1$ and $\mathbf{x}_2$ *are normal with mean vector with mean vector* $\boldsymbol{\mu}_i, i = 1,2$, *and covariance matrix* $\boldsymbol{\Sigma}_{ii}$, i $= 1,2$.

Proof:

See Appendix A.

Using Lemma (2.3), we can now derive the conditional distributions of $\mathbf{x}_1|\mathbf{x}_2$ or equivalently $\mathbf{x}_2|\mathbf{x}_1$.

Theorem 2.1. *The conditional distribution of* $\mathbf{x}_i$ given $\mathbf{x}_j$ *is normal with mean vector*

$$\boldsymbol{\mu}_{i|j} = \boldsymbol{\mu}_i + \boldsymbol{\Sigma}_{ij}\boldsymbol{\Sigma}_{jj}^{-1}(\mathbf{x}_j - \boldsymbol{\mu}_j)$$

and covariance matrix

$$\boldsymbol{\Sigma}_{i|j} = \boldsymbol{\Sigma}_{jj} - \boldsymbol{\Sigma}_{ji}\boldsymbol{\Sigma}_{ii}^{-1}\boldsymbol{\Sigma}_{ij}.$$

Proof:

See Appendix A.

For a multivariate normal distribution based on a sample of n independent observations, the likelihood function is given by

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{X}) = -\frac{p}{2}\log(2\pi) - \frac{n}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu})$$

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{X}) \propto -\frac{n}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu})$$

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{X}) = -\frac{n}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Sigma}^{-1}\left(\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})'\right)\right].$$

The maximum likelihood estimators of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are obtained by differentiating $L(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{X})$. First, to calculate $\frac{\partial}{\partial \boldsymbol{\mu}}L(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{X})$ (Press, 1972) we treat $\boldsymbol{\Sigma}$ as a constant,

$$\frac{\partial}{\partial \boldsymbol{\mu}}L(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{X}) = \frac{\partial}{\partial \boldsymbol{\mu}}\left[-\frac{n}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu})\right]$$

$$= -\frac{1}{2}\sum_{i=1}^{n}\frac{\partial}{\partial \boldsymbol{\mu}}[\mathbf{x}_i'\boldsymbol{\Sigma}^{-1}\mathbf{x}_i - 2\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\mathbf{x}_i + \boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}]$$

$$= -\frac{1}{2}\sum_{i=1}^{n}[-2\boldsymbol{\Sigma}^{-1}\mathbf{x}_i + 2\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}]$$

$$= \boldsymbol{\Sigma}^{-1}\left[\sum_{i=1}^{n}\mathbf{x}_i - n\boldsymbol{\mu}\right].$$

Solving $\boldsymbol{\Sigma}^{-1}[\sum_{i=}^{n}\mathbf{x}_i - n\boldsymbol{\mu}] = \mathbf{0}$, we easily see that the ML estimator $\hat{\boldsymbol{\mu}}_{\text{MLE}}$ of $\boldsymbol{\mu}$ is

$$\hat{\boldsymbol{\mu}}_{\text{MLE}} = n^{-1}\sum_{i=1}^{n}\mathbf{x}_i,$$

which is just the sample mean vector $\bar{\mathbf{x}} = [\bar{\mathbf{x}}_1, \ldots, \bar{\mathbf{x}}_p]'$. Next, to calculate $\frac{\partial}{\partial \boldsymbol{\Sigma}}L(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{x})$ (Magnus & Neudecker, 1999; Wand, 2002) we evaluate $L(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{X})$ at the MLE of $\hat{\boldsymbol{\mu}}_{\text{MLE}}$,

$$\frac{\partial}{\partial \boldsymbol{\Sigma}}L(\boldsymbol{\mu} = \hat{\boldsymbol{\mu}}_{\text{MLE}}, \boldsymbol{\Sigma}|\mathbf{X}) = \frac{\partial}{\partial \boldsymbol{\Sigma}}\left[-\frac{n}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\text{tr}\left[\boldsymbol{\Sigma}^{-1}\left(\sum_{i=1}^{n}(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_{\text{MLE}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_{\text{MLE}})'\right)\right]\right]$$

$$= -\frac{n}{2}\boldsymbol{\Sigma}^{-1} + \frac{1}{2}\boldsymbol{\Sigma}^{-1}\sum_{i=1}^{n}(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_{\text{MLE}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_{\text{MLE}})'\boldsymbol{\Sigma}^{-1}.$$

Solving $-\frac{n}{2}\boldsymbol{\Sigma}^{-1} + \frac{1}{2}\boldsymbol{\Sigma}^{-1}\sum_{i=1}^{n}(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_{\text{MLE}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_{\text{MLE}})'\boldsymbol{\Sigma}^{-1} = \mathbf{0}$, we obtain the ML estimator $\hat{\boldsymbol{\Sigma}}_{\text{MLE}}$ of $\boldsymbol{\Sigma}$ as

$$\hat{\boldsymbol{\Sigma}}_{\text{MLE}} = n^{-1}\sum_{i=1}^{n}(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_{\text{MLE}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_{\text{MLE}})',$$

which is the (biased) sample covariance matrix **S**. Technically, a more formal treatment of maximum likelihood estimators would check the second derivative conditions to ensure that global maximum estimators are obtained (see Press, 1972 for a more detailed discussion).

An alternative form of the multivariate normal likelihood can be constructed by adding and subtracting the sample mean to kernel function in the exponent (Johnson & Wischern, 2001) as $\sum_{i=1}^{n}(\mathbf{x}_i + \bar{\mathbf{x}} - \bar{\mathbf{x}} - \boldsymbol{\mu})(\mathbf{x}_i + \bar{\mathbf{x}} - \bar{\mathbf{x}} - \boldsymbol{\mu})'$ and expanding the product as

$$\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})' + 2(\bar{\mathbf{x}} - \boldsymbol{\mu})'\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}}) + \sum_{i=1}^{n}(\bar{\mathbf{x}} - \boldsymbol{\mu})(\bar{\mathbf{x}} - \boldsymbol{\mu})'. \tag{2.28}$$

Here, the middle term in (2.28) goes to zero and the last term is constant for all $i = 1, \dots, n$. Therefore, (2.28) becomes

$$\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})' + n(\bar{\mathbf{x}} - \boldsymbol{\mu})(\bar{\mathbf{x}} - \boldsymbol{\mu})'$$

or equivalently,

$$n\mathbf{S} + n(\bar{\mathbf{x}} - \boldsymbol{\mu})(\bar{\mathbf{x}} - \boldsymbol{\mu})',$$

where $\mathbf{S} = n^{-1}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'$ is the sample covariance matrix. Rewriting $L(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{X})$, the likelihood becomes

$$\begin{aligned} L(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{X}) &= |\boldsymbol{\Sigma}|^{-n/2}\exp\left\{-\frac{1}{2}\mathrm{tr}[\boldsymbol{\Sigma}^{-1}(n\mathbf{S} + n(\bar{\mathbf{x}} - \boldsymbol{\mu})(\bar{\mathbf{x}} - \boldsymbol{\mu})')]\right\} \\ &= |\boldsymbol{\Sigma}|^{-n/2}\exp\left\{-\frac{n}{2}\mathrm{tr}[\boldsymbol{\Sigma}^{-1}\mathbf{S}]\right\} \times \exp\left\{-\frac{n}{2}\mathrm{tr}[\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu})(\bar{\mathbf{x}} - \boldsymbol{\mu})']\right\} \\ &= |\boldsymbol{\Sigma}|^{-n/2}\exp\left\{-\frac{n}{2}\mathrm{tr}[\boldsymbol{\Sigma}^{-1}\mathbf{S}]\right\} \times \exp\left\{-\frac{n}{2}(\bar{\mathbf{x}} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu})\right\}. \end{aligned}$$

To calculate the posterior distribution of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ given the data $\mathbf{x}$, Schafer (1997) uses Jeffrey's prior for $P(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

$$P(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = |\boldsymbol{\Sigma}|^{-(p+1)/2}, \tag{2.29}$$

which is the limiting form of the conjugate normal-inverted Wishart density. If we assume that $P(\boldsymbol{\mu}|\boldsymbol{\Sigma}) \sim N(\boldsymbol{\mu}_0, \tau^{-1}\boldsymbol{\Sigma})$ where $\boldsymbol{\mu}_0 \in \mathcal{R}^p$ and $\tau > 0$ and $P(\boldsymbol{\Sigma}) \sim IW(m, \boldsymbol{\Lambda})$, where IW denotes inverted-Wishart and $m \geq p$ and $\boldsymbol{\Lambda} > \mathbf{0}$, then the normal-IW density for $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is

proportional to

$$P(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-(m+p+2)/2} \exp\left\{-\frac{1}{2}\mathrm{tr}[\boldsymbol{\Lambda}^{-1}\boldsymbol{\Sigma}^{-1}]\right\} \times \exp\left\{-\frac{\tau}{2}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)\right\}. \tag{2.30}$$

Then, the limiting form of (2.30) leads to Jeffrey's prior in (2.29) as $(\tau, m, \boldsymbol{\Lambda}^{-1}) \to (0, -1, \mathbf{0})$ (Timm, 2002). Combing the reparametrized likelihood with the prior in (2.29), the posterior distribution of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ takes the form

$$\begin{aligned} P(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{X}) &\propto L(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{X})P(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ &= \exp\left\{-\tfrac{n}{2}\mathrm{tr}[\boldsymbol{\Sigma}^{-1}\mathbf{S}]\right\} \exp\left\{-\tfrac{n}{2}(\bar{\mathbf{x}}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}-\boldsymbol{\mu})\right\} \times |\boldsymbol{\Sigma}|^{-(p+1)/2}|\boldsymbol{\Sigma}|^{-n/2} \\ &= |\boldsymbol{\Sigma}|^{-(n+p+1)/2} \exp\left\{-\frac{n}{2}\mathrm{tr}[\boldsymbol{\Sigma}^{-1}\mathbf{S}]\right\} \times \exp\left\{-\frac{n}{2}(\bar{\mathbf{x}}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}-\boldsymbol{\mu})\right\}. \end{aligned}$$

Here, the posterior distribution factors into the product of two distributions, $P(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{X}) \propto P(\boldsymbol{\Sigma}|\mathbf{X})P(\boldsymbol{\mu}|\boldsymbol{\Sigma}, \mathbf{X})$ where $P(\boldsymbol{\Sigma}|\mathbf{X}) \sim IW(n-1, (n\mathbf{S})^{-1})$ and $P(\boldsymbol{\mu}|\boldsymbol{\Sigma}, \mathbf{X}) \sim N(\bar{\mathbf{x}}, n^{-1}\boldsymbol{\Sigma})$.

The DA algorithm is the stochastic analog of the expectation-maximization (EM) algorithm (see Dempster, Laird, & Rubin, 1977), iterating between imputation steps (I-step) and posterior steps (P-step) using a Gibbs-sampler routine. Specifically, the I-step simulates

$$\dot{\mathbf{x}}_{\mathrm{mis}}^{(t+1)} = P\left(\mathbf{x}_{\mathrm{mis}}|\mathbf{x}_{\mathrm{obs}}, \dot{\boldsymbol{\theta}}^{(t)}\right),$$

which are independent simulations of random normal vectors for each row of the data matrix, with means and covariances given by the multivariate normal conditional distribution of $\mathbf{x}_{\mathrm{mis}}$ on $\mathbf{x}_{\mathrm{obs}}$ and $\dot{\boldsymbol{\theta}}^{(t)}$ (see Theorem [2.1]). The P-step simulates

$$\dot{\boldsymbol{\theta}}^{(t+1)} = P\left(\boldsymbol{\theta}|\mathbf{x}_{\mathrm{obs}}, \dot{\mathbf{x}}_{\mathrm{mis}}^{(t+1)}\right).$$

which are draws from the normal-inverted Wishart distribution (i.e., the distribution of the parameters conditioned on the 'pseudo'-complete data). The I- and P-steps alternate for thousands of iterations until convergence to the stationary posterior distribution.

Convergence as defined here is in the context of Markov Chains. In particular, the output of the DA is a sequence $\left\{\left(\boldsymbol{\theta}^{(t)}, \mathbf{x}_{\text{mis}}^{(t)}\right): t = 0,1,2, \ldots\right\}$, where at iteration 0, $\boldsymbol{\theta}^{(t)}$ is initialized by using only observed data estimates based on listwise/pairwise deletion or the EM algorithm. The goal of the DA algorithm is to simulate draws from the distribution $P(\boldsymbol{\theta}, \mathbf{x}_{\text{mis}}|\mathbf{x}_{\text{obs}})$. For the sequence to have converged, it is sufficient for the distribution of $\boldsymbol{\theta}^{(t)}$ to have converged to $P(\boldsymbol{\theta}|\mathbf{x}_{\text{obs}})$, because $\boldsymbol{\theta}^{(t)} \sim P(\boldsymbol{\theta}|\mathbf{x}_{\text{obs}})$ implies that $\left(\boldsymbol{\theta}^{(s+t)}, \mathbf{x}_{\text{mis}}^{(s+t)}\right) \sim P(\boldsymbol{\theta}, \mathbf{x}_{\text{mis}}|\mathbf{x}_{\text{obs}})$ for all $s > 0$. Equivalently, it is sufficient for the distribution of $\mathbf{x}_{\text{mis}}^{(t)}$ to have converged to $P(\mathbf{x}_{\text{mis}}|\mathbf{x}_{\text{obs}})$. Also, convergence by t iterations means that $\boldsymbol{\theta}^{(s)}$ and $\mathbf{x}_{\text{mis}}^{(s)}$ are independent of $\boldsymbol{\theta}^{(s+t)}$ and $\mathbf{x}_{\text{mis}}^{(s+t)}$. In practice, convergence is empirically monitored through the successive values of $\boldsymbol{\theta}$ rather than successive values of $\mathbf{x}_{\text{mis}}$ because the latter is usually in high dimension.

Schafer's JM MI algorithm requires repeated applications of the conditional distributions for the multivariate normal distribution. To simplify the algorithm, a device known as the sweep operator is used. The sweep operator is widely used in linear model computations and provides a useful means of computing MLEs in multivariate missing data problems (Goodnight, 1979). Suppose that $\mathbf{G}$ is a symmetric $p \times p$ matrix with elements g_{ij}. The sweep operator SWP[k] operates on $\mathbf{G}$ by replacing it with another $p \times p$ symmetric matrix $\mathbf{H}$,

$$\mathbf{H} = \text{SWP}[k]\mathbf{G},$$

where the elements of **H** are given by

$$h_{kk} = -\frac{1}{g_{kk}}$$

$$h_{jk} = h_{kj} = \frac{g_{jk}}{g_{kk}} \text{ for } j \neq k$$

$$h_{jl} = h_{lj} = g_{jl} - \frac{g_{jk}g_{kl}}{g_{kk}} \text{ for } j \neq k \text{ and } l \neq k.$$

Algorithm (2.2) presents the (forward) sweep operator. After applying Algorithm (2.2) on the kth position on a matrix, the matrix is said to be swept on position k. To return a swept matrix to its original form, we define a reverse sweep operator denoted by

$$\mathbf{H} = \text{RSWP}[k]\mathbf{G},$$

Algorithm 2.2. Sweep Matrix on Position k

Requires: $\mathbf{G}_{p\times p}$ symmetric matrix

1. *initialize* $\mathbf{H} \leftarrow \mathbf{0}_{p\times p}$
2. $h_{kk} \leftarrow -1/g_{kk}$
3. **for** $j \leftarrow 1$ to p and $j \neq k$ **do**
4. $\quad h_{jk} = h_{kj} \leftarrow -g_{jk}h_{kk}$
5. **end for**
6. **for** $j \leftarrow 1$ to p and $j \neq k$ **do**
7. $\quad$ **for** $l \leftarrow 1$ to p and $l \neq k$ **do**
8. $\quad\quad h_{jl} = h_{lj} \leftarrow g_{jl} - g_{kl}h_{jk}$
9. $\quad$ **end for**
10. **end for**

where the elements of **H** are similar to those given by the (forward) sweep operator,

$$h_{kk} = -\frac{1}{g_{kk}}$$

$$h_{jk} = h_{kj} = -\frac{g_{jk}}{g_{kk}} \text{ for } j \neq k$$

$$h_{jl} = h_{lj} = g_{jl} - \frac{g_{jk}g_{kl}}{g_{kk}} \text{ for } j \neq k \text{ and } l \neq k,$$

with the difference in negating the calculation of $h_{jk} = h_{kj}$. For computational convenience, the rows of matrix **X** should be sorted, from minimum missingness to maximum missingness, into S unique missingness patterns. Figure (2.3) displays one desired form of sorting based on missingness.

Similar to the matrix of binary indicators defined in (2.1), let **R** be an $s \times p$ matrix of binary indicators with elements given by

$$r_{sj} = \begin{cases} 1, & \text{if } \mathbf{x}_j \text{ is observed in pattern } s \\ 0, & \text{if } \mathbf{x}_j \text{ is missing in pattern } s. \end{cases}$$

For each missingness pattern s, let $\mathcal{O}(s)$ denote the subset of column labels $\{1,2, \dots, p\}$ corresponding to variables that are observed,

	Variables				
Patterns	X_1	X_2	X_3	⋯	X_p
s_1	1	1	1		1
s_2	0	1	1		1
s_3	1	0	1		1
.	.	.	.		.
.	.	.	.		.
.	.	.	.		.
s_S	1	0	0		0

Figure 2.3. Matrix of missingness patterns for $\mathbf{X}_{n\times p}$, where 1 denotes an observed value and 0 denotes a missing value. Each row of $\mathbf{X}_{n\times p}$ is grouped into a unique missing data pattern $s = 1,2 \dots, S$.

$$\mathcal{O}(s) = \{j: r_{sj} = 1\}.$$

Similarly, let $\mathcal{M}(s)$ denote the subsets of columns $\{1,2, \dots, p\}$ that are missing,

$$M(s) = \{j: r_{sj} = 0\}.$$

Lastly, let $i(s)$ denote the subset of rows $\{1,2,\dots,n\}$ corresponding to $\mathbf{X}$ that are in missing pattern s.

The most common way to simulate multivariate normal random vectors within the I-step is using the Cholesky decomposition on the square submatrices of SWP$[\mathcal{O}(s)]\boldsymbol{\theta}$ (Schafer, 1997). The Cholesky decomposition is a numerically efficient technique used to decompose a matrix into the product of a lower triangular matrix and its conjugate transpose (Gentle, 1998). The complete-data sufficient statistics for the multivariate normal distribution are given by a vector of column sums

$$\mathbf{T}_1 = \left[\sum_{i=1}^{n} x_{i1}, \sum_{i=1}^{n} x_{i2}, \dots, \sum_{i=1}^{n} x_{ip},\right]'$$

and a matrix of sums of squares and cross-products

$$\mathbf{T}_2 = \begin{bmatrix} \sum_{i=1}^{n} x_{i1}^2 & \sum_{i=1}^{n} x_{i1}x_{i2} & \cdots & \sum_{i=1}^{n} x_{i1}x_{ip} \\ \sum_{i=1}^{n} x_{i2}x_{i1} & \sum_{i=1}^{n} x_{i2}^2 & \cdots & \sum_{i=1}^{n} x_{i2}x_{ip} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{n} x_{ip}x_{i1} & \sum_{i=1}^{n} x_{ip}x_{i2} & \cdots & \sum_{i=1}^{n} x_{ip}^2 \end{bmatrix}.$$

We can arrange the sufficient statistics into a $(p+1)\times(p+1)$ matrix $\mathbf{T}$ as

$$\mathbf{T} = \begin{bmatrix} n & \mathbf{T}_1' \\ \mathbf{T}_1 & \mathbf{T}_2 \end{bmatrix}.$$

Using the sweep operator, we can calculate the MLEs as

$$\begin{aligned} \widehat{\boldsymbol{\theta}}_{\text{MLE}} &= \text{SWP}[0]n^{-1}\mathbf{T} \\ &= \begin{bmatrix} -1 & \mathbf{T}_1'/n \\ \mathbf{T}_1/n & \mathbf{T}_2 - (\mathbf{T}_1\mathbf{T}_1')/n \end{bmatrix} \\ &= \begin{bmatrix} -1 & \widehat{\boldsymbol{\mu}}_{\text{MLE}}' \\ \widehat{\boldsymbol{\mu}}_{\text{MLE}} & \widehat{\boldsymbol{\Sigma}}_{\text{MLE}} \end{bmatrix}. \end{aligned}$$

For missing data, $\mathbf{T}$ is partitioned into observed and missing sufficient statistics such that $\mathbf{T} = \mathbf{T}_{\text{obs}} + \mathbf{T}_{\text{mis}}$ for each missingness pattern. $\mathbf{T}_{\text{obs}}$ is a matrix that contains the sufficient statistics for only fully observed variables and places zeros in rows and columns

corresponding to missing variables, whereas, $\mathbf{T}_{\text{mis}}$ contains the complement of $\mathbf{T}_{\text{obs}}$. As such,

$$\mathbf{T}_{\text{obs}} = \sum_{s=1}^{S} \mathbf{T}_{\text{obs}}(s)$$

and

$$\mathbf{T}_{\text{mis}} = \sum_{s=1}^{S} \mathbf{T}_{\text{mis}}(s).$$

The JM MI algorithm is presented in Algorithm (2.3).

2.3.5. Fully conditional specification multiple imputation. In practice, it is often difficult to specify a realistic joint model $P(\mathbf{X}, \mathbf{R})$ that incorporates the model for generating imputations and the substantive model for which the data was sampled. As such, the fully conditional specification (FCS) approach to missing data imputes missing data on a variable-by-variable basis. That is, as opposed to assuming the joint distribution $P(\mathbf{X}, \mathbf{R}|\boldsymbol{\theta})$ follows a multivariate normal distribution as in the JM approach, the FCS approach specifies this multivariate density as a series of univariate conditional densities (Gelman & Speed, 1993) of the form

$$P\left(\mathbf{x}_j \middle| \mathbf{X}_{-j}, \mathbf{R}, \boldsymbol{\theta}\right), \tag{2.31}$$

where $\mathbf{x}_j$ is the jth variable of $\mathbf{X}$ and $\mathbf{X}_{-j} = \left[\mathbf{x}_1, \dots, \mathbf{x}_{j-1}, \mathbf{x}_{j+1}, \dots, \mathbf{x}_p\right]$ is collection of the $p-1$ variables in $\mathbf{X}$ excluding the jth variable $\mathbf{x}_j$ (van Buuren, 2007). The conditional densities in (2.31) are the form of univariate regression models in which $\mathbf{x}_j$ is regressed on all other variables $\mathbf{X}$.

The multiple imputation by chained equations (MICE) algorithm is one popular algorithm that implements FCS MI (see Algorithm [2.4]). MICE first loops

Algorithm 2.3. Joint Multivariate Normal Model Multiple Imputation

Require: $\mathbf{X}_{n\times p}$ matrix with rows sorted by S missingness patterns

1: *initialize* $\boldsymbol{\theta}^0 \leftarrow (\boldsymbol{\mu}^0, \boldsymbol{\Sigma}^0)$
2: $\mathbf{T} \leftarrow \mathbf{T}_{\text{obs}}$
3: **for** $t \leftarrow 1$ to T
4: **for** $s \leftarrow 1$ to S **do**
5: **for** $j \leftarrow 1$ to p **do**
6: **if** $r_{sj} = 1$ and $\boldsymbol{\theta}_{jj} > 0$
7: $\boldsymbol{\theta} \leftarrow \text{SWP}[j]\boldsymbol{\theta}$
8: **else**
9: $\boldsymbol{\theta} \leftarrow \text{RSWP}[j]\boldsymbol{\theta}$
10: **end for**
11: $\mathbf{C} \leftarrow \text{Cholesky}_{\mathcal{M}(s)}\boldsymbol{\theta}$
12: **for** $i \in \mathcal{i}(s)$ **do**
13: **for** $j \in \mathcal{M}(s)$ **do**
14: $\mathbf{x}_{ij} \leftarrow \boldsymbol{\theta}_{0j}$
15: **for** $k \in \mathcal{O}(s)$ **do**
16: $\mathbf{x}_{ij} \leftarrow \mathbf{x}_{ij} + \boldsymbol{\theta}_{kj}\mathbf{x}_{ik}$
17: **end for**
18: $\dot{\mathbf{z}}_j \leftarrow$ draw from $N(0,1)$
19: **for** $k \in \mathcal{M}(s)$ and $k \leq j$ **do**
20: $\mathbf{x}_{ij} \leftarrow \mathbf{x}_{ij} + \mathbf{C}_{kj}\mathbf{z}_k$
21: **end for**
22: $\mathbf{T}_{0j} \leftarrow \mathbf{T}_{0j} + \mathbf{x}_{ij}$
23: **for** $k \in \mathcal{O}(s)$ **do**
24: $\mathbf{T}_{kj} \leftarrow \mathbf{T}_{kj} + \mathbf{x}_{ij}\mathbf{x}_{ik}$
25: **end for**
26: **for** $k \in \mathcal{M}(s)$ and $k \leq j$ **do**
27: $\mathbf{T}_{kj} \leftarrow \mathbf{T}_{kj} + \mathbf{x}_{ij}\mathbf{x}_{ik}$
28: **end for**
29: **end for**
30: **end for**
31: **end for**
32: $\dot{\boldsymbol{\theta}}^t = (\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t) \leftarrow$ draw from $p(\boldsymbol{\theta}|\mathbf{x}_{\text{obs}}, \mathbf{x}^t_{\text{mis}})$
33: **end for**

through each of the p variables in $\mathbf{X}$ and if missingness is found on $\mathbf{x}_j$ (Line 3), then the missing values of $\mathbf{x}_j$ ($\mathbf{x}_{i_{\text{mis}(j)}}$) are replaced by random draws from the observed values of $\mathbf{x}_j$ ($\mathbf{x}_{i_{\text{obs}(j)}}$) (Line 4). Next, the parameters of interest $\boldsymbol{\theta}_j$ are drawn from the posterior distribution of the current values of $\boldsymbol{\theta}_j^{(t)}$ at the tth iteration conditional on the observed

Algorithm 2.4. Multivariate Imputation by Chained Equations Multiple Imputation

Require: $\mathbf{X}_{n\times p}$ matrix with missing values

1: **for** $t \leftarrow 1$ to T **do**
2: **for** $j \leftarrow 1$ to p **do**
3: $i_{\text{obs}} \leftarrow \{\text{indices observed data in } \mathbf{x}_j\}$
4: $i_{\text{mis}} \leftarrow \{\text{indices missing data in } \mathbf{x}_j\}$
5: **if** $\mathbf{x}_{i_{\text{mis}}(j)} = \emptyset$
6: **pass**
7: **else**
8: **if** any$\left(\mathbf{X}_{i_{\text{obs}}(-j)}\right)$ missing
9: $\mathbf{X}_{i_{\text{obs}}(-j)} \leftarrow$ random draws from observed data in $\mathbf{X}_{(-j)}$
10: **end if**
11: $\dot{\boldsymbol{\theta}}_j^{(t)} \leftarrow$ draw from $P\left(\boldsymbol{\theta}_j^{(t)} \middle| \mathbf{x}_{i_{\text{obs}}(j)}, \dot{\mathbf{X}}_{i_{\text{obs}}(-j)}^{(t)}\right)$
12: $\dot{\mathbf{x}}_{i_{\text{mis}}(j)}^{(t)} \leftarrow$ draw from $P\left(\mathbf{x}_{i_{\text{mis}}(j)} | \mathbf{x}_{i_{\text{obs}}(j)}, \dot{\mathbf{X}}_{i_{\text{obs}}(-j)}^{(t)}, \dot{\boldsymbol{\theta}}_j^{(t)}\right)$
13: **end if**
14: **end for**
15: **end for**

values of $\mathbf{x}_j$ ($\mathbf{x}_{i_{\text{obs}}(j)}$), and the observed and imputed values in $\mathbf{X}_{-j}$ ($\dot{\mathbf{X}}_{(-j)}^{(t)}$) (lines 6), where $\dot{\mathbf{X}}_{(-j)}^{(t)} = \left[\dot{\mathbf{x}}_1^{(t)}, \ldots, \dot{\mathbf{x}}_{j-1}^{(t)}, \dot{\mathbf{x}}_{j+1}^{(t-1)}, \ldots, \dot{\mathbf{x}}_p^{(t-1)}\right]$. Finally, imputations for missing values in $\mathbf{x}_j$ $\mathbf{x}_{i_{\text{mis}}(j)}$ are drawn using the distribution of $\mathbf{x}_{i_{\text{mis}}(j)}$ conditional on the observed values of $\mathbf{x}_{i_{\text{obs}}(j)}$, the observed and imputed values $\dot{\mathbf{X}}_{(-j)}^{(t)}$, and the current parameter draws $\dot{\boldsymbol{\theta}}_j^{(t)}$ (Line 7). As described above, the MICE algorithm can be viewed as a Gibbs sampler used to obtain the posterior distribution of $\boldsymbol{\theta}$ by sampling iteratively from a series of conditional distributions, where the parameters are treated as specific to their respective conditional densities (Azur, Stuart, Frangakis, & Leaf, 2011). van Buuren and Groothuis-Oudshoorn (2011) note that empirically, MICE converges after approximately 5-10 iterations. For continuous-valued variables that are normally distributed, a Bayesian linear regression with non-informative priors can be used to impute missing values (see Algorithm [2.5]). Similarly, for binary-valued variables, an approximate Bayesian logistic regression with non-informative priors

can be used to impute missing values (see Algorithm [2.6]). See van Buuren (2012) for a description of these algorithms.

2.3.6. Model compatibility. One of the widely cited criticisms against FCS is the issue of compatible conditional models. In the simplest case, two conditional densities $f(x|y)$ and $g(y|x)$ are compatible if their density ratio $f(x|y)/g(y|x)$ factorizes into $u(x)v(y)$ for some integrable functions u and v (Arnold & Press, 1989; Besag, 1974). Recently, the convergence properties of FCS under compatible conditionals have begun to be understood (Liu, Gelman, Hill, Su, & Kropko, 2014) and several approaches have been proposed to determine the amount of compatibility (Arnold, Castillo, & Sarabia, 1999; Chen, 2010). However, in FCS MI the 'true' joint distribution is not known and of scientific interest. When incompatible conditionals are used, FCS MI is not guaranteed to converge to the true underlying multivariate density of $P(\mathbf{X}, \mathbf{R}|\boldsymbol{\theta})$ (Liu et al., 2014). However, research has shown that clearly incompatible conditionals generally have little impact on the empirical performance of the FCS MI algorithm (van Buuren, Brands, Groothuis-Oudshoorn, & Rubin, 2006), and may lead to consistent estimates when other technical conditions are met given in Liu et al. (2014). Similarly, Gelman (2004) argues that having a joint distribution in the imputation is less important than incorporating unique effects in the conditional imputation models (e.g., nonlinear effects).

2.3.7. Comparison between JM and FCS. The goal of both JM and FCS MI is to obtain confidence proper estimates in the presence of missing data, where JM MI models the observed and missing data using a multivariate normal distribution and FCS MI models the observed and missing data using a series of conditional densities. In cases with continuous and normally distributed data, studies have consistently shown that FCS MI using linear

Algorithm 2.5. MICE Bayesian Linear Regression Imputation

Require: $\mathbf{X}_{n\times p}$ matrix with missing values only in $\mathbf{x}_j$

for $t \leftarrow 1$ to T **do**
  $i_{\text{obs}} \leftarrow \{\text{indices observed data in } \mathbf{x}_j\}$
  $i_{\text{mis}} \leftarrow \{\text{indices missing data in } \mathbf{x}_j\}$
  if any$\left(\mathbf{X}_{i_{\text{obs}}(-j)}\right)$ missing
    $\mathbf{X}_{i_{\text{obs}}(-j)} \leftarrow$ random draws from observed data in $\mathbf{X}_{(-j)}$
  end if
  $\mathbf{V} \leftarrow \left(\mathbf{X}'_{i_{\text{obs}}(-j)}\mathbf{X}_{i_{\text{obs}}(-j)} + \text{diag}\left(\mathbf{X}'_{i_{\text{obs}}(-j)}\mathbf{X}_{i_{\text{obs}}(-j)}\right)\kappa\right)^{-1}$
  $\widehat{\boldsymbol{\beta}} \leftarrow \mathbf{V}\mathbf{X}'_{i_{\text{obs}}(-j)}\mathbf{x}_{i_{\text{obs}}(j)}$
  $\dot{g} \leftarrow$ draw from $\chi^2_{df=n_1-(p-1)}$
  $\dot{\sigma}^{2(t)} \leftarrow \left(\mathbf{x}_{i_{\text{obs}}(j)} - \mathbf{X}'_{i_{\text{obs}}(-j)}\widehat{\boldsymbol{\beta}}\right)'\left(\mathbf{x}_{i_{\text{obs}}(j)} - \mathbf{X}'_{i_{\text{obs}}(-\text{j})}\widehat{\boldsymbol{\beta}}\right)/\dot{g}$
  $\dot{\mathbf{z}}_1 \sim (p-1)$ independent draws from $N(0,1)$
  $\mathbf{V}^{1/2} \leftarrow \text{Cholesky}(\mathbf{V})$
  $\dot{\boldsymbol{\beta}}^{(t)} \leftarrow \widehat{\boldsymbol{\beta}} + \dot{\sigma}^{(t)}\mathbf{V}^{1/2}\dot{\mathbf{z}}_1$
  $\dot{\mathbf{z}}_2 \sim n_2$ independent draws from $N(0,1)$
  $\dot{\mathbf{x}}^{(t)}_{i_{\text{mis}}(j)} \leftarrow \mathbf{X}_{i_{\text{mis}}(-j)}\dot{\boldsymbol{\beta}}^{(t)} + \dot{\mathbf{z}}_2\dot{\sigma}^{(t)}$
end for

Algorithm 2.6. MICE Approximate Bayesian Logistic Regression Imputation

Require: $\mathbf{X}_{n\times p}$ matrix with missing values only in $\mathbf{x}_j$

for $t \leftarrow 1$ to T **do**
  $i_{\text{obs}} \leftarrow \{\text{indices observed data in } \mathbf{x}_j\}$
  $i_{\text{mis}} \leftarrow \{\text{indices missing data in } \mathbf{x}_j\}$
  if any$\left(\mathbf{X}_{i_{\text{obs}}(-j)}\right)$ missing
    $\mathbf{X}_{i_{\text{obs}}(-j)} \leftarrow$ random draws from observed data in $\mathbf{X}_{(-j)}$
  end if
  $\widehat{\boldsymbol{\beta}} \leftarrow \arg\min_{\boldsymbol{\beta}} L\left(\boldsymbol{\beta}|\mathbf{x}_{i_{\text{obs}}(j)},\mathbf{X}_{i_{\text{obs}}(-j)}\right)$
  $\mathbf{V} \leftarrow \text{Cov}\left(\widehat{\boldsymbol{\beta}}\right)$
  $\mathbf{V}^{1/2} \leftarrow \text{Cholesky}(\mathbf{V})$
  $\dot{\boldsymbol{\beta}}^{(t)} \leftarrow \widehat{\boldsymbol{\beta}} + \mathbf{V}^{1/2}\dot{\mathbf{z}}_1$
  $\dot{\mathbf{p}} \leftarrow n_2$ predicted probabilites $\left(1 + \exp\left(-\mathbf{X}_{i_{\text{mis}}(-j)}\dot{\boldsymbol{\beta}}^{(t)}\right)\right)^{-1}$
  $\mathbf{u} \leftarrow n_2$ independent draws from $U(0,1)$
  $\dot{\mathbf{x}}^{(t)}_{i_{\text{mis}}(j)} = \begin{cases} 1, & u_i \geq \dot{p}_i \\ 0, & u_i < \dot{p}_i \end{cases} \quad i = 1, \ldots, n_2$
end for

regression models with constant variance and the JM approach yield parameter estimates

and standard errors that are approximately similar (Karangwa, 2013; Kropko, Goodrich, Gelman, & Hill 2014; Raghunathan, Lepkowski, Van Hoewyk & Solenberger, 2001). Despite the differences in implementation, in a limited number of simple cases (e.g., the joint model is multivariate normal and the conditional models are all linear regression modes with constant variance) described in (Hughes, White, Seaman, Carpenter, Tilling, & Sterne, 2014) and (Liu et al., 2014), theoretically the JM and FCS approach are equivalent. Thus, the results that FCS MI using linear regression models with constant variance and JM produce similar parameter estimates and standard errors in the case of continuous, normally distributed data are intuitive.

In more common implementations of FCS MI (e.g., using nonlinear or nonparametric regression models for imputation), the theoretical connections between JM and FCS do not hold (Hughes et al., 2014). In such cases, the empirical research has been mixed concerning which method is better for imputing MAR data with different variable types (e.g., binary, ordered categorical, unordered categorical). For instance, in the case of imputing ordered categorical data, Finch (2010) and Lee and Carlin (2010) found that compared to FCS MI, JM MI generally resulted in less biased regression parameter estimates and lower standard errors. Likewise, other research has shown that JM performs slightly better in terms of regression parameter bias and standard errors than FCS for imputing binary variables (Lee & Carlin, 2010) and imputing unordered categorical variables (Karangwa, Kotze, & Blignaut, in press). On the contrary, in a more comprehensive study comparing the performance of JM and FCS for imputing missing data on continuous, binary, ordered categorical, and unordered categorical variables, Kropko et al. (2014) found that FCS MI outperformed JM MI in terms of regression coefficients' accuracy and standard errors.

Compared to the JM MI, one important difference is that under FCS no information about $\mathbf{x}_{\text{mis}(j)}$ is used to draw $\boldsymbol{\theta}_j$. In simpler terms, the FCS algorithm is a concatenation of univariate regression procedures applied to the cases with complete $\mathbf{x}_{\text{obs}(j)}$ and deviates from MCMC theory at this point. A consequence of specifying the sampler as a series of independent conditional densities is a faster rate of convergence (around 5-10 iterations) compared to JM MI, which often requires hundreds to thousands of iterations (Schafer, 1997). Convergence in this context refers to the stability of regression coefficients in the presence of missing data, and not convergence to a 'true' stationary posterior distribution. The independence specification is also a theoretical weakness of FCS in that the conditional densities are usually not derived from a 'true' joint model, so the Gibbs sampler routine is not guaranteed to converge to a 'true' posterior distribution (which is often unknown in MI contexts).

2.3.8. Imputation with nonlinear effects. Despite the mixed empirical performance potential for incompatible conditional models, the FCS framework for MI easily allows the specification of nonlinear effects such as interactions in an imputation model. The JM MI framework also allows for imputing nonlinear effects if the nonlinear effects are first created and then the imputation routine is applied, this method is considerably limited when the assumption of multivariate normality is not reasonable. Regardless of the MI framework, for conditional indirect effects, if the interaction effects are ignored in the imputation models, then estimation and inference for conditional indirect effects are likely to be biased (Allison, 2003). For such cases, there are two scenarios in which nonlinearities can be modeled in an imputation procedure. First, an interaction variable may be explicitly imputed. Consider the linear model,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 I_{X_1 X_2} + \zeta_Y, \quad (2.32)$$

where $I_{X_1 X_2} = X_1 X_2$ is the interaction between X_1 and X_2 and X_2 contains missing data and ζ_Y follows a standard normal error distribution. Von Hippel (2009) describes two common approaches to imputing interactions. The first method is called impute then transform (also called passive imputation [PI]; this is the default procedure in MICE). Here, to impute an interaction term, first impute the main effects with missing values (impute), then manually create the interaction term (transform), and iterate until all missing interactions are imputed. Applied to the model in (2.32), $X_{\text{mis}(2)}$ will be imputed first to obtain $\dot{X}_2$, and the interaction term will be defined as $I_{X_1 \dot{X}_2} = X_1 \dot{X}_2$.

In the second method, called transform then impute, to impute an interaction term, the interaction variable is created and treated as another variable (transform), and an imputation model is used (impute, e.g., using a JM or FCS approach). This method is often called the just another variable (JAV) method.

Applied to the model in (2.32), similar to the impute then transform method, $X_{\text{mis}(2)}$ could be imputed first to obtain $\dot{X}_2$. However, now an imputation model $I_{X_1 \dot{X}_2} = X_1 \dot{X}_2$ will be specified to impute the missing values on this interaction variable (treating it as just another variable). The imputation order, or visit sequence in MICE (van Buuren & Groothuis-Oudshoorn, 2011), can be specified in any order. This latter example demonstrates an FCS approach to imputation. However, a JM approach could be used where the interaction variable is first created, then a joint imputation model (i.e., multivariate normal) is specified for all the variables as in Enders et al. (2014).

With regards to the two methods above for imputing interaction terms in linear regression models, von Hippel (2009) found that the PI method produced biased

regression estimates and incorrect standard errors. However, the JAV method produced inaccurate standard errors but unbiased regression coefficients. Seaman et al. (2012) reported slightly different results to that of von Hippel (2009). Specifically, Seaman et al. (2012) found that compared to PI, JAV gives consistent estimation, but only when the covariate with missing data had a MCAR missingness mechanism; under a MAR missingness mechanism, the JAV approach was biased in some cases, but still outperformed passive imputation. Enders et al. (2014) found that the JAV method led to slightly biased regression coefficient estimates in some cases, but generally nominal levels of confidence interval coverage. From these results, it appears that JAV approach is reasonable to use in linear regression models. In the case of biased standard errors, a resampling technique such as the bootstrap (nonparametric or Bayesian) can be used to obtain more accurate standard error estimates and confidence intervals (Efron & Tibshirani, 1993; Rubin, 1981).

On the contrary, with regards to logistic regression and the JAV imputation method, the research is inconclusive. von Hippel (2009) claims that the JAV approach works well in logistic regression models; however, his analysis and conclusions were based on an existing data set in which the true regression coefficients were unknown. Although Seaman et al. (2012) did not examine logistic regression models with interactions, they did find that both passive imputation and JAV methods for imputing quadratic covariates with MCAR and MAR missingness mechanisms resulted in substantial bias. In addition, the confidence interval coverage of the regression coefficient for the quadratic term was a function of the response distribution, specifically, with better coverage rates when the response had an even distribution in the two classes, $P(Y = 1) = P(Y = 0) = .5$, than when the response had a biased distribution in the two classes $P(Y = 1) = 01, P(Y = 0) = .9$.

A more recent method was developed by Bartlett et al. (2014) to overcome the limitations of PI and JAV in imputing nonlinear covariate terms. The proposed method is an extension of FCS, called substantive model compatibility FCS (SMC-FCS), which derives imputation models for nonlinear covariate terms that are theoretically compatible with the underlying substantive model that the nonlinear term is included in as a covariate. Results from simulation studies show that compared to PI and JAV, SMC-FCS provides more consistent estimates for both linear regression and Cox proportional hazards models that contain nonlinear covariate terms. Although these results are promising, as applied the structural equation models, the application of their methods is ambiguous. Specifically, the SMC-FCS approach derives compatible distributions for variables with missing data based on an underlying substantive model, but in SEMs, a variable with missing data may appear as a covariate in several substantive models and even be the response of another substantive model. Moreover, the empirical performance of their method in the case of logistic regression models with nonlinear effects is unknown.

PI, JAV, and SMC-FCS are imputation procedures designed to impute a nonlinear term that contains missing data. The second scenario in which nonlinearities can be modeled in an imputation procedure is when an endogenous variable contains missing data in which the underlying substantive model contains nonlinear terms. This scenario is particularly relevant for moderated mediation models, in which the response variable may contain missing data and the underlying substantive model contains at least one interaction term. For best practice of MI, the imputation model should at least incorporate the same effects as the substantive model (Allison, 2003; Little & Rubin, 2002; Schafer, 1997, van Buuren, 2012; Zhang, 2003). That is, if the substantive model contains

interactions, then the imputation model should contain interactions. It is important to note that unless directly specified, the JM MI algorithm and MICE implementation of the FCS MI were not designed to automatically impute interaction effects. That is, the interaction variables must be explicitly created and added to the existing data before either of these MI algorithms are used.

For imputing variables in which the substantive model contains nonlinear effects, tree-based imputation models (e.g., classification and regression trees, random forests, gradient boosted trees) have shown to be very effective at automatically retaining the nonlinear effects of such models (Abrahantes, Sotto, Molenberghs, Vromman, & Bierinckx, 2011; Burgette & Reiter, 2010; Doove et al., 2014; Shah, Bartlett, Carpenter, Nicholas, & Hemingway, 2014; Stekhoven & Bühlmann, 2012; Twala, Jones, & Hand, 2008; Wang & Feng, 2009). The application of these models for imputation of missing data will be described in detail in the next chapter.

CHAPTER 3

SUPERVISED MACHINE LEARNING FOR MULTIPLE IMPUTATION

3.1. Supervised Machine Learning Perspective of Missing Data

Machine learning is a subfield of computer science that explores the development and applications of computational algorithms to extract patterns and trends in data (Mitchell, 1997). The applications of machine learning models are often classified into three categories: (1) supervised learning, (2) unsupervised learning, and (3) reinforcement learning. The framework for supervised machine learning applications to predictive modeling and pattern recognition can be paralleled to applications of missing data imputation. This chapter discusses the concept of supervised learning, the application of supervised machine learning algorithms to multiple imputation in structural equation models, and a novel MI algorithm that combines classic linear models with tree-based ensemble machine learning models.

In the simplest case of supervised learning, a data set with N samples is first partitioned into two independent data sets, a training set $\mathcal{D}_{\text{train}} = \{y_i, \mathbf{x}_i\}_1^{n_1}$ and a test set $\mathcal{D}_{\text{test}} = \{y_i, \mathbf{x}_i\}_1^{n_2}$, where n_1 and n_2 denote the sizes of the training and test sets, respectively and $n_1 + n_2 = N$. In both sets, $\mathbf{x}_i = [x_1, \dots, x_p]$ is a $1 \times p$ vector of features (or covariates) for sample i and y_i is the corresponding label (or response variable to be predicted) for sample i. After partitioning the samples, the training set is used to estimate (or train) the model parameters and the trained model is used to predict the labels (or response values) on the test set. When the labels are known in the test set, metrics can be calculated to determine the accuracy of the model predictions. Unfortunately, in most

practical applications every sample i does not have a corresponding label y_i. When the labels are unknown for some samples, often the n_1 samples with labels will correspond to the training set and the n_2 unlabeled samples will correspond to the test set, in which case $\mathcal{D}_{\text{test}} = \{y_i, \mathbf{x}_i\}_1^{n_2}$ becomes $\mathcal{D}_{\text{test}} = \{\mathbf{x}_i\}_1^{n_2}$.

To make the supervised learning connection to a missing data context, consider a data set $\mathbf{X}$ with N samples and p features. In the simplest case only the jth variable $\mathbf{x}_j$ has missing values. If we treat $\mathbf{x}_j$ as our corresponding label variable, then for sample i we can partition the samples based on the n_1 observed values for $\mathbf{x}_j$ and the n_2 missing values for $\mathbf{x}_j$. That is, let the observed data set $\mathcal{D}_{i_{\text{obs}},j} = \left\{x_{ij}, \mathbf{x}_{i(-j)}\right\}_{i=1}^{n_1}$ (or training set) and missing data set $\mathcal{D}_{i_{\text{mis}},j} = \left\{\mathbf{x}_{i(-j)}\right\}_{i=1}^{n_2}$ (or test set), where n_1 and n_2 denote the sizes of the observed and missing test sets, respectively and $n_1 + n_2 = N$. Here, the notation $\mathcal{D}_{i_{\text{obs}},j} = \left\{x_{ij}, \mathbf{x}_{i(-j)}\right\}_{i=1}^{n_1}$ denotes the set of n_1 samples with x_{ij} observed for sample i and variable j and with $\mathbf{x}_{i(-j)} = \left[x_{i1}, \dots, x_{i(j-1)}, x_{i(j+1)}, \dots, x_{ip}\right]$ as the $1 \times p - 1$ vector of features for sample i that does not include variable j. Similarly, $\mathcal{D}_{i_{\text{mis}},j} = \left\{\mathbf{x}_{i(-j)}\right\}_{i=1}^{n_2}$ denotes the set of n_2 samples with x_{ij} missing for sample i and variable j and with $\mathbf{x}_{i(-j)} = \left[x_{i1}, \dots, x_{i(j-1)}, x_{i(j+1)}, \dots, x_{ip}\right]$ as the $1 \times p - 1$ vector of features for sample i that does not include variable j (see Figure [3.1] for a visual).

After partitioning the samples, the observed data $\mathcal{D}_{i_{\text{obs}},j} = \left\{x_{ij}, \mathbf{x}_{i(-j)}\right\}_{i=1}^{n_1}$ can be used as a training set to estimate (or train) the model parameters and the trained model can be used to predict (or impute in statistical language) the labels for the missing data set $\mathcal{D}_{i_{\text{mis}},j} = \left\{\mathbf{x}_{i(-j)}\right\}_{i=1}^{n_2}$. The procedure described here forms the basis of the machine learning-

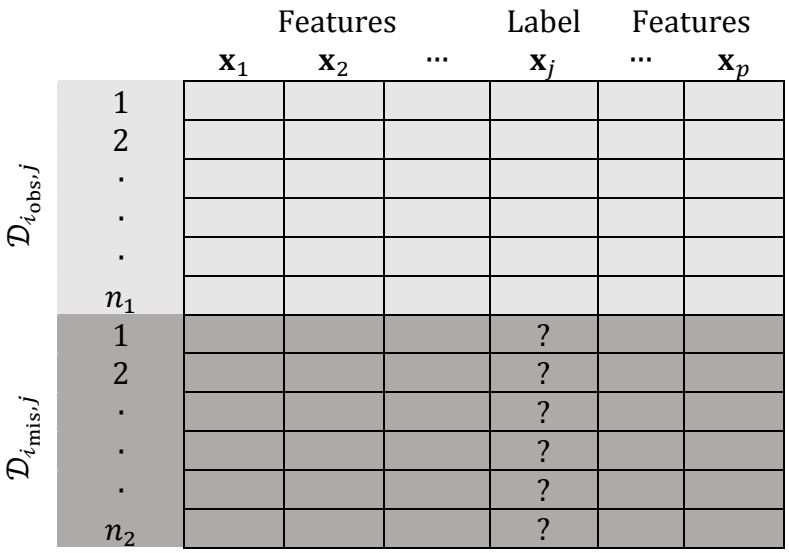


Figure 3.1. Example partitioning for supervised learning application to a missing data problem. The partitioning scheme is based on the n_1 observed values on label $\mathbf{x}_j$ (lighter grey rows) and the n_2 missing values on label $\mathbf{x}_j$ (darker grey rows).

based MI algorithms described later.

In psychological research, the most commonly used models for MI are parametric regression models (e.g., linear regression for continuous responses and logistic regression for discrete responses). Current implementation of these methods for MI involves predicting missing values on a variable or set of variables based on the first-order linear combination of the remaining variables in a data set. Although this approach works well under a variety of settings, there are inherent limitations. Specifically, only including first-order linear combinations precludes their use to imputing nonlinearities (e.g., interactions). This limitation is of practical importance to imputing missing data involving interactions in moderated mediation models. Nonlinear terms can be specified in the imputation models, however, other statistical models exist that automatically incorporate

nonlinearities into imputation routines. In addition, linear regression models are not robust to outliers, which are common in practice. Similarly, logistic regression models have poor performance with sparse data and small sample sizes (Cohen, Cohen, West, & Aiken, 2002). To overcome some of these issues, however, regularization penalties can be added to objective functions (Bishop, 2006).

In cases where parametric regression models with linear main effects are not appropriate due to concerns with nonlinearities in the data, nonparametric methods such as classification and regression tree (CART; Breiman, Friedman, Olsen, & Stone, 1984) models and ensemble CART-based models may be suitable replacements for MI. Unfortunately, these nonparametric methods are not without their limitations as well. In the next sections, we discuss the CART model, a flexible method of additive learning known as gradient boosting, and the applications of these methods to imputing missing data. In addition, we discuss the limitations of using only linear- or tree-based models for imputing missing data and introduce a unified framework that can automatically accommodate both types of imputation models using gradient boosting.

3.2. Classification and Regression Trees

3.2.1. Overview. Classification and regression trees (CART) models are simple, but powerful supervised machine learning models. The CART model is a type of adaptive basis-function model of the form

$$f(\mathbf{x}) = \sum_{j=1}^{J} \gamma_j \phi_j(\mathbf{x}), \tag{3.1}$$

in which $\phi_j(\mathbf{x}) = \phi_j(\mathbf{x}|\boldsymbol{\alpha}_j)$ is a parametric basis function. An advantage of this specification

is that the model is not linear in the parameters, so nonlinearities such as interactions between variables are easily modeled. Each parameter $\boldsymbol{\alpha}_j$ encodes both the feature used for splitting and the corresponding threshold value. The basis functions define the region, and the weights encode the response value in each terminal region. Specifically, for continuous-valued labels, the response value in the *j*th region is the arithmetic mean of the labels in the region, whereas for discrete-valued labels, the response value is the distribution of classes in the *j*th region.

The CART model uses a binary, recursive partition algorithm to partition the space of features based on a training set $\boldsymbol{\mathcal{D}}_{\text{train}}$ into $R_j, j = 1, \dots, J$ disjoint regions, where $\cup_{j=1}^{J} R_j = \boldsymbol{\mathcal{D}}_{\text{train}}$ and $\cap_{j=1}^{J} R_j = \emptyset$, into a piecewise-constant response surface (Hastie et al., 2009). In each partitioned region of the feature space a constant γ_j is assigned based on the predictive rule,

$$\mathbf{x} \in R_j \Longrightarrow f(\mathbf{x}) = \gamma_j. \tag{3.2}$$

In words, (3.2) simply says if the feature vector results in terminal node R_j, the predicted label is γ_j. Following the specification in (3.1), a tree can be formally expressed as

$$f(\mathbf{x}|\boldsymbol{\Theta}) = \sum_{j=1}^{J} \gamma_j I\big(\mathbf{x} \in R_j\big), \tag{3.3}$$

with parameters $\boldsymbol{\Theta} = \{R_j, \gamma_j\}_1^J$, where J denotes the total number of terminal nodes. Here, $I(\cdot)$ is the indicator function defined as

$$I\big(\mathbf{x} \in R_j\big) = \begin{cases} 1, & \mathbf{x} \text{ is member in } R_j \\ 0, & \text{otherwise} \end{cases}.$$

Figure (3.2) presents an example of a partitioned input space based on two features,

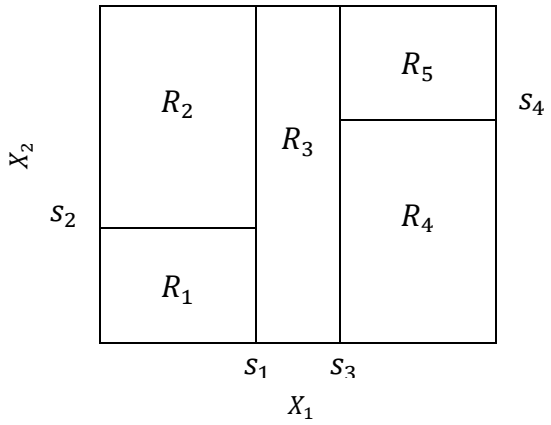


Figure 3.2. Example of partitioned joint input space based on two features, X_1 and X_2.

X_1 and X_2. The binary tree for the CART model corresponding to Figure (3.2) is displayed in Figure (3.3). Using the parameterization of the model given by (3.3), the CART model corresponding to the Figures (3.2) and (3.3) can be written as

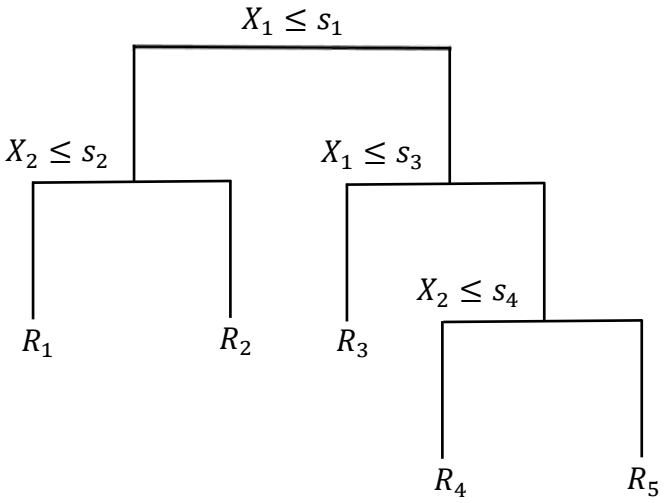


Figure 3.3. CART model based on Figure (3.2).

$$f\left(\mathbf{x}|\{R_j,\gamma_j\}_1^5\right)=\sum_{j=1}^{5}\gamma_j I(\mathbf{x}\in R_j)$$

$$=\gamma_1 I(X_1\le s_1\wedge X_2\le s_2)$$

$$+\gamma_2 I(X_1\le s_1\wedge X_2> s_2)$$

$$+\gamma_3 I(X_1> s_1\wedge X_1\le s_3)$$

$$+\gamma_4 I(X_1> s_1\wedge X_1> s_3\wedge X_2\le s_4)$$

$$+\gamma_5 I(X_1> s_1\wedge X_1> s_3\wedge X_2> s_4),$$

with ∧ denoting the 'and' logical condition. From Figure (3.3), the regions where splitting rules are shown (e.g., $X_1\le s_1$) and the partitioned input space (e.g., R_1) are called nodes. There are two types of nodes, parent nodes and child nodes. Each parent node is partitioned into a left child node and a right child node based on a selected splitting criterion $x_j\le x_j^{s^*}$ (see Figure [3.4]). If a node has no further partitions, it is called a terminal

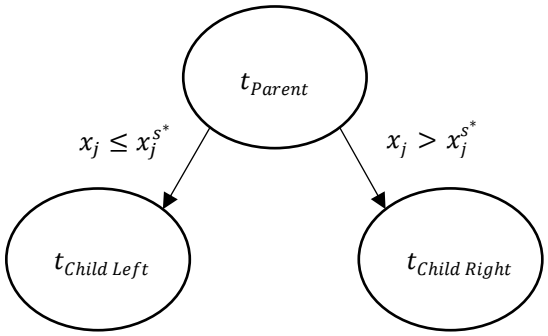


Figure 3.4. Splitting algorithm for CART.

node (denoted by R in Figure [3.3]). The basic idea of tree growing in a CART mode is to choose a binary split among all possible binary splits at each node so that the resulting child nodes are the 'purest'. In the CART model, only univariate splits are considered. That

is, each split depends on the value of only one feature. All possible splits consist of possible splits for each feature. A tree is grown starting from the root node by repeatedly using the following three steps on each node: (1) find each feature's best split, (2) find the node's best split, and (3) split the node using its best split found in Step (2) if stopping rules are not satisfied.

As mentioned above, the parameters in a CART model include the variables (or features) used for splitting, the threshold values for each split, and the expected response within each partitioned region. The parameters are found by minimizing the empirical risk

$$\widehat{\boldsymbol{\Theta}} = \arg\min_{\boldsymbol{\Theta}} \sum_{j=1}^{J} \sum_{\mathbf{x}_i \in R_j} L(y_i, \gamma_j). \tag{3.4}$$

However, minimizing (3.4) is an intractable optimization problem. Fortunately, a greedy, top-down approach is used to approximate global solutions to (3.4) (Brieman et al., 1984). The greedy approach is based on two parts: (1) the challenging part of finding R_j and (2) the easier part of finding γ_j given R_j. To solve both parts, it is often easier to approximate (3.4) by a smoother and more convenient criterion for optimizing R_j,

$$\widehat{\boldsymbol{\Theta}} = \arg\min_{\boldsymbol{\Theta}} \sum_{i=1}^{n_1} \big(y_i, f(\mathbf{x}_i|\boldsymbol{\Theta})\big).$$

Given $\hat{R}_j$, γ_j is often estimated by $\hat{\gamma}_j = \bar{y}_j$ (i.e., the mean of the $y_i \in R_j$) for regression loss functions or by the majority probability class of observations falling in region R_j for classification loss functions. The optimal splitting variables and split points are determined based on a node's impurity measure, $i(t)$, where t denotes the tth node in a tree (Hastie et. al, 2009). Maximum homogeneity of child nodes is defined by the impurity function $i(t)$. Since the impurity of a parent node t_p is constant for any of the possible splits $X_j \leq X_j^s, j =$

$1, \dots, p$, the maximum homogeneity of left and right child nodes will be equivalent to the maximization of change of impurity function $\Delta i(s,t)$

$$\Delta i(s,t) = i(t_P) - E(t_C),$$

where t_C consists of the left and right child nodes of the parent node t_P to be split on (James, Witten, Hastie, & Tibshirani, 2013). Assuming that $P_L = \frac{|t_L|}{|t|}$ and $P_R = \frac{|t_R|}{|t|}$ are the probabilities of the left and right child nodes, respectively, the splitting criterion can be defined as

$$\Delta i(s,t) = i(t_P) - P_L i(t_L) - P_R i(t_R).$$

Therefore, at each node CART solves the following maximization problem

$$\underset{x_j \le x_j^{s^*}, j=1,\dots,p}{\arg\max} \; [i(t_P) - P_L i(t_L) - P_R i(t_R)].$$

Depending on the type of problem, that is regression or classification, different impurity measures are used. For regression problems, the squared-error impurity function is used. The squared-error at node t can be defined as

$$i(t) = \frac{1}{|t|} \sum_{i \in t} (y_i - \gamma_t)^2,$$

where $\gamma_t = \frac{1}{|t|} \sum_{i \in t} y_i$ is the expected mean response value of labels in node t. On the contrary, for classification problems with $c = 1, \dots, C$ classes, generally the Gini index is the impurity function used. The Gini index at node t can be defined as

$$i(t) = \sum_{c=1}^{C} P(c|t)\big(1 - P(c|t)\big), \tag{3.5}$$

where $P(c|t) = \frac{|t_c|}{|t|}$ is the conditional probability of the cth class in node t. In the case of binary classification, (3.5) simplifies to

$$i(t) = 2P(c = 1|t)P(c = 2|t)$$

where the notation $P(c = 1|t)$ and $P(c = 2|t)$ denotes the conditional probability of labels in class 1 and class 2, respectively, at node t. In regularized stochastic gradient boosting (discussed later), we will introduce information gain as the impurity measure of choice.

The algorithm used in the CART model recursively splits tree nodes until specific stopping criteria are met. For instance, if a node is pure, that is, all samples within a node have identical labels, the node will not be split. Similarly, if all samples in a node have identical values for each feature, the node will not be split. Four additional user-specified stopping criteria for node splitting occur if: (1) the tree reaches the specified maximum depth limit, (2) the size of a node is less than the minimum specified node size, (3) the split of a node results in a child node whose node size is less than the specified minimum child node size, and (4) the best split found for a variable $x_j \leq x_j^{s^*}$ at node t is smaller than the specified minimum improvement (Breiman et al., 1984; Hastie et al., 2009). The CART algorithm is presented in Algorithm (3.1).

Algorithm 3.1. Classification and Regression Tree (CART)

Requires: Training data $\boldsymbol{\mathcal{D}}_{\text{train}} = \{y_{\text{i}}, \mathbf{x}_i\}_1^{n_1}$

1. Create node t
2. **if** stopping criteria met for t **then**
3. $\quad \hat{\gamma} \leftarrow E(y_i \in R_t)$
4. **else**
5. $\quad x^{s^*} \leftarrow \underset{x_j \leq x_j^{s^*}, j=1,\ldots,p}{\arg\max} \; [i(t_P) - P_L i(t_L) - P_R i(t_R)]$
6. $\quad \boldsymbol{\mathcal{D}} \leftarrow \boldsymbol{\mathcal{D}}_{t_L} \cup \boldsymbol{\mathcal{D}}_{t_R}$ according to x^{s^*}
7. $\quad t_L \leftarrow$ Build CART using $\boldsymbol{\mathcal{D}}_{t_L}$
8. $\quad t_R \leftarrow$ Build CART using $\boldsymbol{\mathcal{D}}_{t_R}$
9. **end if**

Compared to traditional statistical methods (e.g., linear regression, logistic

regression), CART models have several advantages. First, the CART model makes no formal distributional assumptions and can be used for both regression and classification (Strobl, Malley, & Tutz, 2009). Second, variable selection is automatic and the recursive binary partitioning algorithm can automatically fit nonlinear interactions (Doove et al., 2014). Lastly, CART models can easily handle sparse data (Strobl et al., 2009). On the contrary, although CART models provide relatively unbiased estimates, they are prone to overfitting (Hastie et al., 2009). That is, the model learns both the signal and the noise of a training set. Overfitting on a training set leads to poor model generalization (James et al., 2013). A preferred strategy to combat overfitting is to grow a large tree, then prune the tree using a cost-complexity pruning (see Hastie et al., 2009 for more details). Trees are also highly unstable. That is, a small change in a feature can often lead to different splits. Given the problems of overfitting and high variability in CART models, a clever method to reduce these problems is to combine the predictions of many tree models. The combination of multiple models is more commonly referred to as ensemble learning (Bishop, 2006).

3.2.2. Application to missing data. As discussed in the previous chapter, tree-based models such as CART have been shown to be useful for multiple imputation of missing data. Burgette and Reiter (2010) examined the bias, root-mean square error, and confidence interval coverage for linear regression coefficients estimated with missing data that was imputed using MICE and CART. Interestingly, Burgette and Reiter found that imputation routines using CART generally performed better than MICE on all three metrics, even on linear main effects. Doove et al. (2014) also found that CART outperformed MICE for regression coefficients of nonlinear terms, however, for regression coefficients of linear main effects, they found that MICE outperformed CART in all scenarios in their simulation

studies. Moreover, CART models preserved interaction effects better than random forests (i.e., an ensemble method that combines many CART models).

Although the results of these two studies are contradictory in terms of imputing linear main effects using CART, it is important to note that CART (and other tree-based models) are limited in two common missing data scenarios because of the underlying recursive partitioning algorithm used to build the trees. First, tree-based methods are only capable of data interpolation, that is, the imputed (or predicted) values are bounded by the range of the observed labels in the training data. Second, tree-based methods have difficulty in modeling linear main effects because the recursive-partitioning algorithm is inherently nonlinear due to a series of binary splits of the feature space. Consequently, CART should not be used as a standalone imputation model applied to all variable types. The next section discusses a more general framework, using gradient boosted models, that can incorporate models to impute both linear and nonlinear effects under the same general model.

3.3. Gradient Boosted Learning

3.3.1. Overview. Boosting is a powerful supervised machine learning meta-algorithm that is used to solve function estimation problems and also used to reduce bias and variance in prediction problems (Freund & Schapire, 1997). The function estimation view provides a general framework for its use in supervised machine learning applications. In a function estimation setting, based on a set of training $\boldsymbol{\mathcal{D}}_{\text{train}} = \{y_{\text{i}}, \mathbf{x}_i\}_1^{n_1}$, the goal is to find a function $f^*(\mathbf{x})$ that maps $\mathbf{x}$ to y, such that over the joint distribution of all $(y, \mathbf{x})$-values, the expected value of a specified loss function $L\big(y, f(\mathbf{x})\big)$ is minimized

$$f^*(\mathbf{x}) = \arg\min_{f(\mathbf{x})} E_{y,\mathbf{x}} L\big(y, f(\mathbf{x})\big)$$

$$= \arg\min_{f(\mathbf{x})} \underbrace{E_{\mathbf{x}}[\overbrace{E_y\left[L\left(y, f(\mathbf{x})\right)\right]}^{\text{expectation over } y}|\mathbf{x}]}_{\text{expectation over training data}} .$$

From this vantage, gradient boosting is a method that learns the functional relationship $f(\cdot)$ between inputs and outputs based on the features and labels of the training data (Natekin & Knoll, 2013). Boosting approximates $f^*(\mathbf{x})$ by an additive-type expansion of the form similar to (3.1),

$$f(\mathbf{x}) = \sum_{m=0}^{M} \beta_m f_m(\mathbf{x}|\boldsymbol{\Theta}_m)$$

where the functions $f(\mathbf{x}|\boldsymbol{\Theta})$ are base learners. Often, the base learners $f(\cdot)$ are chosen to be 'weak' learners (e.g., shallow CART models, simple linear models) that by themselves have little predictive power, but when combined in an ensemble form a power predictive analytic model. The expansion coefficients $\{\beta_m\}_0^M$ and parameters $\{\boldsymbol{\Theta}_m\}_0^M$ are jointly fit to the data in forward stage-wise manner, where one starts with an initial guess $f_0(\mathbf{x})$ and then for $m = 1, \dots, M$

$$(\beta_m, \boldsymbol{\Theta}_m) = \arg\min_{\beta,\boldsymbol{\Theta}} \sum_{i=1}^{n_1} L\left(y_i, f_{m-1}(\mathbf{x}_i) + \beta f(\mathbf{x}_i|\boldsymbol{\Theta})\right) \tag{3.6}$$

and

$$f_m(\mathbf{x}) = f_{m-1}(\mathbf{x}) + \beta_m f(\mathbf{x}|\boldsymbol{\Theta}_m). \tag{3.7}$$

A technique introduced by Friedman (2001) called gradient boosting approximately solves (3.6) for arbitrary differential loss functions using a general procedure. Specifically, first the function $h(\mathbf{x}|\boldsymbol{\Theta})$ is fit by least-squares

$$\boldsymbol{\Theta}_m = \arg\min_{\boldsymbol{\Theta},\rho} \sum_{i=1}^{n_1} L\left(\tilde{y}_{im} - \rho f(\mathbf{x}_i|\boldsymbol{\Theta})\right)^2 \tag{3.8}$$

where

$$\tilde{y}_{im} = -\frac{\partial L\big(y_{\mathrm{i}}, f(\mathbf{x}_{\mathrm{i}})\big)}{\partial f(\mathbf{x}_i)}\bigg|_{f(\mathbf{x})=f_{m-1}(\mathbf{x})} \tag{3.9}$$

are the working responses, or "pseudo"-residuals. Then, given $f(\mathbf{x}_i|\boldsymbol{\Theta})$, the optimal value of the coefficient β_m is determined

$$\beta_m = \arg\min_{\boldsymbol{\Theta},\rho} \sum_{i=1}^{n_1} L\big(y_i, f_{m-1}(\mathbf{x}_i) + \beta f(\mathbf{x}_i|\boldsymbol{\Theta}_m)\big). \tag{3.10}$$

This general two-step procedure replaces an intractable optimization problem given in (3.6) by one based on least-squares (3.8), followed by a single parameter optimization (3.10) (Bühlmann & Hothorn, 2007).

The logic behind why gradient boosting works is described in (2001), Hastie et al. (2009), and Bühlmann and Hothorn (2007). First, consider a nonparametric approach to numerical optimize $\arg\min_{f(\mathbf{x})} E_{y,\mathbf{x}} L\big(y, f(\mathbf{x})\big)$ in function space. In this case, we consider $f(\mathbf{x})$ evaluated at each point $\mathbf{x}$ to the parameter and thus want to minimize. Theoretically, in function space there are an infinite number of such parameters, but in finite data sets, only a finite number of parameters $\{f(\mathbf{x}_i)\}_1^{n_1}$ exist. Using the boosting technique of function approximation, the approximation is an additive form

$$f^*(\mathbf{x}) = \sum_{m=0}^{M} f_m(\mathbf{x}). \tag{3.11}$$

Similar to before, in (3.11) $f_0(\mathbf{x})$ is an initial guess and $\{f_m(\mathbf{x})\}_1^m$ are incremental boosts (or steps) defined by an optimization method (Natekin & Knoll, 2013). In particular, using a steepest-descent optimization routine,

$$f_m(\mathbf{x}) = -\rho_m g_m(\mathbf{x}) \tag{3.12}$$

with

$$g_m(\mathbf{x}) = E_{y,\mathbf{x}}\left[\frac{\partial L\big(y_i, f(\mathbf{x}_i)\big)}{\partial f(\mathbf{x}_i)}\bigg|_{f(\mathbf{x})=f_{m-1}(\mathbf{x})}\right] \tag{3.13}$$

as the unconstrained gradient (assuming sufficient regularity conditions hold) with

$$f_{m-1}(\mathbf{x}) = \sum_{i=0}^{m-1} f_i(\mathbf{x}).$$

The negative gradient in (3.13) is said to define the steepest-descent direction and the multiplier ρ_m is given by the line search

$$\rho_m = \arg\min_{\rho} E_{y,\mathbf{x}} L\big(y, f_{m-1}(\mathbf{x}) - \rho g_m(\mathbf{x})\big). \tag{3.14}$$

Consider a case where, for a particular loss function and/or base learner, the solution to (3.6) is intractable. Given some approximation $f_{m-1}(\mathbf{x})$, the function $\beta f(\mathbf{x}|\boldsymbol{\Theta}_m)$ in (3.6) and (3.7) can be viewed as the steepest-descent step in (3.12) based on the training data under the constraint that the step direction $f(\mathbf{x}|\boldsymbol{\Theta}_m)$ be a member of the parametrized class of functions $f(\mathbf{x}|\boldsymbol{\Theta})$ (Friedman, 2001). For a finite set of training data $\{y_i, \mathbf{x}_i\}_1^{n_1}$, the data-based analog of the unconstrained gradient (3.13)

$$-g_m(\mathbf{x}_i) = -\frac{\partial L\big(y_i, f(\mathbf{x}_i)\big)}{\partial f(\mathbf{x}_i)}\bigg|_{f(\mathbf{x})=f_{m-1}(\mathbf{x})} \tag{3.15}$$

gives the best steepest-descent step direction $-\mathbf{g}_m = \{-g_m(\mathbf{x}_i)\}_1^{n_1}$ in the n_1-dimensional data space at $f_{m-1}(\mathbf{x})$. The problem, however, is that the data-based unconstrained gradient can only be estimated at the observed values in the training set. Therefore, to increase generalization to non-observed values, Friedman (2001) proposed that by solving the minimization

$$\boldsymbol{\Theta}_m = \arg\min_{\boldsymbol{\Theta},\beta} \sum_{i=1}^{n_1} \left(-g_m(\mathbf{x}_i) - \beta f(\mathbf{x}_i|\boldsymbol{\Theta})\right)^2$$

one obtains the member of the parametrized class $f(\mathbf{x}|\boldsymbol{\Theta})$ that is most highly correlated with $-g_m(\mathbf{x}) \in \mathcal{R}^{n_1}$ over the data distribution. Now, in the steepest-descent routine the constrained negative gradient $f(\mathbf{x}|\boldsymbol{\Theta}_m)$ is used in place of the unconstrained one $-g_m(\mathbf{x}_i)$ (3.15) and the line search is performed using

$$\rho_m = \arg\min_{\rho} \sum_{i=1}^{n_1} L\big(y_i, f_{m-1}(\mathbf{x}_i) - \rho f(\mathbf{x}_i|\boldsymbol{\Theta}_m)\big). \tag{3.16}$$

Finally, the approximation is updated as

$$f_m(\mathbf{x}) = f_{m-1}(\mathbf{x}) + \rho_m f(\mathbf{x}|\boldsymbol{\Theta}_m). \tag{3.17}$$

Although Friedman (2001) proposed the line search step given in (3.16), Bühlmann and Hothorn (2007) showed that for loss functions in the exponential family (e.g., squared-error loss, negative log-likelihood binomial loss) a good functional estimator of $f^*(\mathbf{x})$ can be obtained by removing the line search step, specifying a small learning rate parameter, and increasing the number of boosting iterations. That is, (3.17) is often replaced with

$$f_m(\mathbf{x}) = f_{m-1}(\mathbf{x}) + \nu f(\mathbf{x}|\boldsymbol{\Theta}_m),$$

where ν is the learning rate and is usually a constant, small number (e.g., .01) (Bühlmann & Hothorn, 2007).

3.3.2. Regularized gradient boosting. Generally, best practice in any predictive modeling to prevent overfitting is to define the objective function $J(\boldsymbol{\Theta})$ as the sum of two components, the training loss $L(\boldsymbol{\Theta})$ and the regularization penalty $\Omega(\boldsymbol{\Theta})$,

$$J(\boldsymbol{\Theta}) = L(\boldsymbol{\Theta}) + \Omega(\boldsymbol{\Theta}). \tag{3.18}$$

In the context of gradient boosted models, $L(\boldsymbol{\Theta})$ is a standard loss function such as the

squared error loss for continuous variables or negative log-likelihood binomial loss for binary variables. The added regularization penalty function $\Omega(\boldsymbol{\Theta})$ is a standard regularization function, such as the L2-norm imposed on the regression weights in a linear or logistic regression model or on the leafs of a CART model, to help the train a model to generalize to independent test data (i.e., help prevent overfitting; Bishop, 2006). As Proposition (3.1) shows, for linear models with a squared error loss or negative log-likelihood binomial loss, imposing an L2-norm on the regression weights leads to an equivalent objective function as specifying a $N(\mathbf{0}, \mathbf{I})$ (i.e., normal distribution with mean vector $\mathbf{0}$ and covariance matrix $\mathbf{I}$) prior on the regression weights.

Proposition 3.1. *The objective functions in linear and logistic regression models with L2-norms imposed on the regression weights are equivalent to the objective functions in linear and logistic regression models with Gaussian* $N(\mathbf{0}, \mathbf{I})$ *priors on the regression coefficients* $\boldsymbol{\beta}$.

Proof:

We will first prove the proposition for linear regression models. Adding the L2-norm on the squared error loss yields,

$$J(\mathbf{y}|\mathbf{x}, \boldsymbol{\beta}, \lambda) = \frac{1}{2}\sum_{i=1}^{n_1}(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2 + \frac{\lambda}{2}\sum_{j=1}^{p}\beta_j^2, \tag{3.19}$$

where λ is the regularization parameter. If we negate the objective function in (3.19) and then exponentiate the result, we obtain

$$\begin{aligned} J(\mathbf{y}|\mathbf{x}, \boldsymbol{\beta}, \lambda) &= e^{-\frac{1}{2}\sum_{i=1}^{n_1}(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2 - \frac{\lambda}{2}\sum_{j=1}^{p}\beta_j^2} \\ &= e^{-\frac{1}{2}\sum_{i=1}^{n_1}(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2} e^{-\frac{\lambda}{2}\sum_{j=1}^{p}\beta_j^2}. \end{aligned}$$

From the theory of linear models, we know that the distribution of the response y_i is normal with mean given by the linear function $\mathbf{x}_i'\boldsymbol{\beta}$ and constant variance σ^2, or $N(\mathbf{x}_i'\boldsymbol{\beta}, \sigma^2)$ (Rao, 1976). Moreover, we know that a $N(\mathbf{0}, \mathbf{I})$ Gaussian prior on $\boldsymbol{\beta}$ is proportional to $e^{-\frac{1}{2}\sum_{j=1}^{p}\beta_j^2}$. Thus, the posterior distribution under of the linear model $N(\mathbf{x}_i'\boldsymbol{\beta}, \sigma^2)$ under a $N(\mathbf{0}, \mathbf{I})$ prior on $\boldsymbol{\beta}$ is

$$J^*(\mathbf{y}|\mathbf{x}, \boldsymbol{\beta}, \sigma^2) = \prod_{i=1}^{n_1} N(\mathbf{x}_i'\boldsymbol{\beta}, \sigma^2) \times N(\mathbf{0}, \mathbf{I})$$

$$\propto e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n_1}(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2} e^{-\frac{1}{2}\sum_{j=1}^{p}\beta_j^2}.$$

Comparing $J^*(\mathbf{y}|\mathbf{x}, \boldsymbol{\beta}, \sigma^2)$ to $J(\mathbf{y}|\mathbf{x}, \boldsymbol{\beta}, \lambda)$, the objective functions are equivalent by setting $\lambda = \frac{1}{\sigma^2}$, that is, the regularization parameter in $J(\mathbf{y}|\mathbf{x}, \boldsymbol{\beta}, \lambda)$ is equal to the precision parameter in $J^*(\mathbf{y}|\mathbf{x}, \boldsymbol{\beta}, \sigma^2)$.

For logistic regression models, adding the L2-norm to negative log-likelihood binomial loss gives

$$J(\mathbf{y}|\mathbf{x}, \boldsymbol{\beta}, \lambda) = -\sum_{i=1}^{n_1}\left[y_i \log\left(\frac{1}{1 + e^{-\mathbf{x}_i'\boldsymbol{\beta}}}\right) + (1 - y_i)\log\left(\frac{1}{1 + e^{\mathbf{x}_i'\boldsymbol{\beta}}}\right)\right] + \frac{\lambda}{2}\sum_{j=1}^{p}\beta_j^2.$$

If we denote the binomial likelihood loss function as $f(\mathbf{y}|\mathbf{x})$, we see that applying a $N(\mathbf{0}, \mathbf{I})$ Gaussian prior on $\boldsymbol{\beta}$ gives the posterior distribution

$$\prod_{i=1}^{n_1} f(\mathbf{y}|\mathbf{x}) \times N(\mathbf{0}, \mathbf{I}) \propto \prod_{i=1}^{n_1}\left(\frac{1}{1 + e^{-\mathbf{x}_i'\boldsymbol{\beta}}}\right)^{y_i}\left(\frac{1}{1 + e^{\mathbf{x}_i'\boldsymbol{\beta}}}\right)^{1-y_i} e^{-\frac{\lambda}{2}\sum_{j=1}^{p}\beta_j^2}. \qquad (3.20)$$

Negating (3.20) and taking the log,

$$J^*(\mathbf{y}|\mathbf{x}, \boldsymbol{\beta}) = -\log\left[\prod_{i=1}^{n_1}\left(\frac{1}{1 + e^{-\mathbf{x}_i'\boldsymbol{\beta}}}\right)^{y_i}\left(\frac{1}{1 + e^{\mathbf{x}_i'\boldsymbol{\beta}}}\right)^{1-y_i} e^{-\frac{\lambda}{2}\sum_{j=1}^{p}\beta_j^2}\right]$$

$$= -\sum_{i=1}^{n_1}\left[y_i \log\left(\frac{1}{1+e^{-\mathbf{x}_i'\boldsymbol{\beta}}}\right) + (1-y_i)\log\left(\frac{1}{1+e^{\mathbf{x}_i'\boldsymbol{\beta}}}\right)\right] + \frac{\lambda}{2}\sum_{j=1}^{p}\beta_j^2,$$

we see that $J(\mathbf{y}|\mathbf{x}, \boldsymbol{\beta}, \lambda)$ is equivalent to $J^*(\mathbf{y}|\mathbf{x}, \boldsymbol{\beta})$. This completes the proof. ■

A corollary of Proposition (3.1) is that the least-squares estimator $\widehat{\boldsymbol{\beta}}_{\text{LS}}$ of $\boldsymbol{\beta}$ for $J(\mathbf{y}|\mathbf{x}, \boldsymbol{\beta}, \lambda)$ equals the maximum a posteriori (MAP) estimator $\widehat{\boldsymbol{\beta}}_{\text{MAP}}$ of $\boldsymbol{\beta}$ for $J^*(\mathbf{y}|\mathbf{x}, \boldsymbol{\beta}, \sigma^2)$ and similarly the maximum likelihood estimator $\widehat{\boldsymbol{\beta}}_{\text{MLE}}$ of $\boldsymbol{\beta}$ for $J(\mathbf{y}|\mathbf{x}, \boldsymbol{\beta}, \lambda)$ equals the MAP estimator $\widehat{\boldsymbol{\beta}}_{\text{MAP}}$ of $\boldsymbol{\beta}$ for $J^*(\mathbf{y}|\mathbf{x}, \boldsymbol{\beta})$.

A more recent framework of gradient boosted, eXtreme gradient boosted (XBG or XGBoost), developed by Chen and Guestrin (2016), makes use of gradient boosted learners with objective functions of the form in (3.18). XGBoost was developed as a large-scale regularized gradient boosting framework that incorporates linear boosters via methods similar to the generalized additive models (GAMs) proposed in Friedman, Hastie and Tibshirani (2000) and tree boosters similar to Friedman's original gradient boosting framework (2001). Specifically, the predicted value $\hat{y}_i^{(M)}$ with M additive functions is

$$\hat{y}_i^{(M)} = f_0(\mathbf{x}_i) + \sum_{m=1}^{M} \nu_m f_m(\mathbf{x}_i), \quad f_m \in \mathcal{F}$$

and the learners $f_m(\mathbf{x}_i)$ are selected based on optimizing the objective function

$$J(\boldsymbol{\Theta}) = \sum_{i=1}^{n_1} L(y_i, \hat{y}_i) + \sum_{m=1}^{M} \Omega(f_m)$$

$$= \sum_{i=1}^{n_1} L\left(y_i, \hat{y}_i^{(M-1)} + \nu_M f_M(\mathbf{x}_i)\right) + \sum_{m=1}^{M} \Omega(f_m),$$

where ν_M is the learning rate for learner M. Note, the learning rate parameter here is denoted with a subscript to indicate that the learning rate can change during the learning process (e.g., using a polynomial decaying rate).

Under the XGBoost framework, linear boosters are important when the underlying function is strictly linear. In this case, the linear boosters can be implemented using the general functional gradient descent (FGD) algorithm for gradient boosted linear models given in Bühlmann and Hothorn (2007) and presented in Algorithm (3.2). Bühlmann and Hothorn's (2007) FGD algorithm is

Algorithm 3.2. Functional Gradient Descent for Boosted Models

Requires: Training data $\mathcal{D}_{\text{train}} = \{y_i, \mathbf{x}_i\}_1^{n_1}$

1. *initialize* $f_0(\cdot) \leftarrow \arg\min_c n_1^{-1} \sum_{i=1}^{n_1} L(y_i, c)$
2. **for** $m \leftarrow 1$ to M **do**
3. $\quad \mathbf{u}^{(m)} \leftarrow -\frac{\partial}{\partial f} L(\mathbf{y}, f)\Big|_{f=f_{m-1}(\mathbf{x})}$
4. $\quad \boldsymbol{\Theta}_m \leftarrow \arg\min_{\boldsymbol{\Theta}} \sum_{i=1}^{n_1} J\left(u_i^{(m)}, f_m(\mathbf{x}_i|\boldsymbol{\Theta})\right)$
5. $\quad f_m(\mathbf{x}) \leftarrow f_{m-1}(\mathbf{x}) + \nu f(\mathbf{x}|\boldsymbol{\Theta}_m)$
6. **end for**

similar to Friedman's (2001) gradient boosted algorithm, except the line search step is omitted in between Lines 4 and 5 due to the reasons given above. For linear boosters, a componentwise linear least squares variable selection procedure is implemented at each boosting iteration. If we consider the base procedure at any given boosting round m, then the 'best' prediction is

$$f_m(\mathbf{x}) = \sum_{i=1}^{n_1} x_{is^*} \hat{\beta}_{s^*},$$

where $\hat{\beta}_{s^*}$ is

$$\hat{\beta}_{s^*} = \frac{\sum_{i=1}^{n_1} x_{i,s^*} u_{i,s^*}}{\sum_{i=1}^{n_1} x_{i,s^*}^2}$$

and s^* represents the index of the selected variable in $\mathbf{x}$ and $\widehat{\boldsymbol{\beta}}$ that minimizes the squared error,

$$s^* = \arg\min_{1 \le j \le p} \sum_{i=1}^{n_1} \left(u_i - x_{ij}\hat{\beta}_j\right)^2.$$

In other words, s^* is the index that represents the variable (and corresponding regression coefficient) that result in the closest prediction to the negative gradient $\mathbf{u}$ at boosting round m. The regression coefficient estimate can be updated as

$$\hat{\beta}_m = \hat{\beta}_{m-1} + \nu_m \hat{\beta}_{s_m^*}.$$

In the case of continuous labels, where the squared-error loss function is used and an L2-norm is imposed on the regression weights, the initial function $f_0(\cdot)$ is the sample mean $\bar{y}$ in Line 1 of Algorithm (3.2). The negative gradient (Line 3) is simply the residuals,

$$-\frac{\partial L(y_i, \hat{y}_i)}{\partial \hat{y}_i} = \frac{\partial}{\partial \hat{y}_i}\left[\frac{1}{2}(y_i - \hat{y}_i)^2\right]$$

$$= (y_i - \hat{y}_i)$$

and the objective function $J(\cdot)$ that needs to be minimized (Line 4) is given is the squared-error loss with the L2-norm. For this minimization, the componentwise linear least squares variable selection procedure described above is implemented. In the case of binary labels, the loss function used is still the negative log-likelihood binomial loss, however, there the function is reparametrized as

$$L(y_i, \hat{y}_i) = \log\left(1 + e^{-2\tilde{y}_i \hat{y}_i}\right) \tag{3.21}$$

where $\tilde{y}_i = 2y_i - 1$ and $y_i \in \{-1,1\}$ or

$$L(y_i, \hat{y}_i) = \log\left(1 + e^{-2(2y_i - 1)\hat{y}_i}\right) \tag{3.22}$$

where $y_i \in \{0,1\}$. In (3.21) and (3.22), $\hat{y}_i$ is the predicted label for the ith sample. To obtain this result, if we also reparametrize $P(Y = 1) = p$ as

$$p(x_i) = \frac{e^{\hat{y}_i}}{e^{\hat{y}_i} + e^{-\hat{y}_i}}$$

$$= \frac{1}{1 + e^{-2\hat{y}_i}},$$

then

$$1 - p(x_i) = \frac{e^{-2\hat{y}_i}}{1 + e^{-2\hat{y}_i}}.$$

Rewriting the likelihood using indicator an indicator function, we obtain

$$\left(\frac{1}{1 + e^{-2\hat{y}_i}}\right)^{I(y_i=1)} \left(\frac{e^{-2\hat{y}_i}}{1 + e^{-2\hat{y}_i}}\right)^{I(y_i=-1)}.$$

The negative log-likelihood then becomes

$$L(y_i, \hat{y}_i) = -\log\left[\left(\frac{1}{1 + e^{-2\hat{y}_i}}\right)^{I(y_i=1)} \left(\frac{e^{-2\hat{y}_i}}{1 + e^{-2\hat{y}_i}}\right)^{I(y_i=-1)}\right]$$

$$= I(y_i = 1)\log\left(1 + e^{-2\hat{y}_i}\right) + I(y_i = -1)\left(\log\left(1 + e^{-2\hat{y}_i}\right) + 2\hat{y}_i\right).$$

Note that $\left(\log\left(1 + e^{-2\hat{y}_i}\right) + 2\hat{y}_i\right) = \log\left(1 + e^{2\hat{y}_i}\right)$ since

$$\frac{1 + e^{a}}{1 + e^{-a}} = \frac{e^{a}(e^{-a} + 1)}{1 + e^{-a}}$$

$$= e^{a}$$

which implies

$$1 + e^{a} = (1 + e^{-a})e^{a}$$

$$\log(1 + e^{a}) = \log(1 + e^{-a}) + a.$$

Finally, the loss function becomes

$$L(y_i, \hat{y}_i) = I(y_i = 1)\log\left(1 + e^{-2\hat{y}_i}\right) + I(y_i = -1)\log\left(1 + e^{2\hat{y}_i}\right) \tag{3.23}$$

and we can rewrite (3.23) as

$$\log\left(1 + e^{-2\tilde{y}_i\hat{y}_i}\right) = \begin{cases} \log\left(1 + e^{2\hat{y}_i}\right), & y_i = -1 \\ \log\left(1 + e^{-2\hat{y}_i}\right), & y_i = 1 \end{cases}$$

or equivalently,

$$\log\left(1 + e^{-2(2y_i-1)\hat{y}_i}\right) = \begin{cases} \log\left(1 + e^{2\hat{y}_i}\right), & y_i = 0 \\ \log\left(1 + e^{-2\hat{y}_i}\right), & y_i = 1. \end{cases}$$

The initial function $f_0(\cdot)$ for modeling binary labels, using the negative log-likelihood binomial loss, is set to 0 and the initial probabilities $p_0(x)$ are set to $\frac{1}{2}$ (Dettling & Bühlmann, 2002). Instead of using the negative gradient directly as in Bühlmann and Hothorn's (2007) BinomialBoosting algorithm, a Newton-type update (Friedman et al., 2000) is used to calculate the working response as

$$z_i = \frac{y_i - p(x_i)}{p(x_i)\left(1 - p(x_i)\right)},$$

where the numerator is the partial derivative of the expected log-likelihood with respect to the update $\Delta(\hat{y}_i)$ (with the constant 2 omitted),

$$\frac{\partial E\left(\hat{y}_i + \Delta(\hat{y}_i)\right)}{\partial \Delta(\hat{y}_i)} = \frac{\partial E\left[2y\left(\hat{y}_i + \Delta(\hat{y}_i)\right) - \log\left(1 + e^{2\left(\hat{y}_i + \Delta(\hat{y}_i)\right)}\right)\right]}{\partial \Delta(\hat{y}_i)},$$

and the denominator is the second partial derivative with respect to the update $\Delta(x)$ (with the constant -4 omitted),

$$\frac{\partial^2 E\left(\hat{y}_i + \Delta(\hat{y}_i)\right)}{\partial \Delta(\hat{y}_i)^2} = \frac{\partial^2 E\left[2y\left(\hat{y}_i + \Delta(\hat{y}_i)\right) - \log\left(1 + e^{2\left(\hat{y}_i + \Delta(\hat{y}_i)\right)}\right)\right]}{\partial \Delta(\hat{y}_i)^2}. \tag{3.24}$$

The objective function $J(\cdot)$ that needs to be minimized (Algorithm [3.2], Line 4) is a weighted least-squares regression of z_i to x_i using weights w_i, where w_i is proportional to the second partial derivative in (given by the second partial derivative in (3.24), or

$$w_i = p(x_i)\big(1 - p(x_i)\big).$$

In Friedman et al. (2000), the learning rate is set to $\frac{1}{2}$ to make the update a true Newton step (Line 5; $\nu = \frac{1}{2}$), but in the XGB implementation the learning rate is a hyperparameter that needs to be specified. Similar to the case with continuous labels, the L2-norm can be added to the objective function to help prevent overfitting and the minimization can be accomplished using the componentwise linear least squares variable selection procedure.

For tree boosters, Chen and Guestrin (2016) approximate an objective function in a general setting using the second order Taylor series expansion at the mth boosting iteration by

$$J(\boldsymbol{\Theta})^{(m)} = \sum_{i=1}^{n_1}\left[L\left(y_i, \hat{y}_i^{(m-1)}\right) + g_i f_m(\mathbf{x}_i) + \frac{1}{2} h_i f_m^2(\mathbf{x}_i)\right] + \Omega(f_m) + C,$$

where $g_i = \partial_{\hat{y}_i^{(m-1)}} L\left(y_i, \hat{y}_i^{(m-1)}\right)$ and $h_i = \partial^2_{\hat{y}_i^{(m-1)}} L\left(y_i, \hat{y}_i^{(m-1)}\right)$ are the first and second order gradient statistics of a loss function and C is a constant term that does not involve $f_m(\mathbf{x}_i)$. For squared error loss $L(y_i, \hat{y}_i) = \sum_{i=1}^{n_1}(y_i - \hat{y}_i)^2$, the first and second order gradient statistics are

$$g_i = -2(y_i - \hat{y}_i)$$

and

$$h_i = 2$$

respectively, and for the negative log-likelihood of the binomial loss

$$L(y_i, \hat{y}_i) = -\sum_{i=1}^{n_1} y_i \log\left(1 + e^{-\hat{y}_i}\right) - \sum_{i=1}^{n_1} (1 - y_i) \log\left(1 + e^{\hat{y}_i}\right)$$

the first and second order gradient statistics are

$$g_i = \frac{e^{\hat{y}_i} - y_i e^{\hat{y}_i} - y_i}{1 + e^{\hat{y}_i}}$$

and

$$h_i = \frac{e^{\hat{y}_i}}{(1 + e^{\hat{y}_i})^2},$$

respectively. Omitting all constant terms not involving $f_m(\mathbf{x}_i)$, we obtain

$$J(\boldsymbol{\Theta})^{(m)} = \sum_{i=1}^{n_1} \left[g_i f_m(\mathbf{x}_i) + \frac{1}{2} h_i f_m^2(\mathbf{x}_i) \right] + \Omega(f_m). \tag{3.25}$$

By reparametrizing the tree as a score within a leaf j, $I_j = \{i | q(\mathbf{x}_i) = j\}$, (3.25) can be rewritten to include the L2-norm on the leaf scores as

$$\begin{aligned} J(\boldsymbol{\Theta})^{(m)} &= \sum_{i=1}^{n_1} \left[g_i w_j + \frac{1}{2} h_i w_j^2 \right] + \gamma M + \frac{1}{2} \lambda \sum_{j=1}^{M} w_j^2 \\ &= \sum_{j=1}^{M} \left[\left(\sum_{i \in I_j} g_i \right) w_j + \frac{1}{2} \left(\sum_{i \in I_j} h_i + \lambda \right) w_j^2 \right] + \gamma M \end{aligned}$$

where w_j is the score in the jth leaf, M is the number of trees in the model, γ is the complexity cost by introducing an additional leaf in the model, and λ is the regularization hyperparameter. With a fixed structure of the tree is $q(\mathbf{x})$ is fixed, the derivative of $J(\boldsymbol{\Theta})^{(m)}$ with respect to w_j is

$$\frac{\partial J(\boldsymbol{\Theta})^{(m)}}{\partial w_j} = \sum_{i \in I_j} g_i + \left(\sum_{i \in I_j} h_i + \lambda \right) w_j \tag{3.26}$$

and therefore the optimal weight in leaf j w_j^* is given by

$$w_j^* = -\frac{\sum_{i\in I_j} g_i}{\sum_{i\in I_j} h_i + \lambda}. \tag{3.27}$$

Substituting (3.27) into (3.26), we see that the optimal objective function value is

$$J(\boldsymbol{\Theta})^{(m)} = \sum_{j=1}^{M}\left[\left(\sum_{i\in I_j} g_i\right)\left(-\frac{\sum_{i\in I_j} g_i}{\sum_{i\in I_j} h_i + \lambda}\right) + \frac{1}{2}\left(\sum_{i\in I_j} h_i + \lambda\right)\left(-\frac{\sum_{i\in I_j} g_i}{\sum_{i\in I_j} h_i + \lambda}\right)^2\right] + \gamma M$$

$$= -\frac{1}{2}\sum_{j=1}^{M}\frac{\left(\sum_{i\in I_j} g_i\right)^2}{\sum_{i\in I_j} h_i + \lambda} + \gamma M.$$

This objective function measures how 'good' a tree structure is, with smaller scores indicating a better structure. Since it is normally impossible to enumerate all possible tree structures to optimize $J(\boldsymbol{\Theta})$, a greedy algorithm is used that starts with a single leaf and iteratively adds branches to the tree (Chen & Guestrin, 2016). Specifically, for each leaf node of the tree, the algorithm tries to add a split and the change of objective after adding the split is given by the information gain

$$J(\boldsymbol{\Theta})_{\text{split}} = \frac{1}{2}\left[\frac{\left(\sum_{i\in I_L} g_i\right)^2}{\sum_{i\in I_L} h_i + \lambda} + \frac{\left(\sum_{i\in I_R} g_i\right)^2}{\sum_{i\in I_R} h_i + \lambda} - \frac{(\sum_{i\in I} g_i)^2}{\sum_{i\in I} h_i + \lambda}\right] - \gamma,$$

where $I = I_L \cup I_R$ is the union of the left and right instance sets after the split.

3.3.3. Application to missing data. The use of gradient boosted for imputing missing data is limited as far as we know. One study, however, examined the performance of using gradient boosting for single imputation of missing predictors in high-dimensional classification problems. In this study, Wang and Feng (2010) implemented boosting as a form of additive logistic regression and found that imputation methods based on boosting

outperformed naïve methods such as mean imputation in terms of both training and testing sensitivity, specificity, and prediction error.

In some cases, tree boosting algorithms are implemented to handle missing values using a surrogate variable approach described in Hastie et al. (2009). The surrogate variable approach can be summarized as two steps: (1) Find the best split among predictor variables using only the observed observations, and (2) After choosing the best predictors and split point, form a list of surrogate predictors and split points that best mimic the splits obtained in (1). When applied to a testing data set, the surrogate splits from (2) are used in order if the primary splitting predictors in (1) are missing. In practice, MI has been shown to lead to higher predictive accuracy than the surrogate approach in CART models in part because MI profits from variance reductions by averaging over the M multiply imputed estimates (Feelders, 1999).

3.4. Proposed Multiple Imputation Algorithm

The MI algorithm proposed in this study combines the Bayesian bootstrap and FCS MI (denoted as BB-FCS) using regularized gradient boosted models to impute missing data and estimate the posterior distributions of model parameters (e.g., regression coefficients; functions of regression coefficients [e.g., indirect effects]). With a slight change of notation, let $\boldsymbol{\mathcal{D}} = \{\boldsymbol{\mathcal{D}}_{\text{obs}}, \boldsymbol{\mathcal{D}}_{\text{mis}}\}$ denote the $n \times p$ sample data set consisting of both observed and missing values. A general overview of the algorithm is as follows:

1. Generate a Bayesian bootstrap sample from $\boldsymbol{\mathcal{D}}$ denoted as $\boldsymbol{\mathcal{D}}^{*b}$.
2. Create M imputed data sets $\{\boldsymbol{\mathcal{D}}_m^{*b}\}_{m=1}^{M}$ using FCS MI with regularized gradient boosting machine learning models.
3. For each of the M imputed data sets in Step (2), estimate the regression model

parameters using complete-data methods and calculate the indirect effect for a total of M estimates, $\{\widehat{\boldsymbol{\theta}}_m^{*b}\}_{m=1}^{M}$.

4. Calculate the point estimate of the multiply imputed indirect estimate of the data in Step (3) as $\overline{\widehat{\boldsymbol{\theta}}}^{*b} = \frac{1}{M}\sum_{m=1}^{M}\widehat{\boldsymbol{\theta}}_m^{*b}$.

5. Repeat steps (1) – (4) B times to obtain B indirect effect estimates $\left\{\overline{\widehat{\boldsymbol{\theta}}}^{*b}\right\}_{b=1}^{B}$.

6. Calculate the point estimate $\overline{\widehat{\boldsymbol{\theta}}}^{**}$ using the mean $\frac{1}{B}\sum_{b=1}^{B}\overline{\widehat{\boldsymbol{\theta}}}^{*b}$ or median $P\left(\overline{\widehat{\boldsymbol{\theta}}}^{*b} \leq q\right) = P\left(\overline{\widehat{\boldsymbol{\theta}}}^{*b} \geq q\right) = \frac{1}{2}$, where q is the median. Also, calculate the $100(1-\alpha)\%$ confidence intervals with the lower limit $\overline{\widehat{\boldsymbol{\theta}}}_{\alpha/2}^{**}$ as $\text{percentile}_{\alpha/2}\left(\overline{\widehat{\boldsymbol{\theta}}}^{*b\cdot}\right)$ and upper limit $\overline{\widehat{\boldsymbol{\theta}}}_{1-\alpha/2}^{**}$ as $\text{percentile}_{1-\alpha/2}\left(\overline{\widehat{\boldsymbol{\theta}}}^{*b\cdot}\right)$.

Figure (3.5) presents a graphical depiction of the algorithm described above. The multiply imputed estimates for each Bayesian bootstrap sample are denoted as $\overline{\widehat{\boldsymbol{\theta}}}^{*b}$, where b indicates the bth Bayesian boostrapped sample. Algorithm (3.3) presents the pseudocode. As discussed in Chapter 2, a two-stage Bayesian bootstrapping sampling scheme is implemented where B bootstrap samples are drawn of size n_2 (Lines 1 and 2). The convergence criteria ρ in Line 9 is based on Stekhoven and Bühlmann (2012) who use a random forest algorithm to impute missing data. Specifically, ρ is updated as convergent (or true) when both the continuous imputed variables and binary imputed variables are greater than or equal to for the first time with respect to both variable types or the maximum number of iterations is reached (e.g., 5-10). Here, the difference in set of continuous variables $\mathbf{C}$ is defined as

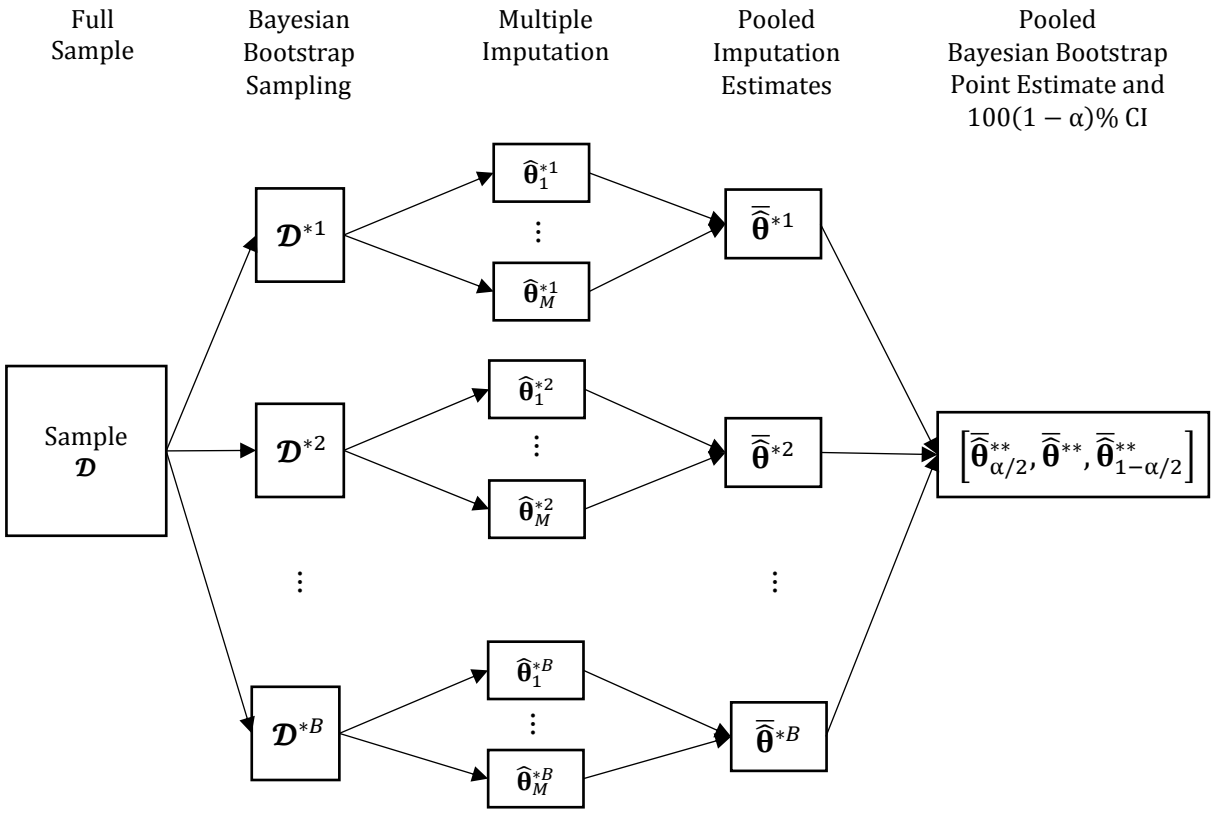


Figure 3.5. A graphical depiction of the Bayesian bootstrapped fully-conditional specification multiple imputation algorithm.

$$\Delta_{\mathbf{C}} = \frac{\sum_{j \in \mathbf{C}} \left(\dot{\mathbf{X}}_{\text{new}} - \dot{\mathbf{X}}_{\text{old}}\right)^2}{\sum_{j \in \mathbf{C}} \left(\dot{\mathbf{X}}_{\text{new}}\right)^2}$$

and for the set of binary variables **D** as

$$\Delta_{\mathbf{D}} = \frac{\sum_{j \in \mathbf{D}} I\left(\dot{\mathbf{X}}_{\text{new}} \neq \dot{\mathbf{X}}_{\text{old}}\right)}{n_{\text{mis}}},$$

where n_{mis} is the number of missing values in the binary variables.

The logic behind the Bayesian bootstrap, MI combination comes from the purpose of inference with missing data using MI. Recall in Chapter 2, for inference in MI we are interested in the observed data posterior distribution of the parameters given the

Algorithm 3.3. BB-FCS Multiple Imputation Using Gradient Boosted Learners

Require: $\mathbf{X}_{n\times p} = (\mathbf{X}_{\text{obs}}, \mathbf{X}_{\text{mis}})$ matrix sorted by missingness on p columns

for $b \leftarrow 1$ to B
  idx $\leftarrow$ draw Bayesian bootstrap sample of size n_2 $\{1{:}\,n\}$
  $\boldsymbol{\mathcal{D}}^{*b} \leftarrow \mathbf{X}$[idx, :]
  if all rows($\boldsymbol{\mathcal{D}}^{*b}$) observed
    $\widehat{\boldsymbol{\theta}}^{*b} \leftarrow f_{\text{SEM}}(\boldsymbol{\mathcal{D}}^{*b})$
  else
    $\mathbf{X}^{*b}_{\text{mis}} \leftarrow$ random draws from $\mathbf{X}^{*b}_{\text{obs}}$
    for $m \leftarrow 1$ to M **do**
      while not ρ **do**
        for $j \leftarrow 1$ to p **do**
          $i_{\text{obs}} \leftarrow \{\text{indices observed data in original } \mathbf{x}^{*b}_j\}$
          $i_{\text{mis}} \leftarrow \{\text{indices missing data in original } \mathbf{x}^{*b}_j\}$
          if length$(\mathbf{x}^{*b}_{i_{\text{mis}}(j)}) = 0$
            pass
          else
            $f^{\text{train}}_{\text{GBM}} \leftarrow \text{Train } f_{\text{GBM}}\left(\dot{\mathbf{x}}^{*b}_{i_{\text{obs}}(j)} \middle| \dot{\mathbf{X}}^{*b}_{i_{\text{obs}}(-j)}\right)$
            $\dot{\boldsymbol{\mathcal{D}}}^{*b}_{i_{\text{mis}}(j)} \leftarrow \text{Test } f^{\text{train}}_{\text{GBM}}\left(\dot{\mathbf{X}}^{*b}_{i_{\text{mis}}(-j)}\right)$
          end if
        end for
        update ρ
      end while
      $\widehat{\boldsymbol{\theta}}^{*bm} \leftarrow f_{\text{SEM}}\left(\dot{\boldsymbol{\mathcal{D}}}^{*b}\right)$
    end for
    $\overline{\widehat{\boldsymbol{\theta}}}^{*b} \leftarrow \frac{1}{M}\sum_{m=1}^{M} \widehat{\boldsymbol{\theta}}^{*bm}$
  end if
end for
sort $\overline{\widehat{\boldsymbol{\theta}}}^{*b\cdot}$
$\left[\overline{\widehat{\boldsymbol{\theta}}}^{**}_{\alpha/2} \leftarrow \text{percentile}_{\alpha/2}\left(\overline{\widehat{\boldsymbol{\theta}}}^{*b}\right), \overline{\widehat{\boldsymbol{\theta}}}^{**} \leftarrow \text{point}\left(\overline{\widehat{\boldsymbol{\theta}}}^{*b}\right), \overline{\widehat{\boldsymbol{\theta}}}^{**}_{1-\alpha/2} \leftarrow \text{percentile}_{1-\alpha/2}\left(\overline{\widehat{\boldsymbol{\theta}}}^{*b}\right)\right]$

observed data,

$$P(\boldsymbol{\theta}|\mathbf{X}_{\text{obs}}) = \int P(\boldsymbol{\theta}|\mathbf{X}_{\text{obs}}, \mathbf{X}_{\text{mis}}) P(\mathbf{X}_{\text{mis}}|\mathbf{X}_{\text{obs}}) d\mathbf{X}_{\text{mis}}. \tag{3.28}$$

The algorithm described above approximates (3.28). Here, $P(\mathbf{X}_{\text{mis}}|\mathbf{X}_{\text{obs}})$ is a regularized gradient boosted model used to create imputations $\dot{\mathbf{X}}_{\text{mis}}$ of the missing data $\mathbf{X}_{\text{mis}}$. After

imputations are drawn, the $P(\boldsymbol{\theta}|\mathbf{X}_{\text{obs}}, \mathbf{X}_{\text{mis}})$ represents a complete-data method that is used to calculate the regression parameters and indirect effect from the pseudo-complete data. This process is repeated for B Bayesian bootstrapped samples, with M multiply imputed data sets in each sample, and each imputed data set a complete-data method is applied. Within each sample, the pooled multiple imputation estimate is one approximation to (3.28) and the set of B multiple imputation estimates is the posterior distribution of the indirect effect we are interested in. We can calculate the point estimate of the posterior distribution to obtain an estimate of (2.15) and also the variance to estimate (2.16) or calculate confidence intervals directly.

To account for both linear and nonlinear effects in substantive models using the same imputation framework, we propose using regularized gradient boosted imputation models in which the learners are selected based on the underlying substantive model. For imputing missing data in which the endogenous variable in a substantive model is predicted by only linear effects, we propose using linear boosters, whereas, for imputing missing data in which the substantive model also contains nonlinear predictors, we propose using tree boosters. The loss function for each gradient boosted model should be appropriately selected based on the data to be imputed. For instance, continuous normally distributed variables that contain missing data can be imputed under a squared-error loss and binary variables that contain missing data can be imputed under a negative log-likelihood binomial loss.

To demonstrate the logic behind the choice of linear versus tree boosters, consider the following examples. In one extreme case, suppose we have a linear model

$$Y = 1 + 1.5X + \zeta_Y,$$

where $\zeta_Y \sim N(0,1)$ and all Y values above $X = 0$ are missing. Figure (3.6) shows the

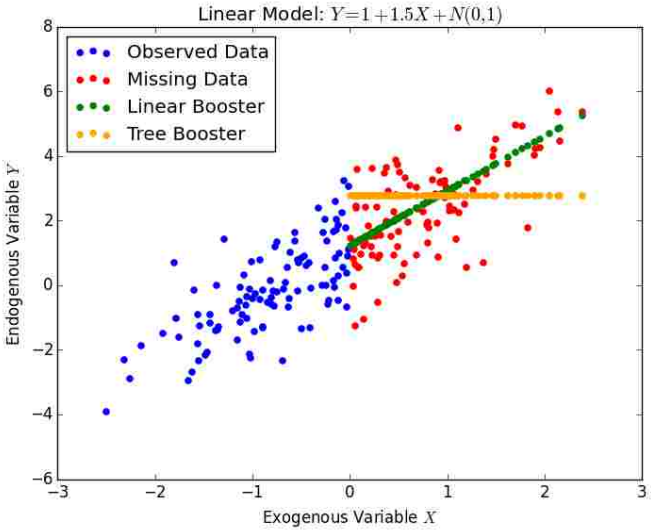


Figure 3.6. Scenario in which imputations are extrapolated outside the range of the observed data and the substantive model contains linear effects.

results of linear boosters and tree boosters to predict the missing Y values above $X = 0$. Clearly, the tree booster is inadequate in this situation because tree-based models are unable to extrapolate outside the range of the observed training data. Although the linear boosters were able to accurately capture the linear relationship between X and Y above $X = 0$, the method as is will not be useful for multiple imputation due to its deterministic form. To make the predictions more stochastic, we propose two sources of randomness. First, stochastic subsampling is used to train a model on a random subset of the training data. Compared to regular gradient boosting, stochastic gradient boosting (Friedman, 2002) has shown to be computationally more efficient and lead to more accurate predictions. Furthermore, Friedman found that subsampling 50% of the data resulted in

accurate predictions with low variance. Second, we propose adding a normal random error to each imputed value, where the variance of the random variable is calculated from the training data error variance,

$$\hat{\sigma}^2_{\text{train}} = \frac{1}{n_1}\sum_{i=1}^{n_1}(y_i - \hat{y}_i)^2,$$

Note, $\hat{\sigma}^2_{\text{train}}$ is the maximum likelihood estimator of σ^2_{train}; we use the maximum likelihood estimator here to be consistent with other machine learning research (Bishop, 2006), but the unbiased estimator could be used here as well without loss of generality.

Continuing with the same example, Figure (3.7) shows the results of our imputation procedure with the two sources of randomness from 10 multiple imputations. The plot shows that the stochastic processes help generate plausible

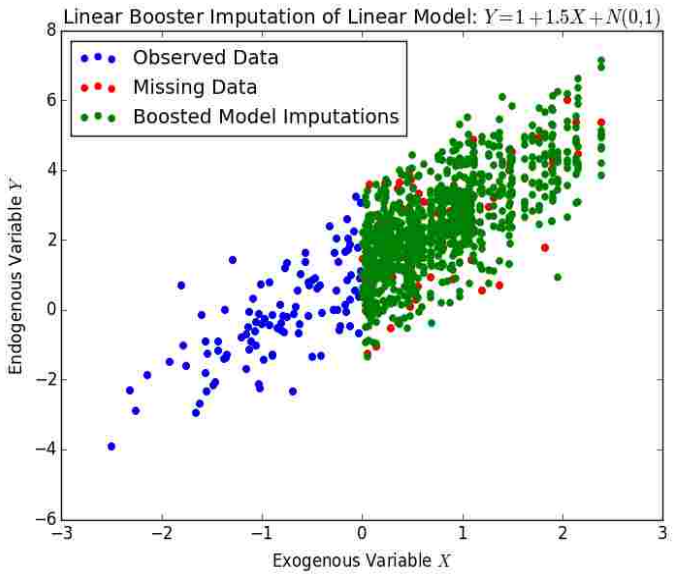


Figure 3.7. Ten rounds of multiple imputation using a linear booster with stochastic subsampling and added Gaussian noise in a substantive model with linear effects.

imputed values.

Consider another linear model,

$$Y = 1 + 0.1X + 0.9X^2 + \zeta_Y,$$

where $\zeta_Y \sim N(0, 1)$ and Y is simulated to have 25% missingness based on a MCAR mechanism. Figure (3.8) shows the results of linear boosters and tree boosters to predict the missing Y values based on only X as a feature (and a vector of ones for the linear booster). We can see that the linear booster based only the feature X, and not X^2 is unable to capture the nonlinearities of Y. Although this is to be expected, the purpose is to demonstrate how tree-based models can automatically capture nonlinearities, with and without explicit interaction features as input. Similar to the linear booster, we can introduce a form of randomness to make the predictions less deterministic for tree boosters. Figure (3.9) shows the results from 10 imputations using 50% subsampling. We

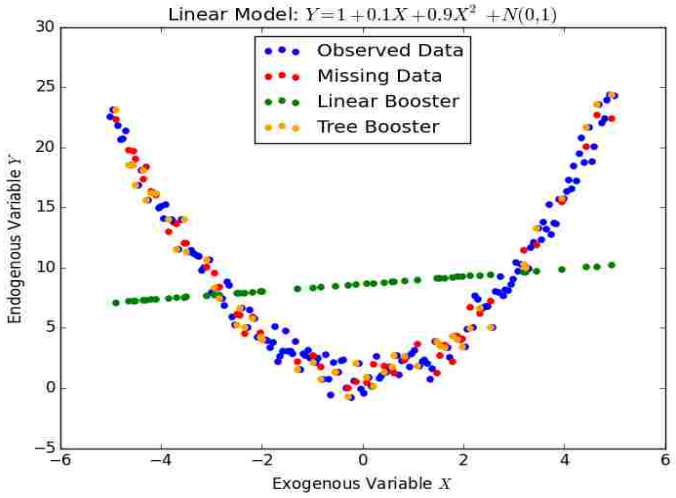


Figure 3.8. Scenario in which imputations are interpolated outside the range of the observed data and the substantive model contains nonlinear effects.

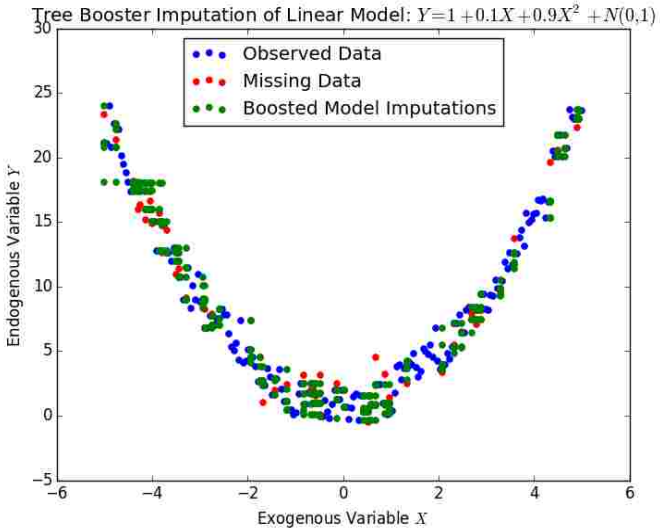


Figure 3.9. Ten rounds of multiple imputation using a tree booster with stochastic subsampling in a substantive model with nonlinear effects.

can see that the stochastic subsampling helps ensure the imputed values cover more of the range of missing values. Unlike the linear boosters, since tree boosters are nonparametric, we do not add an additional random error term to each imputed value. It is important to note that similar scenarios exist for linear and tree boosters for imputing missing data in the case when the endogenous variable is binary.

With regards to the mediation-type models, we demonstrate how our algorithm can be implemented. First, consider a simple mediation model given by

the three equations

1. $X = \alpha_X + \zeta_X$
2. $M = \alpha_M + \beta_{M \cdot X} X + \zeta_M$
3. $Y = \alpha_Y + \beta_{Y \cdot X} X + \beta_{Y \cdot M} M + \zeta_Y$.

If missing data are present on X, M and Y, we propose that linear boosters should be used

to impute missing data, with other variables specified as features and appropriate loss functions selected based on variable types (i.e., we use the squared-error loss for continuous, normally-distributed variables and the negative log-likelihood binomial loss for binary variables). A more complicated case exists when the substantive model contains interactions, such as a moderated mediation model given by the six equations,

1. $X = \alpha_X + \zeta_X$
2. $W = \alpha_W + \zeta_W$
3. $XW = \alpha_{XW} + \zeta_{XW}$
4. $MW = \alpha_{MW} + \zeta_{MW}$
5. $M = \alpha_M + \beta_{M \cdot X} X + \beta_{M \cdot W} W + \beta_{M \cdot XW} XW + \zeta_M$
6. $Y = \alpha_Y + \beta_{Y \cdot X} X + \beta_{Y \cdot W} W + \beta_{Y \cdot XW} XW + \beta_{Y \cdot MW} MW + \beta_{Y \cdot M} M + \zeta_Y$.

Currently, there is no gold standard for imputing missing data on interaction terms in SEMs. At best, the JAV approach appears to be the most applicable for moderated mediation models with continuous response variables and questionable with non-continuous response variables (e.g., binary; Enders et al., 2014; Seaman et al., 2012; von Hippel, 2009). It is import to mention that the JAV approach using FCS MI can be limited with sparse data. For instance, consider the interaction XW and a simple, extreme case with $n = 5$, where X is missing values on the first two observations and $W$is missing values on the last three observations (see Figure 3.10). Using a procedure similar to MICE (van Buuren & Groothuis-Oudshoorn, 2011), we propose that missing values should be replaced by random draws of observed data for each variable. Next, iterative regression models are trained in which the original indices of observed data are used to subset the data as the training set, and the original indices of missing data are used to subset the data as the

id	X	W	XW
1	?	w_1	?
2	?	w_2	?
3	x_3	?	?
4	x_4	?	?
5	x_5	?	?

Figure 3.10. Simple scenario in which the JAV imputation method always fails because of sparse data on the interaction term XW.

testing set, where imputed values are estimated. As applied to the scenario in Figure (3.10), the random replacement of missing values and selection of observed and missing indices are summarized below:

1. x_1 and x_2 are randomly replaced by draws from $\{x_3, x_4, x_5\}$. The observed indices are $X_{i_{\text{obs}}} = \{3, 4, 5\}$ and missing indices are $X_{i_{\text{mis}}} = \{1, 2\}$.
2. w_2, w_3, and w_4 are randomly replaced by draws from $\{w_1, w_2\}$. The observed indices are $W_{i_{\text{obs}}} = \{1, 2\}$ and missing indices are $W_{i_{\text{mis}}} = \{3, 4, 5\}$.
3. XW is created as the product of the randomly replaced values in X and W. The observed indices are based on the union of missing indices in both X and W as $XW_{i_{\text{mis}}} = X_{i_{\text{mis}}} \cup W_{i_{\text{mis}}} = \{1, 2, 3, 4, 5\}$, whereas, the observed indices are based on the intersection, $XW_{i_{\text{obs}}} = X_{i_{\text{obs}}} \cap W_{i_{\text{obs}}} = \emptyset$.

We can see that in Step (3), an imputation cannot be trained for XW because the observed indices are the empty set, that is, there is no 'true' interaction based on the original X and W. Although an extreme case such as this one is unlikely to occur in practice, if X has p_X percent missing data and W has p_W percent missing data, then the interaction variable XW will have a missingness rate p_{XW} bounded by

$$\min(p_X, p_W) \leq p_{XW} \leq p_X + p_W. \quad (3.29)$$

Applied to Figure (3.10), we see that $p_X = .4$, $p_W = .6$, and $p_{XW} = 1.0$.

Continuing with the moderated mediation model, if we have missing data on all variables X, W, M and Y (which implies missingness on the interaction variables), we propose that two imputation methods, imputing only main effects (passive imputation [PI]) and the JAV method. For PI imputation, we recommend that linear boosters are used to impute missing data on X and W, whereas tree boosters are used to impute missing data on M and Y. In both cases, all other variables should be used as features and appropriate loss functions selected based on the variable type. For the JAV method, linear boosters could be used to impute missing data on X, W, XW, and MW with all other variables as features and appropriate loss functions based on the variable types. Alternatively, tree boosters could be used to impute the interaction variables, however, the performance of tree boosters for imputing interaction variables is an open research question. Note, unless both variables in an interaction are binary, we propose using the squared-error loss function to impute missing values. On the contrary, since the substantive models for M and Y contain interactive effects, we propose that tree boosters should be used to impute missing data on these variables, with all other variables as features and appropriate loss functions.

Although the use of the BB with MI has been limited in research, the use of the nonparametric bootstrap and MI has been investigated in other studies with promising results. Recently, Schomaker and Heumann (2016) examined the performance of four different bootstrap (with percentile-based CIs) and MI methods to impute missing data in linear models: (1) MI then bootstrap based on pooling, (2) MI then bootstrap without

pooling, (3) bootstrap then MI with pooling, and (4) bootstrap then MI without pooling. Results of the study demonstrated that among the four methods, only the bootstrap then MI with pooled estimates provided efficient and randomization valid confidence intervals. Randomization valid here indicates that the actual confidence interval coverage equaled the normal confidence interval coverage.

In a more relevant study, Wang and Wang (2014) applied a similar approach, but with bias-corrected confidence intervals for estimating unconditional indirect effects in mediation and found the method performed well under MCAR and MAR missingness mechanisms. Wu and Jia (2013) examined the performance of performing MI first and then using a nonparametric bootstrap sample within each imputed data set. Results demonstrated that their algorithm performed comparably to bootstrapping with FIML. Although the MI then bootstrap method is computationally faster than the bootstrap then MI approach, the underlying method is theoretically flawed. As described in Schomaker and Heumann (2016) and Enders et al. (2013), inferences based on the MI then bootstrap method are inappropriate because the empirical sampling distributions reflect the variation of the complete-data indirect effect estimates (i.e., the bootstrap sampling distributions are narrower than their missing-data counterparts). Among all these studies, the JM MI framework was used.

CHAPTER 4

MONTE CARLO SIMULATIONS STUDIES

4.1. Study 1

4.1.1. Method. Despite its similarity to the nonparametric bootstrap, little research has examined the performance of the Bayesian bootstrap (BB) for indirect effects analysis. The first purpose of the current study is fill this gap, that is, to empirically examine the performance of the BB for estimating and testing unconditional and conditional indirect effects. A Monte Carlo (MC) simulation study is conducted to examine: (1) biases, (2) mean squared errors (MSEs), (3) confidence interval coverage probabilities, (4) length of confidence intervals, and (5) Type I error/power rates of the BB for estimating and testing indirect effects under different conditions. For comparison, the relative performance of the BB is compared to that of some conventional methods, including the nonparametric bootstrap with bias-corrected confidence intervals (current best practice), first-order standard errors, and second-order standard errors. For both Bayesian (both stage one and stage two sampling) and nonparametric bootstrapping, 1,000 bootstrap samples are used. All parameters are estimated with ML and α is set to .05 to generate 95% confidence intervals or credible intervals for BB. Note, we refer to both frequentist confidence intervals and Bayesian credible intervals as confidence intervals (CIs) to keep language similar in text, tables, and figures. However, with BB, the CIs reflect credible intervals. For BB, the mean and median indirect effect estimates are computed for comparisons (Wang & Preacher, 2015). Three sample sizes of 100, 500, and 1,000 are chosen to represent small,

medium, and large sample sizes for mediation and moderated mediation models, respectively.

For simulating data, two models are considered. First, a simple mediation model (see Figure [1.2]),

$$\begin{bmatrix} X \\ M \\ Y \end{bmatrix} = \begin{bmatrix} \alpha_X \\ \alpha_M \\ \alpha_Y \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ \beta_{M\cdot X} & 0 & 0 \\ \beta_{Y\cdot X} & \beta_{Y\cdot M} & 0 \end{bmatrix} \begin{bmatrix} X \\ M \\ Y \end{bmatrix} + \begin{bmatrix} \zeta_X \\ \zeta_M \\ \zeta_Y \end{bmatrix},$$

and second, a moderated mediation model given in Figure (1.4) Model 5,

$$\begin{bmatrix} X \\ W \\ XW \\ MW \\ M \\ Y \end{bmatrix} = \begin{bmatrix} \alpha_X \\ \alpha_W \\ \alpha_{XW} \\ \alpha_{MW} \\ \alpha_M \\ \alpha_Y \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ \beta_{M\cdot X} & \beta_{M\cdot W} & \beta_{M\cdot XW} & 0 & 0 & 0 \\ \beta_{Y\cdot X} & \beta_{Y\cdot W} & \beta_{Y\cdot XW} & \beta_{Y\cdot MW} & \beta_{Y\cdot M} & 0 \end{bmatrix} \begin{bmatrix} X \\ W \\ XW \\ MW \\ M \\ Y \end{bmatrix} + \begin{bmatrix} \zeta_X \\ \zeta_W \\ \zeta_{XW} \\ \zeta_{MW} \\ \zeta_M \\ \zeta_Y \end{bmatrix}.$$

For all models, four combination of variables types are used for the mediator variable M and endogenous variable Y: (1) continuous M, continuous Y, (2), continuous M, binary Y, (3) binary M, continuous Y, and (4) binary M, binary Y. Using the generalized linear model framework, a linear regression is used to model continuous variables and a logistic regression is used to model binary variables.

All regression parameters are simulated with four different combinations of effect sizes (null = .00, small = .14, medium = .39, large = .59; Cohen, 1988) and only true indirect effects are simulated (Preacher et al., 2007). Exogenous variables are simulated to have normal distributions with zero mean and unit variance. Mediator and endogenous variables are simulated based on linear combinations described by the four different combinations of effect sizes. For continuous mediator and/or endogenous variables, the errors are simulated to have normal distributions with zero mean and unit variance. For

binary mediator and/or endogenous variables, the linear combination $\boldsymbol{\eta}$ that are used to simulate the variable will be converted into a probability using the logit transformation as

$$P(M = 1|\boldsymbol{\eta}_M, \boldsymbol{\theta}_M) = \frac{1}{1 + \exp(-\boldsymbol{\eta}_M)} \quad (4.1)$$

and

$$P(Y = 1|\boldsymbol{\eta}_Y, \boldsymbol{\theta}_Y) = \frac{1}{1 + \exp(-\boldsymbol{\eta}_Y)} \quad (4.2)$$

for the mediator and endogenous variable models, respectively. Then, the probabilities in (4.1) and (4.2) are used as the success probability for which random draws from a Binomial distribution are sampled using one trial (i.e., to simulate a 0/1 response). Similar to other studies (Preacher et al., 2007; Preacher & Wang, 2015), the conditional indirect effects from the moderated mediation model is tested at a value of +1 (i.e., approximately 1 standard deviation above the mean) on the moderator variable. The simulation consists of 2 (types of models) x 4 (mediator and endogenous variable type combinations) x 3 (sample sizes) x 4 (regression coefficient effect sizes) = 96 unique conditions. To help combat the effects of MC error, 10,000 replications are simulated per condition (Koehler, Brown, & Haneuse, 2009).

It is important to note that in this simulation study (and the next simulation study), parameters are fixed and only the data are resampled or generated. As such, the simulations are conducted from a frequentist perspective, which parallels previous research in this area (e.g., Enders et al., 2014; Yuan & MacKinnon, 2009; Wang & Preacher, 2015). As Yuan and MacKinnon (2009, p. 10) note, it is of interest to evaluate the performance of any Bayesian analysis from the frequentist point of view because provided

that the Bayesian model is correctly specified, Bayesian estimates are consistent and credible intervals have exact nominal coverage rates regardless of sample size.

Let γ denote the true indirect effect and $\hat{\gamma}_i$ denote the estimated indirect effect for the ith replicate, then to evaluate the performance of the studied methods, the empirical bias (EB) is calculated as

$$\text{EB} = \begin{cases} 100\left[\frac{\sum_{i=1}^{T}\hat{\gamma}_i}{T\gamma} - 1\right], & \gamma \neq 0 \\ \frac{\sum_{i=1}^{T}\hat{\gamma}_i}{T} - \gamma, & \gamma = 0 \end{cases}$$

where T is the total number of replications per condition. Unbiased estimates should have EB estimates around zero. MSE is calculated as

$$\text{MSE} = \frac{1}{T}\sum_{i=1}^{T}(\hat{\gamma}_i - \gamma)^2.$$

For MSE estimates, the lower the estimate the better the performance. Let $\widehat{LL}_i$ and $\widehat{UL}_i$ denote the estimated lower and upper limits of a 95% CI of γ in the ith replicate, then the coverage probability (CP) is calculated as

$$\text{CP} = \frac{1}{T}\sum_{i=1}^{T} I\left(\widehat{LL}_i < \gamma < \widehat{UL}_i\right),$$

where $I\left(\widehat{LL}_i < \gamma < \widehat{UL}_i\right)$ is the indicator function defined as

$$I\left(\widehat{LL}_i < \gamma < \widehat{UL}_i\right) = \begin{cases} 1, & \gamma \in \left(\widehat{LL}_i, \widehat{UL}_i\right) \\ 0, & \gamma \notin \left(\widehat{LL}_i, \widehat{UL}_i\right) \end{cases}.$$

At $\alpha = .05$, CP should be approximately .95. Confidence interval length (CIL) is calculated as

$$\text{CIL} = \frac{1}{T}\sum_{i=1}^{T}\left|\widehat{UL}_i - \widehat{LL}_i\right|.$$

For CIL estimates, assuming other metrics are constant, the narrower the interval the better. Lastly, the rejection rate (RR) is calculated as

$$\text{RR} = \frac{1}{T}\sum_{i=1}^{T} I\left(\left(\widehat{LL}_i > 0\right) \vee I\left(\widehat{UL}_i < 0\right)\right),$$

where $I\left(\left(\widehat{LL}_i > 0\right) \vee I\left(\widehat{UL}_i < 0\right)\right)$ is the indicator function defined as

$$I\left(\left(\widehat{LL}_i > 0\right) \vee I\left(\widehat{UL}_i < 0\right)\right) = \begin{cases} 1, & 0 \notin \left(\widehat{LL}_i, \widehat{UL}_i\right) \\ 0, & 0 \in \left(\widehat{LL}_i, \widehat{UL}_i\right) \end{cases}. \quad (4.3)$$

Here, when $\gamma = 0$, (4.3) is the Type I Error rate, and when $\gamma \neq 0$, (4.3) is the power. When $\gamma = 0$, RR should be .05 or less (for conservative estimates), whereas, when $\gamma \neq 0$, higher RR indicate higher empirical power.

Given that the point estimate of the BC bootstrap is based on the sample data, as opposed to the bootstrapped distribution, empirical biases and MSEs are equivalent for the BC bootstrap and delta methods. Despite this fact, however, we report empirical biases and MSEs results for each method to keep reporting consistent with Study 2.

4.1.2. Results: Mediation models. Results of the MC simulation for a mediation model with continuous mediator and endogenous variables are displayed in Table 4.1. As can be seen in the table, all methods are relatively unbiased. The BB with median estimator, however, tends to underestimate the true indirect effect with small effect sizes (i.e., effect = .14). For confidence interval lengths, all methods have similar lengths with null and small (i.e., effect = .14) effect sizes. The BB methods have slightly wider intervals than other methods at medium (i.e., effect = .39) and large (i.e., effect = .59) effect sizes in both medium ($n = 500$) and large ($n = 1000$) sample sizes. Across all conditions, the BB methods have approximately nominal coverage rates (i.e., 95%) or higher. For small effect

Table 4.1
Mediation Model Metrics – Mediator Continuous, Endogenous Continuous

		Empirical Bias			*Confidence Interval Length*			*Coverage Probability*			*Rejection Rate*			*Mean Squared Error*		
Method	*Effect*	*100*	*500*	*1000*	*100*	*500*	*1000*	*100*	*500*	*1000*	*100*	*500*	*1000*	*100*	*500*	*1000*
Delta-1	.00	0.000	0.000	0.000	0.050	0.001	0.005	1.000	1.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000
Delta-1	.14	0.027	0.000	0.003	0.091	0.036	0.025	0.912	0.925	0.936	0.019	0.580	0.968	0.001	0.000	0.000
Delta-1	.39	0.007	0.002	0.001	0.224	0.097	0.069	0.934	0.950	0.949	0.866	1.000	1.000	0.003	0.001	0.000
Delta-1	.59	0.001	-0.001	0.001	0.335	0.147	0.104	0.944	0.945	0.953	1.000	1.000	1.000	0.007	0.001	0.001
Delta-2	.00	0.000	0.000	0.000	0.066	0.013	0.007	1.000	1.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000
Delta-2	.14	0.027	0.000	0.003	0.101	0.037	0.025	0.992	0.935	0.940	0.015	0.542	0.963	0.001	0.000	0.000
Delta-2	.39	0.007	0.002	0.001	0.228	0.098	0.069	0.940	0.951	0.949	0.850	1.000	1.000	0.003	0.001	0.000
Delta-2	.59	0.001	-0.001	0.001	0.338	0.147	0.104	0.946	0.945	0.953	1.000	1.000	1.000	0.007	0.001	0.001
BC-Boot	.00	0.000	0.000	0.000	0.075	0.015	0.007	0.992	0.995	0.994	0.008	0.005	0.006	0.000	0.000	0.000
BC-Boot	.14	0.027	0.000	0.003	0.108	0.037	0.026	0.913	0.955	0.952	0.134	0.824	0.988	0.001	0.000	0.000
BC-Boot	.39	0.007	0.002	0.001	0.231	0.098	0.069	0.949	0.951	0.951	0.950	1.000	1.000	0.003	0.001	0.000
BC-Boot	.59	0.001	-0.001	0.001	0.338	0.146	0.103	0.947	0.943	0.948	1.000	1.000	1.000	0.007	0.001	0.001
Bayes-1	.00	0.000	0.000	0.000	0.071	0.019	0.012	0.998	1.000	1.000	0.002	0.000	0.000	0.000	0.000	0.000
Bayes-1	.14	0.025	0.000	0.003	0.104	0.045	0.036	0.966	0.979	0.993	0.067	0.552	0.885	0.001	0.000	0.000
Bayes-1	.39	0.007	0.002	0.001	0.228	0.118	0.097	0.944	0.982	0.994	0.922	1.000	1.000	0.003	0.001	0.000
Bayes-1	.59	0.001	-0.001	0.000	0.338	0.178	0.146	0.946	0.981	0.993	1.000	1.000	1.000	0.007	0.001	0.001
Bayes-2	.00	0.000	0.000	0.000	0.071	0.019	0.012	0.998	1.000	1.000	0.002	0.000	0.000	0.000	0.000	0.000
Bayes-2	.14	-0.106	-0.065	-0.044	0.104	0.044	0.036	0.965	0.979	0.992	0.069	0.557	0.885	0.000	0.000	0.000
Bayes-2	.39	-0.024	-0.008	-0.005	0.228	0.118	0.097	0.945	0.982	0.994	0.921	1.000	1.000	0.003	0.001	0.000
Bayes-2	.59	-0.012	-0.005	-0.002	0.338	0.178	0.146	0.949	0.979	0.994	1.000	1.000	1.000	0.007	0.001	0.001

Note. Delta-1 = first-order delta method; Delta-2 = second-order delta method; BC-Boot = bias-corrected nonparametric bootstrap; Bayes-1 = Bayesian bootstrap with mean estimator; Bayes-2 = Bayesian bootstrap with median estimator.

sizes in sample sizes, however, both the first-order delta method and BC bootstrap have lower than nominal coverage rates (i.e., 95%). Similarly, for small effects in medium sample sizes, both delta methods have slightly less than nominal coverage rates; other conditions achieve approximately nominal coverage rates for each method. For each method, Type I Error rates are well below nominal rates (i.e., 5%). In regards to power, with larger effect sizes in larger samples, all methods generally have similar performance. For small effect sizes, however, the BC bootstrap has highest power across all sample sizes, whereas, the BB methods have slightly higher power than delta methods in small samples; in large samples, the delta methods have higher power than the BB methods. The mean squared errors (MSEs) are approximately similar across all methods and conditions. Supplemental figures for the results are presented in Appendix C.

Results of the MC simulation for a mediation model with continuous mediator variable and categorical endogenous variable are displayed in Table 4.2. With regards to empirical bias, all methods are biased in small sample sizes with non-null effect sizes. For other conditions examined, however, all methods aside from the BB with median estimator are relatively unbiased. Similar to the previous simulation results, the BB with median estimator tends to underestimate the true indirect effect with small effect sizes in medium and large sample sizes. The confidence interval lengths are slightly wider for the BB methods with medium and large effect sizes at larger sample sizes, whereas narrower than the BC bootstrap in small sample sizes.

With regards to coverage probabilities, the second-order delta method and the BB methods have approximately nominal or higher coverage levels, with the BB methods having slightly higher coverage levels. For small samples, the BC bootstrap has slightly

Table 4.2
Mediation Model Metrics – Mediator Continuous, Endogenous Categorical

		Empirical Bias			*Confidence Interval Length*			*Coverage Probability*			*Rejection Rate*			*Mean Squared Error*		
Method	*Effect*	*100*	*500*	*1000*	*100*	*500*	*1000*	*100*	*500*	*1000*	*100*	*500*	*1000*	*100*	*500*	*1000*
Delta-1	.00	0.000	0.000	0.000	0.105	0.020	0.010	1.000	1.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000
Delta-1	.14	0.063	0.005	0.000	0.160	0.058	0.040	0.966	0.934	0.937	0.003	0.140	0.472	0.002	0.000	0.000
Delta-1	.39	0.058	0.010	0.005	0.385	0.162	0.114	0.943	0.951	0.948	0.282	0.989	1.000	0.010	0.002	0.001
Delta-1	.59	0.049	0.010	0.006	0.607	0.259	0.182	0.947	0.947	0.949	0.720	1.000	1.000	0.027	0.004	0.002
Delta-2	.00	0.000	0.000	0.000	0.137	0.026	0.013	1.000	1.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000
Delta-2	.14	0.063	0.005	0.000	0.184	0.060	0.041	0.999	0.950	0.944	0.001	0.120	0.446	0.002	0.000	0.000
Delta-2	.39	0.058	0.010	0.005	0.395	0.163	0.114	0.952	0.953	0.949	0.256	0.988	1.000	0.010	0.002	0.001
Delta-2	.59	0.049	0.010	0.006	0.614	0.260	0.182	0.950	0.947	0.949	0.707	1.000	1.000	0.027	0.004	0.002
BC-Boot	.00	0.000	0.000	0.000	0.170	0.030	0.015	0.994	0.994	0.995	0.007	0.007	0.005	0.000	0.000	0.000
BC-Boot	.14	0.063	0.005	0.000	0.219	0.063	0.042	0.933	0.942	0.950	0.059	0.384	0.642	0.002	0.000	0.000
BC-Boot	.39	0.058	0.010	0.005	0.445	0.166	0.115	0.940	0.947	0.948	0.560	0.993	1.000	0.010	0.002	0.001
BC-Boot	.59	0.049	0.010	0.006	0.688	0.264	0.183	0.927	0.941	0.947	0.835	1.000	1.000	0.027	0.004	0.002
Bayes-1	.00	0.000	0.000	0.000	0.159	0.038	0.024	0.998	1.000	1.000	0.002	0.000	0.000	0.001	0.000	0.000
Bayes-1	.14	0.119	0.021	0.010	0.206	0.077	0.059	0.989	0.980	0.993	0.022	0.130	0.255	0.002	0.000	0.000
Bayes-1	.39	0.115	0.024	0.015	0.427	0.201	0.163	0.954	0.984	0.994	0.439	0.970	0.999	0.011	0.002	0.001
Bayes-1	.59	0.107	0.025	0.016	0.662	0.320	0.259	0.948	0.980	0.994	0.776	1.000	1.000	0.031	0.005	0.002
Bayes-2	.00	0.000	0.000	0.000	0.159	0.038	0.024	0.998	1.000	1.000	0.002	0.000	0.000	0.000	0.000	0.000
Bayes-2	.14	-0.069	-0.084	-0.068	0.206	0.076	0.059	0.988	0.980	0.994	0.022	0.132	0.252	0.001	0.000	0.000
Bayes-2	.39	0.049	0.005	0.002	0.427	0.201	0.163	0.956	0.984	0.994	0.440	0.969	0.999	0.010	0.002	0.001
Bayes-2	.59	0.068	0.015	0.009	0.662	0.321	0.259	0.948	0.980	0.994	0.777	1.000	1.000	0.029	0.005	0.002

Note. Delta-1 = first-order delta method; Delta-2 = second-order delta method; BC-Boot = bias-corrected nonparametric bootstrap; Bayes-1 = Bayesian bootstrap with mean estimator; Bayes-2 = Bayesian bootstrap with median estimator.

lower than nominal coverage rates for non-null effects; other conditions the method achieves approximately nominal coverage levels. Similarly, in small samples, the first-order delta method has lower than nominal coverage levels for small effects; this method achieves nominal coverage in other conditions. In terms of Type I error rates, all methods have lower than nominal coverage. For empirical power, the bootstrap methods outperform the delta methods in small sample sizes, with the BC bootstrap having the highest power. In larger samples with small effect sizes, the BC bootstrap has the highest power; in large samples the delta methods have higher power than the BB methods. Power in other conditions are similar across methods. All methods have comparable MSEs. Supplemental figures for the results are presented in Appendix C.

Results of the MC simulation for a mediation model with categorical mediator variable and continuous endogenous variable are displayed in Table 4.3. In small sample sizes with non-null effects, all methods are biased except for the BB with median estimator at large effects. On the contrary, for medium and large sample sizes, all methods are relatively unbiased except for the BB with median estimator at small effect sizes. In small samples, the delta methods have the narrowest confidence interval lengths, whereas, the BC bootstrap has the widest. For medium to large sample sizes, the BB methods have slightly wider confidence intervals than the BC bootstrap and delta methods, which are approximately similar. In terms of coverage probabilities, the BB methods have the highest coverage rates across conditions, whereas the first-order delta method has the lowest coverage rate across conditions. The second-order delta method has higher power than the BC bootstrap and first-order delta method in for smaller effect sizes and approximately similar power at medium and large effect sizes.

Table 4.3
Mediation Model Metrics – Mediator Categorical, Endogenous Continuous

		Empirical Bias			*Confidence Interval Length*			*Coverage Probability*			*Rejection Rate*			*Mean Squared Error*		
Method	*Effect*	*100*	*500*	*1000*	*100*	*500*	*1000*	*100*	*500*	*1000*	*100*	*500*	*1000*	*100*	*500*	*1000*
Delta-1	.00	0.000	0.000	0.000	0.206	0.040	0.020	1.000	1.000	1.000	0.000	0.000	0.000	0.002	0.000	0.000
Delta-1	.14	0.053	0.018	-0.001	0.255	0.078	0.052	0.983	0.906	0.908	0.001	0.026	0.141	0.004	0.000	0.000
Delta-1	.39	0.037	0.010	0.008	0.503	0.205	0.144	0.911	0.937	0.950	0.055	0.952	1.000	0.017	0.003	0.001
Delta-1	.59	0.032	0.006	0.004	0.758	0.321	0.225	0.923	0.947	0.945	0.355	1.000	1.000	0.038	0.007	0.003
Delta-2	.00	0.000	0.000	0.000	0.271	0.052	0.026	1.000	1.000	1.000	0.000	0.000	0.000	0.002	0.000	0.000
Delta-2	.14	0.053	0.018	-0.001	0.312	0.085	0.055	1.000	0.965	0.929	0.000	0.021	0.119	0.004	0.000	0.000
Delta-2	.39	0.037	0.010	0.008	0.536	0.208	0.145	0.943	0.942	0.952	0.041	0.944	1.000	0.017	0.003	0.001
Delta-2	.59	0.032	0.006	0.004	0.783	0.323	0.226	0.934	0.949	0.946	0.317	1.000	1.000	0.038	0.007	0.003
BC-Boot	.00	0.000	0.000	0.000	0.330	0.060	0.030	0.993	0.994	0.995	0.007	0.006	0.006	0.002	0.000	0.000
BC-Boot	.14	0.053	0.018	-0.001	0.375	0.093	0.057	0.980	0.927	0.954	0.026	0.175	0.452	0.004	0.000	0.000
BC-Boot	.39	0.037	0.010	0.008	0.616	0.215	0.147	0.939	0.952	0.953	0.324	0.987	1.000	0.017	0.003	0.001
BC-Boot	.59	0.032	0.006	0.004	0.887	0.330	0.228	0.950	0.950	0.947	0.744	1.000	1.000	0.038	0.007	0.003
Bayes-1	.00	0.000	0.000	0.000	0.313	0.076	0.047	0.999	1.000	1.000	0.001	0.000	0.000	0.002	0.000	0.000
Bayes-1	.14	0.099	0.031	0.007	0.354	0.112	0.082	0.995	0.995	0.993	0.007	0.029	0.055	0.004	0.000	0.000
Bayes-1	.39	0.081	0.022	0.016	0.578	0.256	0.205	0.955	0.980	0.995	0.173	0.935	0.999	0.019	0.003	0.001
Bayes-1	.59	0.076	0.018	0.012	0.833	0.396	0.321	0.954	0.982	0.994	0.594	1.000	1.000	0.042	0.007	0.003
Bayes-2	.00	0.000	0.000	0.000	0.313	0.076	0.047	0.999	1.000	1.000	0.001	0.000	0.000	0.001	0.000	0.000
Bayes-2	.14	-0.124	-0.150	-0.148	0.354	0.112	0.082	0.996	0.996	0.993	0.008	0.027	0.054	0.003	0.000	0.000
Bayes-2	.39	-0.039	-0.020	-0.013	0.578	0.256	0.205	0.956	0.982	0.995	0.178	0.935	0.999	0.017	0.003	0.001
Bayes-2	.59	0.003	-0.004	-0.003	0.833	0.396	0.321	0.954	0.982	0.994	0.591	1.000	1.000	0.040	0.007	0.003

Note. Delta-1 = first-order delta method; Delta-2 = second-order delta method; BC-Boot = bias-corrected nonparametric bootstrap; Bayes-1 = Bayesian bootstrap with mean estimator; Bayes-2 = Bayesian bootstrap with median estimator.

Similar to previous results, all methods have lower than nominal Type I Error rates. The BC bootstrap has higher power relative to other methods at larger effects in small samples and at small effects in larger samples. The BB methods have higher power than the delta methods in small samples with larger effect sizes; the delta methods have higher power than the BB methods in large samples at small effect sizes. Power across other conditions and methods is comparable. All methods have comparable MSEs in the conditions examined. Supplemental figures for the results are presented in Appendix C.

Results of the MC simulation for a mediation model with categorical mediator variable and categorical endogenous variable are displayed in Table 4.4. In small sample sizes, the BB with median estimator is the least biased method for non-null effects; the median estimator is slightly biased for small and large effect sizes. In larger samples, however, the BB with median estimator is biased for small and medium effects. Other methods are relatively unbiased for effects in medium and large samples. In terms of confidence interval lengths, results echo previous findings such that in small samples, the BC bootstrap has the widest intervals and the delta methods have the narrowest intervals. Moreover, in medium and large samples, the BB methods have slightly wider intervals than other methods.

The BB methods have the highest coverage probabilities across all non-null effect sizes examined. The BC bootstrap has lower than nominal coverage probabilities in small samples with medium and large effect sizes, whereas, in medium and large sample sizes, coverage rates are below nominal levels for small effect sizes. The second-order delta method generally has higher coverage rates than the first-order delta method. All methods have lower than nominal Type I Error rates. With regards to power, in small samples the

Table 4.4
Mediation Model Metrics – Mediator Categorical, Endogenous Categorical

		Empirical Bias			*Confidence Interval Length*			*Coverage Probability*			*Rejection Rate*			*Mean Squared Error*		
Method	*Effect*	*100*	*500*	*1000*	*100*	*500*	*1000*	*100*	*500*	*1000*	*100*	*500*	*1000*	*100*	*500*	*1000*
Delta-1	.00	0.001	0.000	0.000	0.420	0.081	0.040	1.000	1.000	1.000	0.000	0.000	0.000	0.008	0.000	0.000
Delta-1	.14	0.153	0.010	0.010	0.490	0.132	0.085	0.996	0.953	0.922	0.000	0.005	0.030	0.013	0.001	0.000
Delta-1	.39	0.082	0.016	-0.003	0.859	0.328	0.227	0.944	0.943	0.942	0.011	0.423	0.818	0.051	0.007	0.003
Delta-1	.59	0.075	0.002	0.006	1.265	0.508	0.356	0.930	0.945	0.946	0.064	0.841	0.992	0.113	0.017	0.008
Delta-2	.00	0.001	0.000	0.000	0.553	0.106	0.052	1.000	1.000	1.000	0.000	0.000	0.000	0.008	0.000	0.000
Delta-2	.14	0.153	0.010	0.010	0.612	0.149	0.092	1.000	0.999	0.971	0.000	0.004	0.023	0.013	0.001	0.000
Delta-2	.39	0.082	0.016	-0.003	0.942	0.335	0.229	0.998	0.950	0.944	0.008	0.397	0.808	0.051	0.007	0.003
Delta-2	.59	0.075	0.002	0.006	1.331	0.513	0.358	0.950	0.947	0.947	0.051	0.834	0.991	0.113	0.017	0.008
BC-Boot	.00	0.001	0.000	0.000	0.711	0.122	0.060	0.992	0.994	0.995	0.008	0.006	0.005	0.008	0.000	0.000
BC-Boot	.14	0.153	0.010	0.010	0.778	0.165	0.098	0.984	0.931	0.922	0.018	0.069	0.173	0.013	0.001	0.000
BC-Boot	.39	0.082	0.016	-0.003	1.160	0.351	0.234	0.913	0.948	0.946	0.136	0.634	0.874	0.051	0.007	0.003
BC-Boot	.59	0.075	0.002	0.006	1.615	0.532	0.363	0.922	0.948	0.941	0.334	0.894	0.994	0.113	0.017	0.008
Bayes-1	.00	0.001	0.000	0.000	0.667	0.155	0.095	0.999	1.000	1.000	0.001	0.000	0.000	0.009	0.000	0.000
Bayes-1	.14	0.240	0.032	0.024	0.730	0.204	0.144	0.998	0.998	0.999	0.003	0.006	0.007	0.015	0.001	0.000
Bayes-1	.39	0.167	0.037	0.011	1.085	0.423	0.331	0.979	0.983	0.993	0.054	0.373	0.584	0.060	0.007	0.004
Bayes-1	.59	0.162	0.024	0.020	1.508	0.642	0.514	0.954	0.983	0.994	0.187	0.763	0.948	0.135	0.018	0.009
Bayes-2	.00	0.001	0.000	0.000	0.667	0.155	0.096	0.999	1.000	1.000	0.001	0.000	0.000	0.006	0.000	0.000
Bayes-2	.14	-0.022	-0.188	-0.186	0.730	0.204	0.144	0.997	0.998	0.999	0.003	0.007	0.008	0.010	0.001	0.000
Bayes-2	.39	0.000	-0.030	-0.036	1.085	0.423	0.331	0.979	0.984	0.993	0.053	0.372	0.585	0.049	0.007	0.003
Bayes-2	.59	0.047	-0.013	-0.005	1.509	0.643	0.514	0.952	0.984	0.995	0.191	0.762	0.947	0.117	0.017	0.008

Note. Delta-1 = first-order delta method; Delta-2 = second-order delta method; BC-Boot = bias-corrected nonparametric bootstrap; Bayes-1 = Bayesian bootstrap with mean estimator; Bayes-2 = Bayesian bootstrap with median estimator.

BC bootstrap has the highest power for non-null effects, whereas, the delta methods have the lowest power for non-null effects. In medium sample sizes, the BC bootstrap has higher power than the BB methods and delta methods; the delta methods have slightly higher power than the BB methods. In large samples with small effect sizes, the BC bootstrap has higher power than other methods. Interestingly, in large samples the BB methods have lower power than other methods with medium effect sizes. All methods have approximately similar MSEs. Supplemental figures for the results are presented in Appendix C.

4.1.3. Results: Moderated mediation models. Results of the MC simulation for a moderated mediation model with continuous mediator and continuous endogenous variable are displayed in Table 4.5. As can be seen in the table, except for the BB with median estimator with smaller effect sizes in smaller samples, all methods are relatively unbiased. In terms of confidence interval lengths, all methods have approximately similar lengths in small samples, but the BB methods have slightly wider intervals in medium and large sample sizes. In small samples, the first-order delta method has considerably lower coverage than the other methods for small effect sizes, all of which have below nominal coverage levels; at medium and large effects, all methods have approximately nominal coverage rates, with slightly lower rates for BB methods. In medium and large sample sizes, all methods have approximately nominal or higher coverage rates, with BB methods having the highest coverage. All methods have lower than nominal Type I Error rates. In terms of power, the bootstrap methods outperform the delta methods in small samples with small effect sizes; in all other conditions the methods have approximately similar power.

Table 4.5
Moderated Mediation Model Metrics – Mediator Continuous, Endogenous Continuous

		Empirical Bias			*Confidence Interval Length*			*Coverage Probability*			*Rejection Rate*			*Mean Squared Error*		
Method	*Effect*	*100*	*500*	*1000*	*100*	*500*	*1000*	*100*	*500*	*1000*	*100*	*500*	*1000*	*100*	*500*	*1000*
Delta-1	.00	0.000	0.000	0.000	0.110	0.020	0.010	1.000	1.000	1.000	0.000	0.000	0.000	0.001	0.000	0.000
Delta-1	.14	0.000	0.001	-0.001	0.246	0.100	0.070	0.901	0.937	0.941	0.087	0.965	1.000	0.004	0.001	0.000
Delta-1	.39	0.002	0.002	0.000	0.623	0.264	0.185	0.944	0.949	0.950	0.996	1.000	1.000	0.025	0.005	0.002
Delta-1	.59	-0.001	0.001	0.000	0.907	0.388	0.272	0.940	0.949	0.952	1.000	1.000	1.000	0.057	0.010	0.005
Delta-2	.00	0.000	0.000	0.000	0.145	0.027	0.013	1.000	1.000	1.000	0.000	0.000	0.000	0.001	0.000	0.000
Delta-2	.14	0.000	0.001	-0.001	0.263	0.101	0.070	0.933	0.941	0.942	0.073	0.959	1.000	0.004	0.001	0.000
Delta-2	.39	0.002	0.002	0.000	0.628	0.264	0.185	0.945	0.949	0.951	0.995	1.000	1.000	0.025	0.005	0.002
Delta-2	.59	-0.001	0.001	0.000	0.910	0.388	0.272	0.940	0.949	0.952	1.000	1.000	1.000	0.057	0.010	0.005
BC-Boot	.00	0.000	0.000	0.000	0.173	0.030	0.015	0.994	0.994	0.995	0.006	0.007	0.005	0.001	0.000	0.000
BC-Boot	.14	0.000	0.001	-0.001	0.284	0.102	0.070	0.932	0.953	0.950	0.290	0.990	1.000	0.004	0.001	0.000
BC-Boot	.39	0.002	0.002	0.000	0.645	0.264	0.185	0.947	0.948	0.949	0.997	1.000	1.000	0.025	0.005	0.002
BC-Boot	.59	-0.001	0.001	0.000	0.932	0.387	0.271	0.939	0.948	0.949	1.000	1.000	1.000	0.057	0.010	0.005
Bayes-1	.00	0.000	0.000	0.000	0.148	0.037	0.023	0.997	1.000	1.000	0.003	0.000	0.000	0.001	0.000	0.000
Bayes-1	.14	-0.001	0.001	-0.001	0.257	0.121	0.098	0.928	0.979	0.993	0.226	0.950	0.999	0.004	0.001	0.000
Bayes-1	.39	0.002	0.002	0.000	0.612	0.318	0.260	0.937	0.981	0.992	0.997	1.000	1.000	0.026	0.005	0.002
Bayes-1	.59	-0.001	0.001	0.000	0.892	0.466	0.382	0.929	0.980	0.993	1.000	1.000	1.000	0.057	0.010	0.005
Bayes-2	.00	0.000	0.000	0.000	0.148	0.037	0.023	0.998	1.000	1.000	0.002	0.000	0.000	0.000	0.000	0.000
Bayes-2	.14	-0.086	-0.033	-0.024	0.257	0.121	0.098	0.929	0.979	0.993	0.225	0.949	0.999	0.004	0.001	0.000
Bayes-2	.39	-0.011	-0.002	-0.003	0.612	0.318	0.260	0.938	0.982	0.994	0.998	1.000	1.000	0.026	0.005	0.002
Bayes-2	.59	-0.007	-0.001	-0.001	0.891	0.466	0.381	0.929	0.981	0.994	1.000	1.000	1.000	0.057	0.010	0.005

Note. Delta-1 = first-order delta method; Delta-2 = second-order delta method; BC-Boot = bias-corrected nonparametric bootstrap; Bayes-1 = Bayesian bootstrap with mean estimator; Bayes-2 = Bayesian bootstrap with median estimator.

Furthermore, MSEs are approximately similar for all methods in each condition. Supplemental figures for the results are presented in Appendix C.

Results of the MC simulation for a moderated mediation model with continuous mediator variable and categorical endogenous variable are displayed in Table 4.6. In small samples, all methods are highly biased for non-null effects. In medium sample sizes, BB methods tend to be slightly more biased than delta methods. For larger samples, however, this difference still exists, but to a smaller extent; the BB with median estimator has lower biases than mean estimator. With regards to confidence interval lengths, in small samples the BC bootstrap has the widest intervals, whereas, the delta methods have the smallest intervals. In medium and large sample sizes, the BB methods have slightly wider intervals than other methods.

In small samples, the BC bootstrap has lower than nominal coverage rates as effect size increases. Surprisingly, the delta methods obtain nominal coverage rates or higher across all conditions, whereas, the BB methods also obtain nominal coverage rates or higher across all conditions except in small samples with medium and large effects. All methods have lower than nominal Type I Error rates. For power in small samples, the bootstrap methods outperform the delta methods, especially at medium effect sizes. In medium sample sizes, the BC bootstrap obtains higher power with small effects; the power for all other effects are similar across methods. Moreover, the power is similar across methods large sample sizes at medium and large effects; at small effect sizes the BB methods have lower power than other methods. With regards to MSEs, in small samples, the MSEs for the bootstrap methods are higher than the delta methods for large effects, with the BC bootstrap having the most variability. All other MSEs are similar across

Table 4.6
Moderated Mediation Model Metrics – Mediator Continuous, Endogenous Categorical

		Empirical Bias			*Confidence Interval Length*			*Coverage Probability*			*Rejection Rate*			*Mean Squared Error*		
Method	*Effect*	*100*	*500*	*1000*	*100*	*500*	*1000*	*100*	*500*	*1000*	*100*	*500*	*1000*	*100*	*500*	*1000*
Delta-1	.00	0.000	0.000	0.000	0.241	0.041	0.020	1.000	1.000	1.000	0.000	0.000	0.000	0.003	0.000	0.000
Delta-1	.14	0.165	0.030	0.013	0.458	0.165	0.114	0.947	0.943	0.945	0.012	0.462	0.849	0.015	0.002	0.001
Delta-1	.39	0.142	0.025	0.013	1.278	0.510	0.356	0.949	0.952	0.946	0.602	1.000	1.000	0.129	0.017	0.009
Delta-1	.59	0.148	0.029	0.015	2.259	0.900	0.628	0.958	0.952	0.950	0.917	1.000	1.000	0.430	0.056	0.026
Delta-2	.00	0.000	0.000	0.000	0.314	0.054	0.026	1.000	1.000	1.000	0.000	0.000	0.000	0.003	0.000	0.000
Delta-2	.14	0.165	0.030	0.013	0.502	0.169	0.115	0.996	0.951	0.949	0.008	0.437	0.843	0.015	0.002	0.001
Delta-2	.39	0.142	0.025	0.013	1.297	0.511	0.356	0.953	0.953	0.946	0.579	1.000	1.000	0.129	0.017	0.009
Delta-2	.59	0.148	0.029	0.015	2.275	0.901	0.628	0.960	0.952	0.950	0.915	1.000	1.000	0.430	0.056	0.026
BC-Boot	.00	0.000	0.000	0.000	0.458	0.064	0.031	0.991	0.993	0.994	0.009	0.007	0.006	0.003	0.000	0.000
BC-Boot	.14	0.165	0.030	0.013	0.684	0.179	0.119	0.907	0.946	0.948	0.142	0.664	0.895	0.015	0.002	0.001
BC-Boot	.39	0.142	0.025	0.013	1.740	0.532	0.363	0.894	0.940	0.937	0.809	1.000	1.000	0.129	0.017	0.009
BC-Boot	.59	0.148	0.029	0.015	3.263	0.941	0.640	0.852	0.926	0.936	0.970	1.000	1.000	0.430	0.056	0.026
Bayes-1	.00	0.000	0.000	0.000	0.375	0.079	0.049	0.998	1.000	1.000	0.002	0.000	0.000	0.003	0.000	0.000
Bayes-1	.14	0.311	0.071	0.041	0.579	0.214	0.168	0.967	0.984	0.994	0.077	0.412	0.650	0.019	0.002	0.001
Bayes-1	.39	0.292	0.065	0.040	1.457	0.639	0.513	0.925	0.980	0.991	0.754	1.000	1.000	0.187	0.020	0.010
Bayes-1	.59	0.321	0.069	0.041	2.579	1.126	0.904	0.899	0.971	0.989	0.957	1.000	1.000	0.602	0.068	0.030
Bayes-2	.00	0.000	0.000	0.000	0.375	0.079	0.049	0.998	1.000	1.000	0.002	0.000	0.000	0.002	0.000	0.000
Bayes-2	.14	0.133	0.005	-0.005	0.579	0.214	0.168	0.966	0.983	0.994	0.076	0.414	0.649	0.016	0.002	0.001
Bayes-2	.39	0.227	0.049	0.029	1.457	0.639	0.513	0.926	0.980	0.992	0.754	1.000	1.000	0.164	0.019	0.009
Bayes-2	.59	0.255	0.057	0.033	2.578	1.126	0.904	0.899	0.972	0.990	0.955	1.000	1.000	0.597	0.064	0.029

Note. Delta-1 = first-order delta method; Delta-2 = second-order delta method; BC-Boot = bias-corrected nonparametric bootstrap; Bayes-1 = Bayesian bootstrap with mean estimator; Bayes-2 = Bayesian bootstrap with median estimator.

conditions and methods. Supplemental figures for the results are presented in Appendix C.

Results of the MC simulation for a moderated mediation model with categorical mediator variable and continuous endogenous variable are displayed in Table 4.7. From the table, we can see that in small samples at small effect sizes, the BB with median estimator is unbiased, whereas, for medium and large effect sizes, all methods are biased; at larger effect sizes all methods overestimate the true indirect effect. In larger sample sizes, the BB with median estimator underestimates the true indirect at small samples; in other conditions both BC bootstrap and delta methods tend to have lower biases compared to BB methods. In terms of confidence interval lengths, in small samples the BC bootstrap has the widest intervals and the delta methods have the narrowest intervals. On the contrary, in medium and large sample sizes the BB methods have slightly wider intervals than other methods.

For coverage probabilities, only the second-order delta method and BB methods obtain approximately nominal coverage rates or higher in small samples across all effect sizes. For medium and large sample sizes, however, the bootstrap methods obtain approximately nominal coverage or higher, whereas the delta methods have slightly lower coverage rates for small effect sizes. All methods obtain lower than nominal Type I Error rates. With respect to power, the bootstrap methods have higher power for small samples than the delta methods, with the BC bootstrap demonstrating higher power. In medium and large size samples, except for small effect sizes where the BC bootstrap has higher power, all methods have similar power across conditions. Similar to previous simulations, the BB methods have lower power than other methods in large sample sizes with small effects. For MSEs, there are small differences across methods in small samples with large effect sizes,

Table 4.7
Moderated Mediation Model Metrics – Mediator Categorical, Endogenous Continuous

		Empirical Bias			*Confidence Interval Length*			*Coverage Probability*			*Rejection Rate*			*Mean Squared Error*		
Method	*Effect*	*100*	*500*	*1000*	*100*	*500*	*1000*	*100*	*500*	*1000*	*100*	*500*	*1000*	*100*	*500*	*1000*
Delta-1	.00	0.000	0.000	0.000	0.462	0.082	0.040	1.000	1.000	1.000	0.000	0.000	0.000	0.010	0.000	0.000
Delta-1	.14	0.131	0.017	0.006	0.676	0.218	0.148	0.961	0.907	0.927	0.002	0.125	0.536	0.029	0.003	0.001
Delta-1	.39	0.109	0.022	0.013	1.655	0.650	0.453	0.926	0.947	0.950	0.175	1.000	1.000	0.198	0.028	0.014
Delta-1	.59	0.090	0.018	0.009	2.713	1.094	0.764	0.942	0.943	0.945	0.675	1.000	1.000	0.544	0.081	0.039
Delta-2	.00	0.000	0.000	0.000	0.605	0.108	0.053	1.000	1.000	1.000	0.000	0.000	0.000	0.010	0.000	0.000
Delta-2	.14	0.131	0.017	0.006	0.791	0.229	0.152	1.000	0.930	0.936	0.001	0.104	0.497	0.029	0.003	0.001
Delta-2	.39	0.109	0.022	0.013	1.726	0.656	0.455	0.940	0.949	0.951	0.145	1.000	1.000	0.198	0.028	0.014
Delta-2	.59	0.090	0.018	0.009	2.777	1.099	0.766	0.947	0.943	0.945	0.634	1.000	1.000	0.544	0.081	0.039
BC-Boot	.00	0.000	0.000	0.000	0.858	0.126	0.061	0.990	0.992	0.994	0.010	0.008	0.006	0.010	0.000	0.000
BC-Boot	.14	0.131	0.017	0.006	1.086	0.249	0.159	0.948	0.950	0.956	0.061	0.431	0.816	0.029	0.003	0.001
BC-Boot	.39	0.109	0.022	0.013	2.250	0.685	0.464	0.936	0.946	0.948	0.613	1.000	1.000	0.198	0.028	0.014
BC-Boot	.59	0.090	0.018	0.009	3.599	1.142	0.779	0.924	0.943	0.946	0.915	1.000	1.000	0.544	0.081	0.039
Bayes-1	.00	0.000	0.000	0.000	0.717	0.158	0.097	0.998	1.000	1.000	0.002	0.000	0.000	0.013	0.000	0.000
Bayes-1	.14	0.240	0.049	0.027	0.915	0.294	0.222	0.990	0.979	0.992	0.022	0.136	0.324	0.036	0.003	0.002
Bayes-1	.39	0.214	0.051	0.032	1.893	0.814	0.651	0.948	0.981	0.994	0.452	0.999	1.000	0.250	0.030	0.014
Bayes-1	.59	0.193	0.045	0.027	3.025	1.362	1.095	0.943	0.980	0.993	0.862	1.000	1.000	0.712	0.088	0.042
Bayes-2	.00	0.000	0.000	0.000	0.717	0.158	0.097	0.998	1.000	1.000	0.002	0.000	0.000	0.008	0.000	0.000
Bayes-2	.14	0.012	-0.081	-0.071	0.915	0.294	0.222	0.990	0.980	0.992	0.021	0.136	0.326	0.027	0.003	0.001
Bayes-2	.39	0.107	0.019	0.010	1.892	0.814	0.652	0.948	0.982	0.993	0.455	0.999	1.000	0.216	0.029	0.014
Bayes-2	.59	0.121	0.026	0.014	3.024	1.361	1.095	0.942	0.980	0.992	0.862	1.000	1.000	0.622	0.084	0.040

Note. Delta-1 = first-order delta method; Delta-2 = second-order delta method; BC-Boot = bias-corrected nonparametric bootstrap; Bayes-1 = Bayesian bootstrap with mean estimator; Bayes-2 = Bayesian bootstrap with median estimator.

such that the BB with mean estimator has the highest variability, and the delta methods have the lowest variability; across other conditions all methods had similar MSEs. Supplemental figures for the results are presented in Appendix C.

Results of the MC simulation for a moderated mediation model with categorical mediator variable and categorical endogenous variable are displayed in Table 4.8. For small samples, the frequentist methods (i.e., delta methods and BC bootstrap) and BB with median estimator provide the most unbiased estimates, with frequentist methods less biased. For medium and large sample sizes, at small effect sizes the BB with median estimator underestimates the true indirect effect; all other methods demonstrate slight bias, with less bias occurring in larger effect sizes. Bootstrap methods have wider confidence interval lengths than the delta methods for small sample sizes; in medium and large sample sizes, the BB methods have slightly wider intervals than other methods. For coverage probabilities, in small samples all methods have approximately nominal coverage rates, with slightly lower rates for the BC bootstrap and first-order delta method; in medium sample sizes, both the first-order delta method and BC bootstrap obtain lower than nominal coverage rates for small effects. Across other conditions, all methods obtain approximately nominal coverage rates or higher, with BB methods having the highest coverage rates.

All methods have lower than nominal Type I Error rates. With respect to power, in small samples the bootstrap methods have higher power than the delta methods, with the BC bootstrap demonstrating the highest power. In medium and large samples, the BC bootstrap has the highest power, with more prominent differences among small effect

Table 4.8
Moderated Mediation Model Metrics – Mediator Categorical, Endogenous Categorical

		Empirical Bias			*Confidence Interval Length*			*Coverage Probability*			*Rejection Rate*			*Mean Squared Error*		
Method	*Effect*	*100*	*500*	*1000*	*100*	*500*	*1000*	*100*	*500*	*1000*	*100*	*500*	*1000*	*100*	*500*	*1000*
Delta-1	.00	-0.003	0.001	0.000	0.990	0.163	0.081	1.000	1.000	1.000	0.000	0.000	0.000	0.049	0.001	0.000
Delta-1	.14	0.155	0.052	0.027	1.296	0.361	0.240	0.986	0.930	0.936	0.000	0.025	0.133	0.103	0.009	0.004
Delta-1	.39	0.174	0.043	0.013	2.853	1.038	0.712	0.939	0.945	0.947	0.022	0.742	0.971	0.631	0.072	0.034
Delta-1	.59	0.218	0.038	0.016	4.736	1.739	1.200	0.948	0.952	0.950	0.126	0.967	1.000	1.845	0.203	0.095
Delta-2	.00	-0.003	0.001	0.000	1.289	0.216	0.106	1.000	1.000	1.000	0.000	0.000	0.000	0.049	0.001	0.000
Delta-2	.14	0.155	0.052	0.027	1.561	0.390	0.250	1.000	0.978	0.950	0.000	0.020	0.114	0.103	0.009	0.004
Delta-2	.39	0.174	0.043	0.013	3.050	1.054	0.718	0.969	0.949	0.949	0.015	0.727	0.969	0.631	0.072	0.034
Delta-2	.59	0.218	0.038	0.016	4.934	1.754	1.206	0.960	0.954	0.951	0.095	0.966	1.000	1.845	0.203	0.095
BC-Boot	.00	-0.003	0.001	0.000	1.789	0.260	0.124	0.991	0.994	0.994	0.011	0.006	0.006	0.049	0.001	0.000
BC-Boot	.14	0.155	0.052	0.027	2.210	0.441	0.267	0.969	0.919	0.943	0.036	0.177	0.376	0.103	0.009	0.004
BC-Boot	.39	0.174	0.043	0.013	3.958	1.136	0.744	0.940	0.942	0.947	0.268	0.839	0.980	0.631	0.072	0.034
BC-Boot	.59	0.218	0.038	0.016	5.982	1.880	1.246	0.937	0.941	0.945	0.523	0.978	1.000	1.845	0.203	0.095
Bayes-1	.00	-0.003	0.000	0.000	1.717	0.328	0.200	0.998	1.000	1.000	0.002	0.000	0.000	0.072	0.001	0.000
Bayes-1	.14	0.378	0.112	0.066	2.043	0.532	0.387	0.993	0.993	0.995	0.010	0.030	0.054	0.152	0.010	0.004
Bayes-1	.39	0.403	0.099	0.050	3.829	1.362	1.056	0.955	0.981	0.994	0.135	0.663	0.879	0.967	0.083	0.037
Bayes-1	.59	0.473	0.093	0.052	6.222	2.256	1.768	0.938	0.982	0.994	0.364	0.935	0.996	3.638	0.240	0.107
Bayes-2	.00	-0.003	0.000	0.000	1.718	0.327	0.200	0.998	1.000	1.000	0.002	0.000	0.000	0.043	0.001	0.000
Bayes-2	.14	0.087	-0.073	-0.084	2.044	0.533	0.387	0.993	0.993	0.995	0.009	0.030	0.052	0.102	0.008	0.004
Bayes-2	.39	0.223	0.047	0.014	3.825	1.363	1.056	0.954	0.981	0.994	0.137	0.664	0.881	0.766	0.077	0.035
Bayes-2	.59	0.323	0.058	0.029	6.218	2.256	1.768	0.939	0.983	0.995	0.363	0.935	0.996	2.444	0.221	0.101

Note. Delta-1 = first-order delta method; Delta-2 = second-order delta method; BC-Boot = bias-corrected nonparametric bootstrap; Bayes-1 = Bayesian bootstrap with mean estimator; Bayes-2 = Bayesian bootstrap with median estimator.

sizes; the delta methods and BB methods have approximately similar power in these conditions. In small sample sizes, MSEs are similar across methods except at large effect sizes, where the BB with mean estimator has the largest variability and the frequentist methods have the lowest variability. MSEs are similar across methods in medium and large sample sizes. Supplemental figures for the results are presented in Appendix C.

4.2. Study 2

4.2.1. Method. The purpose of the previous MC simulation study is to examine the performance of the BB for indirect effects analysis with complete data. As such, the purpose of Study 2 is to extend the scope of Study 1 by examining the performance of proposed FCS-BB MI algorithm (Algorithm [3.3]) using gradient boosted models for missing data under the same substantive models considered in Study 1. A MC simulation study is conducted to examine the same metrics as in Study 1: (1) the biases, (2) mean squared errors (MSEs), (3) confidence interval coverage probabilities, (4) length of confidence intervals, (5) Type I Error/power of the FCS-BB MI algorithms using gradient boosted models for estimating and testing indirect effects in the presence of missing data. In addition, the fraction of missing information (FMI) is computed (Chapter 2, Equation [2.21]) as

$$\gamma = \frac{\left(\left(\frac{\mathbf{B}_M + \mathbf{B}_M/M}{\bar{\mathbf{U}}_M}\right) + 2\right)/(df + 3)}{1 + \left(\frac{\mathbf{B}_M + \mathbf{B}_M/M}{\bar{\mathbf{U}}_M}\right)}.$$

In the FMI, df are finite, sample-adjusted degrees of freedom given by $df = \frac{\vartheta\varphi}{\vartheta+\varphi}$, where

$$\vartheta = (M-1)\left(1 + \left(\frac{\mathbf{B}_M + \mathbf{B}_M/M}{\bar{\mathbf{U}}_M}\right)^{-2}\right)$$

and

$$\varphi = \frac{n-k+1}{n-k+3}(n-k)\left(1-\frac{\mathbf{B}_M+\mathbf{B}_M/M}{\mathbf{T}_M}\right)$$

and k is the number of parameters fit to the data. For comparison, the relative performance of the imputation algorithms is compared to complete case (CC) analysis (i.e., listwise deletion), mean imputation (Algorithm [2.1]), model-based estimation (MBE) using bootstrap with BC CIs, JM MI (Algorithm [2.3]) or data augmentation (DA), and MICE using the Bayesian imputation algorithms (Algorithm [2.5] for continuous variables, Algorithm [2.6] for binary variables). For our FCS-BB and MICE implementations, we set maximum iterations to 5 due to computational complexity and 10, respectively. For our DA implementation, we used 10,000 iterations and took the values of the last I-step of the algorithm as the imputed values. For imputed values with categorical variables, we thresholded the value using the following rule: If the imputed value is greater than or equal to .5, its final value is 1, otherwise its final value is 0.

For Bayesian (stage one and stage two sampling) and nonparametric bootstrapping, 1,000 bootstrap samples are used. The number of multiply imputed data sets is set to match the percent of simulated missing data described below. The rationale for M is based on Rubin's derivation showing the relationship between asymptotic (ideal) variance of the multiply imputed estimates $\mathbf{T}_\infty$ compared to the finite variance $\mathbf{T}_\infty$

$$\mathbf{T}_M = \left(1+\frac{\gamma}{M}\right)\mathbf{T}_\infty. \qquad (4.4)$$

If a p% percent missing data mechanism is observed (e.g., $p = 20$), a crude estimate of γ is $p/100$ and we recommend that M should set to the integer value of p (e.g., $M = 20$ for 20% missingness). Using this crude estimate for γ and value for M, when applied to (4.4),

$$\begin{aligned} \mathbf{T}_M &= \left(1 + \frac{p/100}{p}\right)\mathbf{T}_\infty \\ &= 1.01\mathbf{T}_\infty, \end{aligned} \tag{4.5}$$

where this equation holds for all p. Equation (4.5) says that the finite M variance $\mathbf{T}_M$ is 1.01 times larger than the ideal variance $\mathbf{T}_\infty$.

The simulation models and parameters are identical to those in Study 1. All parameters are estimated with ML and α will be set to .05 to generate 95% CIs. Based on the simulation results from Study 1, for Bayesian bootstrapping mean estimators are used to estimate indirect effects when both the mediator and endogenous variables are continuous and when the mediator and/or the endogenous variables are categorical and the effect size is small. On the contrary, the median estimator is used to estimate indirect effects when the mediator and/or the endogenous variables are categorical and the sample size is small and in larger samples with larger effect sizes.

A MAR missing data mechanism is simulated for M and Y that depends on X similar to Enders et al. (2014). In particular, starting with the highest value of X, observed values of M are deleted with probability .75 until the desired missingness rate is achieved. The same procedure is independently performed with Y. To extend previous research (Enders et al., 2014), for mediation models a MCAR mechanism is simulated for X such that the missingness probability is approximately the same across all variables. For moderated mediation models, however, MAR mechanisms are simulated for only M and Y since by

creating missingness on M, the interaction effect MW will also contain missingness. Both the PI imputation and JAV method are used for imputing data under the moderated mediation model using the FCS-BB algorithm.

Two missing data percentages are examined: 10% and 20%. The simulation consists of 2 (types of models) x 4 (mediator and endogenous variable type combinations) x 3 (sample sizes) x 4 (regression coefficient effect sizes) x 2 (missing data conditions) = 192 unique conditions. Due to computational complexity, 1,000 replications are simulated per condition. For gradient boosted learners, we use the following hyperparameter specifications given in Table 4.9.

Table 4.9.
Hyperparameters for Boosted Imputation Models

Booster	Iterations	Tree Depth	L2-norm	Learning rate	Subsampling
Linear	10	-	.01	.9	.75
Tree	50	6	.01	.1	.50

4.2.2. Results: Mediation models. Results of the MC simulation for a mediation model with continuous mediator and endogenous variables and missingness rates of 10% and 20% (for each variable) are displayed in Table 4.10 and Table 4.11, respectively. As can be seen, except for mean imputation, all methods are relatively unbiased, with the smaller biases occurring with data augmentation (DA). Compared to other methods, mean imputation has slightly narrower intervals across all conditions and FCS-BB has slightly wider intervals at larger effect sizes. Aside from mean imputation, all methods have approximately nominal or higher than nominal coverage rates across conditions. Type I

Table 4.10
Mediation Model Metrics – Mediator Continuous, Endogenous Continuous, 10% Missingness Per Variable

		Empirical Bias			*Confidence Interval Length*			*Coverage Probability*			*Rejection Rate*			*Mean Squared Error*		
Method	*Effect*	*100*	*500*	*1000*	*100*	*500*	*1000*	*100*	*500*	*1000*	*100*	*500*	*1000*	*100*	*500*	*1000*
CC	.00	0.000	0.000	0.000	0.065	0.012	0.006	1.000	1.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000
CC	.14	-0.013	0.035	0.015	0.106	0.041	0.029	0.922	0.931	0.953	0.011	0.415	0.903	0.001	0.000	0.000
CC	.39	0.017	0.012	0.003	0.260	0.111	0.078	0.935	0.960	0.948	0.693	1.000	1.000	0.005	0.001	0.000
CC	.59	-0.022	-0.003	-0.003	0.381	0.166	0.117	0.922	0.951	0.945	0.994	1.000	1.000	0.010	0.002	0.001
Mean	.00	0.000	0.000	0.000	0.050	0.010	0.005	0.999	1.000	1.000	0.001	0.000	0.000	0.000	0.000	0.000
Mean	.14	-0.155	-0.108	-0.113	0.085	0.034	0.024	0.879	0.884	0.902	0.014	0.485	0.946	0.000	0.000	0.000
Mean	.39	-0.142	-0.121	-0.130	0.209	0.092	0.064	0.868	0.847	0.739	0.759	1.000	1.000	0.003	0.001	0.001
Mean	.59	-0.170	-0.142	-0.139	0.308	0.138	0.097	0.821	0.685	0.495	0.999	1.000	1.000	0.010	0.004	0.003
MBE	.00	0.000	0.000	0.000	0.088	0.017	0.009	0.993	0.997	0.991	0.007	0.003	0.009	0.000	0.000	0.000
MBE	.14	-0.012	0.026	0.018	0.120	0.042	0.028	0.912	0.969	0.965	0.105	0.756	0.982	0.001	0.000	0.000
MBE	.39	0.018	0.015	0.003	0.241	0.103	0.072	0.954	0.955	0.950	0.919	1.000	1.000	0.004	0.001	0.000
MBE	.59	-0.016	-0.002	-0.004	0.333	0.147	0.104	0.943	0.948	0.951	0.999	1.000	1.000	0.007	0.001	0.001
DA	.00	0.000	0.000	0.000	0.069	0.013	0.007	1.000	1.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000
DA	.14	-0.027	-0.010	0.000	0.107	0.039	0.027	0.959	0.933	0.929	0.009	0.420	0.903	0.001	0.000	0.000
DA	.39	-0.007	0.009	-0.001	0.250	0.107	0.075	0.936	0.947	0.937	0.720	1.000	1.000	0.004	0.001	0.000
DA	.59	-0.001	0.000	0.002	0.374	0.160	0.112	0.932	0.956	0.944	0.996	1.000	1.000	0.009	0.002	0.001
MICE	.00	0.000	0.000	0.000	0.069	0.013	0.007	1.000	1.000	0.999	0.000	0.000	0.001	0.000	0.000	0.000
MICE	.14	-0.018	0.023	0.018	0.108	0.040	0.028	0.941	0.931	0.949	0.011	0.449	0.925	0.001	0.000	0.000
MICE	.39	0.011	0.012	0.002	0.255	0.107	0.075	0.935	0.958	0.954	0.710	1.000	1.000	0.004	0.001	0.000
MICE	.59	-0.025	-0.005	-0.004	0.371	0.160	0.112	0.924	0.951	0.946	0.997	1.000	1.000	0.009	0.002	0.001
FCS-BB	.00	0.000	0.000	0.000	0.087	0.022	0.014	0.999	1.000	1.000	0.001	0.000	0.000	0.000	0.000	0.000
FCS-BB	.14	-0.013	-0.019	-0.016	0.121	0.050	0.040	0.971	0.984	0.994	0.052	0.447	0.789	0.001	0.000	0.000
FCS-BB	.39	0.011	0.021	0.012	0.263	0.134	0.109	0.944	0.988	0.995	0.849	1.000	1.000	0.005	0.001	0.000
FCS-BB	.59	-0.002	0.020	0.020	0.388	0.202	0.165	0.933	0.981	0.994	0.999	1.000	1.000	0.010	0.002	0.001

Note. CC = complete case analysis; Mean = mean imputation; MBE = model-based estimation; DA = data augmentation; MICE = multiple imputation by chained equations; FCS-BB = fully conditional specification using Bayesian bootstrap.

Table 4.11
Mediation Model Metrics – Mediator Continuous, Endogenous Continuous, 20% Missingness Per Variable

		Empirical Bias			*Confidence Interval Length*			*Coverage Probability*			*Rejection Rate*			*Mean Squared Error*		
Method	*Effect*	*100*	*500*	*1000*	*100*	*500*	*1000*	*100*	*500*	*1000*	*100*	*500*	*1000*	*100*	*500*	*1000*
CC	.00	0.000	0.000	0.000	0.088	0.017	0.009	1.000	1.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000
CC	.14	-0.042	-0.051	-0.001	0.130	0.047	0.032	0.950	0.910	0.908	0.002	0.199	0.717	0.001	0.000	0.000
CC	.39	-0.021	-0.009	0.000	0.298	0.127	0.089	0.913	0.944	0.946	0.444	1.000	1.000	0.006	0.001	0.001
CC	.59	0.001	-0.002	0.002	0.449	0.191	0.134	0.945	0.946	0.937	0.942	1.000	1.000	0.012	0.002	0.001
Mean	.00	0.000	0.000	0.000	0.052	0.011	0.005	1.000	1.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000
Mean	.14	-0.280	-0.285	-0.262	0.082	0.031	0.022	0.841	0.779	0.742	0.005	0.336	0.822	0.000	0.000	0.000
Mean	.39	-0.291	-0.270	-0.260	0.193	0.085	0.060	0.740	0.501	0.282	0.594	1.000	1.000	0.005	0.002	0.002
Mean	.59	-0.290	-0.269	-0.264	0.290	0.129	0.091	0.647	0.225	0.053	0.981	1.000	1.000	0.016	0.010	0.009
MBE	.00	0.000	0.000	0.000	0.110	0.022	0.011	0.993	0.997	0.997	0.007	0.003	0.003	0.000	0.000	0.000
MBE	.14	-0.007	-0.034	-0.006	0.141	0.046	0.031	0.928	0.951	0.940	0.069	0.597	0.920	0.001	0.000	0.000
MBE	.39	-0.003	-0.005	0.000	0.266	0.112	0.079	0.944	0.956	0.947	0.828	1.000	1.000	0.004	0.001	0.000
MBE	.59	0.010	-0.001	0.001	0.367	0.160	0.113	0.939	0.947	0.950	0.997	1.000	1.000	0.009	0.002	0.001
DA	.00	0.001	0.000	0.000	0.093	0.017	0.008	1.000	0.999	1.000	0.000	0.001	0.000	0.000	0.000	0.000
DA	.14	0.010	0.012	0.016	0.130	0.045	0.031	0.992	0.922	0.924	0.007	0.307	0.807	0.001	0.000	0.000
DA	.39	0.009	-0.008	-0.003	0.282	0.117	0.082	0.937	0.942	0.948	0.574	1.000	1.000	0.005	0.001	0.000
DA	.59	-0.015	-0.004	-0.007	0.416	0.175	0.123	0.926	0.939	0.944	0.978	1.000	1.000	0.011	0.002	0.001
MICE	.00	0.000	0.000	0.000	0.091	0.017	0.009	1.000	1.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000
MICE	.14	-0.039	-0.042	-0.008	0.129	0.044	0.030	0.991	0.913	0.906	0.004	0.264	0.787	0.001	0.000	0.000
MICE	.39	-0.032	-0.011	-0.002	0.282	0.117	0.082	0.931	0.945	0.949	0.521	1.000	1.000	0.005	0.001	0.000
MICE	.59	-0.010	-0.002	0.001	0.418	0.176	0.123	0.940	0.934	0.947	0.978	1.000	1.000	0.011	0.002	0.001
FCS-BB	.00	0.000	0.000	0.000	0.113	0.028	0.017	0.997	1.000	1.000	0.003	0.000	0.000	0.000	0.000	0.000
FCS-BB	.14	-0.037	-0.041	-0.007	0.149	0.056	0.044	0.983	0.981	0.990	0.026	0.278	0.603	0.001	0.000	0.000
FCS-BB	.39	0.023	0.008	0.005	0.309	0.153	0.125	0.942	0.976	0.994	0.710	1.000	1.000	0.006	0.001	0.001
FCS-BB	.59	0.010	0.004	0.001	0.465	0.237	0.193	0.940	0.964	0.986	0.985	1.000	1.000	0.015	0.003	0.002

Note. CC = complete case analysis; Mean = mean imputation; MBE = model-based estimation; DA = data augmentation; MICE = multiple imputation by chained equations; FCS-BB = fully conditional specification using Bayesian bootstrap.

Error levels are below nominal levels for all methods across all conditions. With regards to power in small samples, MBE and FCS-BB tend to have higher power than other methods, with MBE having the highest power. In medium sample sizes, MBE has the highest power across missingness; other methods have similar power with 10% missingness, but with 20% missingness CC has lower power than other methods. In large samples, MBE has the highest power, whereas, FCS-BB has the lowest power; other methods have similar rejection rates with 10% missingness, but at 20% missingness CC has lower power than other methods. Across most conditions, MSEs are comparable across methods. At larger effect sizes, however, MSEs are larger for mean imputation, especially with higher rates of missingness. For MI methods (i.e., DA, MICE, FCS-BB), the fraction of missing information (FMI) is more consistently estimated by the FCS-BB method; DA and MICE overestimate the FMI (see Table 4.12). Supplemental figures for the results are presented in Appendix D.

Initial results of the MC simulation for a mediation model with continuous mediator variable and categorical endogenous variable for the FCS-BB method demonstrated high bias, wide confidence intervals, poor coverage, and low power. As van Buuren (2012) notes, to obtain more accurate imputations, an imputation model should incorporate prediction error and parameter uncertainty (if possible). In the case of linear boosters for categorical variables, asymptotically the models converge to estimates from linear logistic regression models (Bühlmann & Hothorn, 2007). As such, using ideas from Bayesian logistic regression (Gelman et al., 2013), we incorporated parameter uncertainty into the linear gradient boosted imputers (for categorical variables) by following the steps presented in Lines 9 to 13 of Algorithm 2.6. For the remainder of this section, we present

Table 4.12
Mediation Model FMI Metric – Mediator Continuous, Endogenous Continuous

		FMI – 10% Missingness			*FMI – 20% Missingness*		
Method	*Effect*	*100*	*500*	*1000*	*100*	*500*	*1000*
DA	.00	0.149	0.138	0.136	0.278	0.255	0.259
DA	.14	0.163	0.160	0.157	0.295	0.307	0.302
DA	.39	0.167	0.156	0.155	0.310	0.300	0.298
DA	.59	0.164	0.147	0.145	0.311	0.294	0.293
MICE	.00	0.149	0.138	0.137	0.270	0.259	0.263
MICE	.14	0.168	0.162	0.154	0.299	0.308	0.308
MICE	.39	0.177	0.153	0.152	0.322	0.302	0.299
MICE	.59	0.171	0.153	0.147	0.315	0.293	0.291
FCS-BB	.00	0.148	0.143	0.136	0.276	0.266	0.256
FCS-BB	.14	0.148	0.143	0.147	0.278	0.270	0.277
FCS-BB	.39	0.134	0.128	0.129	0.265	0.249	0.250
FCS-BB	.59	0.114	0.113	0.111	0.232	0.226	0.222

Note. DA = data augmentation; MICE = multiple imputation by chained equations; FCS-BB = fully conditional specification using Bayesian bootstrap.

results of both original and modified versions of the FCS-BB for comparison purposes. The results of the MC simulation for a mediation model with continuous mediator variable and categorical and missingness rates of 10% and 20% (for each variable) are displayed in Table 4.13 and Table 4.14, respectively.

For non-null effects, mean imputation, MBE, and FCS-BB are biased in all conditions, with FCS-BB the most biased and mean imputation the least biased. For DA, MICE, CC, and the modified FCS-BB (i.e., FCS-BB*), however, these methods have relatively lower biases than other methods; MICE and FCS-BB have comparable performance. MBE and mean imputation have the narrowest confidence interval lengths across all conditions, whereas, the FCS methods have the widest; other methods have similar lengths across conditions. In terms of coverage probabilities, across all conditions FCS-BB* is the only method that has approximately nominal or above nominal coverage rates. At larger effect sizes mean

Table 4.13
Mediation Model Metrics – Mediator Continuous, Endogenous Categorical, 10% Missingness Per Variable

		Empirical Bias			*Confidence Interval Length*			*Coverage Probability*			*Rejection Rate*			*Mean Squared Error*		
Method	*Effect*	*100*	*500*	*1000*	*100*	*500*	*1000*	*100*	*500*	*1000*	*100*	*500*	*1000*	*100*	*500*	*1000*
CC	.00	0.001	0.000	0.000	0.139	0.026	0.013	1.000	1.000	1.000	0.000	0.000	0.000	0.001	0.000	0.000
CC	.14	0.131	0.043	0.003	0.199	0.068	0.045	0.977	0.934	0.926	0.000	0.092	0.344	0.002	0.000	0.000
CC	.39	0.078	0.021	0.007	0.452	0.184	0.129	0.941	0.947	0.950	0.169	0.959	1.000	0.015	0.002	0.001
CC	.59	0.066	0.003	0.005	0.703	0.292	0.206	0.943	0.947	0.952	0.549	1.000	1.000	0.036	0.006	0.003
Mean	.00	0.001	0.000	0.000	0.114	0.022	0.011	1.000	0.999	1.000	0.000	0.001	0.000	0.001	0.000	0.000
Mean	.14	-0.039	-0.068	-0.117	0.163	0.057	0.038	0.964	0.916	0.909	0.002	0.099	0.392	0.002	0.000	0.000
Mean	.39	-0.070	-0.108	-0.115	0.372	0.154	0.108	0.906	0.893	0.877	0.188	0.975	1.000	0.010	0.002	0.001
Mean	.59	-0.099	-0.127	-0.127	0.566	0.242	0.170	0.909	0.837	0.805	0.636	1.000	1.000	0.023	0.006	0.004
MBE	.00	0.001	0.000	0.000	0.129	0.024	0.012	0.996	0.994	0.995	0.004	0.006	0.005	0.000	0.000	0.000
MBE	.14	-0.312	-0.353	-0.378	0.156	0.045	0.029	0.942	0.909	0.852	0.033	0.303	0.549	0.001	0.000	0.000
MBE	.39	-0.366	-0.388	-0.393	0.260	0.100	0.070	0.879	0.420	0.112	0.409	0.986	1.000	0.007	0.004	0.004
MBE	.59	-0.408	-0.423	-0.420	0.336	0.136	0.095	0.638	0.018	0.001	0.787	1.000	1.000	0.027	0.023	0.022
DA	.00	0.001	0.000	0.000	0.142	0.026	0.013	1.000	1.000	1.000	0.000	0.000	0.000	0.001	0.000	0.000
DA	.14	0.152	-0.036	-0.020	0.196	0.064	0.044	0.997	0.938	0.948	0.003	0.096	0.350	0.002	0.000	0.000
DA	.39	0.027	-0.008	-0.018	0.434	0.176	0.123	0.946	0.949	0.941	0.151	0.964	1.000	0.012	0.002	0.001
DA	.59	0.028	-0.015	-0.023	0.684	0.281	0.196	0.946	0.947	0.952	0.578	1.000	1.000	0.031	0.005	0.002
MICE	.00	0.001	0.000	0.000	0.145	0.027	0.013	1.000	0.999	1.000	0.000	0.001	0.000	0.001	0.000	0.000
MICE	.14	0.125	0.049	0.002	0.198	0.066	0.044	0.998	0.937	0.928	0.002	0.094	0.388	0.002	0.000	0.000
MICE	.39	0.065	0.017	0.003	0.440	0.177	0.124	0.952	0.946	0.947	0.164	0.975	1.000	0.013	0.002	0.001
MICE	.59	0.049	0.002	0.001	0.684	0.282	0.198	0.947	0.940	0.951	0.570	1.000	1.000	0.031	0.005	0.002
FCS-BB	.00	0.001	0.000	0.000	0.252	0.070	0.047	0.998	0.999	1.000	0.002	0.001	0.000	0.001	0.000	0.000
FCS-BB	.14	0.242	0.430	0.498	0.311	0.124	0.097	0.991	0.985	0.990	0.021	0.100	0.182	0.003	0.001	0.000
FCS-BB	.39	0.327	0.337	0.337	0.597	0.265	0.210	0.939	0.919	0.905	0.353	0.946	0.996	0.023	0.006	0.004
FCS-BB	.59	0.311	0.242	0.239	0.896	0.404	0.325	0.934	0.921	0.907	0.695	1.000	1.000	0.058	0.014	0.010
FCS-BB*	.00	0.001	0.000	0.000	0.195	0.045	0.028	1.000	1.000	1.000	0.000	0.000	0.000	0.001	0.000	0.000
FCS-BB*	.14	-0.038	-0.017	-0.005	0.248	0.085	0.065	0.987	0.984	0.990	0.017	0.085	0.183	0.002	0.000	0.000
FCS-BB*	.39	0.051	0.013	0.000	0.490	0.225	0.180	0.953	0.976	0.994	0.341	0.925	0.998	0.013	0.002	0.001
FCS-BB*	.59	0.042	0.004	0.019	0.778	0.366	0.292	0.957	0.981	0.993	0.689	0.999	1.000	0.037	0.006	0.003

Note. CC = complete case analysis; Mean = mean imputation; MBE = model-based estimation; DA = data augmentation; MICE = multiple imputation by chained equations; FCS-BB = fully conditional specification using Bayesian bootstrap; FCS-BB* = modified fully conditional specification using Bayesian bootstrap.

Table 4.14
Mediation Model Metrics – Mediator Continuous, Endogenous Categorical, 20% Missingness Per Variable

		Empirical Bias			*Confidence Interval Length*			*Coverage Probability*			*Rejection Rate*			*Mean Squared Error*		
Method	*Effect*	*100*	*500*	*1000*	*100*	*500*	*1000*	*100*	*500*	*1000*	*100*	*500*	*1000*	*100*	*500*	*1000*
CC	.00	-0.001	0.000	0.000	0.187	0.034	0.017	1.000	1.000	1.000	0.000	0.000	0.000	0.002	0.000	0.000
CC	.14	-0.040	0.016	0.015	0.242	0.078	0.053	0.985	0.936	0.946	0.002	0.039	0.208	0.004	0.000	0.000
CC	.39	0.064	-0.001	0.002	0.529	0.211	0.148	0.925	0.941	0.947	0.069	0.873	0.995	0.019	0.003	0.001
CC	.59	0.090	0.023	0.015	0.841	0.339	0.237	0.940	0.951	0.947	0.352	0.998	1.000	0.055	0.008	0.004
Mean	.00	-0.001	0.000	0.000	0.125	0.023	0.011	0.999	1.000	1.000	0.001	0.000	0.000	0.001	0.000	0.000
Mean	.14	-0.241	-0.211	-0.218	0.163	0.055	0.037	0.957	0.881	0.873	0.002	0.062	0.264	0.002	0.000	0.000
Mean	.39	-0.206	-0.238	-0.226	0.358	0.147	0.103	0.861	0.769	0.715	0.109	0.935	1.000	0.010	0.003	0.002
Mean	.59	-0.207	-0.223	-0.228	0.543	0.231	0.162	0.837	0.682	0.498	0.496	1.000	1.000	0.026	0.009	0.008
MBE	.00	0.000	0.000	0.000	0.171	0.031	0.015	0.997	0.994	0.996	0.003	0.006	0.004	0.001	0.000	0.000
MBE	.14	-0.405	-0.375	-0.372	0.195	0.053	0.034	0.967	0.907	0.886	0.015	0.218	0.455	0.001	0.000	0.000
MBE	.39	-0.375	-0.402	-0.395	0.309	0.115	0.080	0.904	0.502	0.212	0.286	0.936	1.000	0.008	0.005	0.004
MBE	.59	-0.401	-0.411	-0.417	0.403	0.156	0.109	0.736	0.077	0.001	0.640	1.000	1.000	0.029	0.022	0.022
DA	.00	0.001	0.000	0.000	0.186	0.033	0.016	1.000	1.000	1.000	0.000	0.000	0.000	0.001	0.000	0.000
DA	.14	-0.007	0.040	-0.036	0.244	0.073	0.048	1.000	0.942	0.934	0.000	0.065	0.261	0.003	0.000	0.000
DA	.39	0.021	-0.037	-0.041	0.494	0.190	0.133	0.943	0.945	0.951	0.067	0.927	0.998	0.014	0.002	0.001
DA	.59	0.017	-0.046	-0.050	0.766	0.305	0.212	0.950	0.937	0.930	0.388	0.999	1.000	0.037	0.006	0.003
MICE	.00	-0.001	0.000	0.000	0.191	0.034	0.017	0.999	1.000	0.999	0.001	0.000	0.001	0.001	0.000	0.000
MICE	.14	-0.066	0.012	0.002	0.240	0.073	0.048	0.999	0.939	0.950	0.001	0.052	0.259	0.003	0.000	0.000
MICE	.39	0.033	-0.009	-0.004	0.496	0.194	0.136	0.935	0.944	0.954	0.071	0.926	1.000	0.015	0.003	0.001
MICE	.59	0.060	0.014	0.008	0.781	0.312	0.218	0.955	0.955	0.948	0.417	1.000	1.000	0.041	0.006	0.003
FCS-BB	.00	-0.001	0.000	0.000	0.391	0.112	0.078	0.998	1.000	1.000	0.002	0.000	0.000	0.002	0.000	0.000
FCS-BB	.14	0.284	0.795	0.998	0.461	0.183	0.147	0.992	0.988	0.986	0.014	0.062	0.110	0.006	0.001	0.001
FCS-BB	.39	0.570	0.664	0.708	0.855	0.371	0.289	0.956	0.836	0.696	0.265	0.843	0.984	0.043	0.016	0.014
FCS-BB	.59	0.602	0.540	0.536	1.304	0.551	0.437	0.899	0.738	0.584	0.523	0.993	1.000	0.137	0.047	0.041
FCS-BB*	.00	0.000	0.000	0.000	0.256	0.056	0.035	0.997	0.999	1.000	0.003	0.001	0.000	0.001	0.000	0.000
FCS-BB*	.14	-0.060	-0.015	-0.005	0.313	0.097	0.074	0.996	0.988	0.993	0.015	0.064	0.139	0.002	0.000	0.000
FCS-BB*	.39	0.035	0.009	0.004	0.616	0.258	0.205	0.944	0.973	0.990	0.293	0.887	0.987	0.026	0.003	0.001
FCS-BB*	.59	0.051	0.021	0.016	0.973	0.424	0.339	0.946	0.975	0.989	0.610	0.995	1.000	0.065	0.009	0.004

Note. CC = complete case analysis; Mean = mean imputation; MBE = model-based estimation; DA = data augmentation; MICE = multiple imputation by chained equations; FCS-BB = fully conditional specification using Bayesian bootstrap; FCS-BB* = modified fully conditional specification using Bayesian bootstrap.

imputation, FCS-BB, and MBE have poor coverage performance, with the lowest coverages for MBE and the highest coverages for mean imputation; DA, MICE, and CC have similar coverage rates across conditions.

All methods have lower than nominal Type I Error rates. Power is highest among MBE across all conditions; generally, all other methods have comparable power except in small samples with medium effect sizes and in large samples with small effect sizes. For the former, FCS-BB and FCS-BB* have higher power compared to other methods, whereas for the latter, these two methods have lower power. At larger effect sizes, MSEs are highest among the FCS-BB method; other methods have comparable mean squared errors. With regards to MI methods, the FMI estimates (see Table 4.15) are comparable across all

Table 4.15
Mediation Model FMI Metric – Mediator Continuous, Endogenous Categorical

		FMI – 10% Missingness			*FMI – 20% Missingness*		
Method	*Effect*	*100*	*500*	*1000*	*100*	*500*	*1000*
DA	.00	0.140	0.135	0.138	0.266	0.248	0.244
DA	.14	0.150	0.139	0.143	0.275	0.270	0.269
DA	.39	0.156	0.145	0.147	0.297	0.279	0.278
DA	.59	0.167	0.154	0.149	0.307	0.292	0.286
MICE	.00	0.146	0.141	0.137	0.272	0.257	0.257
MICE	.14	0.154	0.146	0.144	0.283	0.285	0.282
MICE	.39	0.162	0.146	0.146	0.305	0.292	0.292
MICE	.59	0.168	0.154	0.148	0.315	0.295	0.298
FCS-BB	.00	0.167	0.198	0.215	0.304	0.366	0.381
FCS-BB	.14	0.168	0.192	0.219	0.321	0.384	0.409
FCS-BB	.39	0.133	0.117	0.105	0.285	0.270	0.260
FCS-BB	.59	0.118	0.088	0.085	0.236	0.204	0.197
FCS-BB*	.00	0.148	0.149	0.141	0.281	0.272	0.258
FCS-BB*	.14	0.151	0.141	0.139	0.285	0.278	0.271
FCS-BB*	.39	0.151	0.137	0.136	0.283	0.265	0.266
FCS-BB*	.59	0.149	0.136	0.136	0.278	0.264	0.264

Note. DA = data augmentation; MICE = multiple imputation by chained equations; FCS-BB = fully conditional specification using Bayesian bootstrap; FCS-BB* = modified fully conditional specification using Bayesian bootstrap.

methods except FCS-BB, which overestimates the amount of missing information at smaller effect sizes. Supplemental figures for the results are presented in Appendix D.

Results of the MC simulation for a mediation model with categorical mediator variable and continuous endogenous variable and missingness rates of 10% and 20% (for each variable) are displayed in Table 4.16 and Table 4.17, respectively. Similar to the previous simulation, for non-null effects, FCS-BB, mean imputation, and MBE methods are biased, with FCS-BB exhibiting the most bias and mean imputation the least bias; other methods demonstrate slight bias, with higher bias occurring with small effect sizes. In particular, MICE and FCS-BB* have comparable levels of bias. Confidence interval lengths are widest for the FCS-BB method and narrowest for the MBE method; other methods demonstrate similar trends in lengths with FCS-BB* demonstrating slightly wider intervals than other methods. With regards to coverage probabilities, FCS-BB* is the only method that has nominal or above nominal coverage rates across all conditions. At larger effect sizes, mean imputation, FCS-BB, and MBE tend to have poor coverage performance, with the lowest coverages for MBE and the highest coverages for FCS-BB; all other methods have similar nominal coverage rates across conditions. All methods have lower than nominal Type I Error rates. Power is highest among MBE and in general, comparable across other methods and conditions, except with small samples. In small samples with larger effects, FCS-BB and FCS-BB* have higher power. For MSEs, in smaller samples with larger effect sizes, both FCS-BB and MBE have larger MSEs compared to other methods; across conditions other methods have comparable MSEs. With regards to FMI, estimates are comparable across all methods except FCS-BB, which overestimates the amount of missing

Table 4.16
Mediation Model Metrics – Mediator Categorical, Endogenous Continuous, 10% Missingness Per Variable

		Empirical Bias			*Confidence Interval Length*			*Coverage Probability*			*Rejection Rate*			*Mean Squared Error*		
Method	*Effect*	*100*	*500*	*1000*	*100*	*500*	*1000*	*100*	*500*	*1000*	*100*	*500*	*1000*	*100*	*500*	*1000*
CC	.00	-0.001	0.000	0.000	0.277	0.050	0.025	1.000	1.000	1.000	0.000	0.000	0.000	0.003	0.000	0.000
CC	.14	0.002	0.005	-0.020	0.327	0.091	0.060	0.984	0.900	0.895	0.002	0.014	0.081	0.006	0.001	0.000
CC	.39	0.002	0.004	0.013	0.582	0.233	0.163	0.912	0.932	0.949	0.028	0.847	1.000	0.023	0.003	0.002
CC	.59	0.045	-0.006	-0.005	0.875	0.361	0.254	0.915	0.946	0.958	0.182	0.999	1.000	0.057	0.009	0.004
Mean	.00	-0.001	0.000	0.000	0.216	0.039	0.020	1.000	1.000	1.000	0.000	0.000	0.000	0.002	0.000	0.000
Mean	.14	-0.146	-0.196	-0.213	0.254	0.072	0.048	0.984	0.850	0.843	0.000	0.014	0.074	0.004	0.000	0.000
Mean	.39	-0.226	-0.207	-0.196	0.447	0.184	0.129	0.848	0.834	0.801	0.020	0.861	1.000	0.014	0.003	0.002
Mean	.59	-0.199	-0.221	-0.218	0.670	0.282	0.199	0.823	0.740	0.638	0.197	0.999	1.000	0.038	0.011	0.008
MBE	.00	0.000	0.000	0.000	0.156	0.029	0.014	0.994	0.989	0.992	0.006	0.011	0.008	0.000	0.000	0.000
MBE	.14	-0.618	-0.621	-0.630	0.169	0.041	0.025	0.948	0.778	0.638	0.012	0.112	0.309	0.001	0.000	0.000
MBE	.39	-0.653	-0.641	-0.637	0.228	0.082	0.056	0.683	0.040	0.000	0.177	0.952	1.000	0.012	0.010	0.010
MBE	.59	-0.664	-0.666	-0.663	0.282	0.110	0.077	0.229	0.000	0.000	0.535	0.999	1.000	0.058	0.055	0.054
DA	.00	-0.001	0.000	0.000	0.281	0.051	0.025	1.000	1.000	1.000	0.000	0.000	0.000	0.003	0.000	0.000
DA	.14	0.024	-0.034	-0.069	0.327	0.088	0.057	0.999	0.912	0.898	0.001	0.012	0.081	0.005	0.000	0.000
DA	.39	-0.056	-0.045	-0.033	0.561	0.221	0.154	0.911	0.933	0.934	0.020	0.853	1.000	0.019	0.003	0.001
DA	.59	-0.012	-0.052	-0.054	0.842	0.342	0.239	0.915	0.938	0.940	0.167	0.999	1.000	0.048	0.008	0.003
MICE	.00	-0.001	0.000	0.000	0.288	0.052	0.026	1.000	1.000	1.000	0.000	0.000	0.000	0.003	0.000	0.000
MICE	.14	0.055	0.005	-0.031	0.337	0.090	0.058	0.999	0.918	0.908	0.002	0.015	0.086	0.006	0.001	0.000
MICE	.39	-0.009	-0.002	0.010	0.576	0.226	0.158	0.915	0.938	0.952	0.027	0.875	1.000	0.021	0.003	0.002
MICE	.59	0.035	-0.006	-0.006	0.866	0.351	0.246	0.908	0.946	0.955	0.192	0.999	1.000	0.055	0.008	0.004
FCS-BB	.00	-0.001	0.000	0.000	0.509	0.145	0.101	0.999	1.000	1.000	0.001	0.000	0.000	0.004	0.000	0.000
FCS-BB	.14	0.155	0.333	0.434	0.556	0.188	0.143	0.996	0.996	0.998	0.005	0.015	0.032	0.008	0.001	0.001
FCS-BB	.39	0.150	0.268	0.301	0.804	0.341	0.269	0.973	0.971	0.969	0.105	0.842	0.996	0.030	0.006	0.004
FCS-BB	.59	0.209	0.185	0.189	1.099	0.491	0.392	0.948	0.958	0.971	0.452	0.998	1.000	0.077	0.014	0.009
FCS-BB*	.00	0.000	0.000	0.000	0.380	0.090	0.056	0.997	1.000	1.000	0.003	0.000	0.000	0.002	0.000	0.000
FCS-BB*	.14	-0.051	-0.007	-0.032	0.428	0.129	0.094	0.995	0.996	0.996	0.006	0.018	0.033	0.004	0.000	0.000
FCS-BB*	.39	-0.005	-0.002	-0.018	0.679	0.281	0.226	0.962	0.978	0.996	0.150	0.850	0.993	0.024	0.004	0.002
FCS-BB*	.59	-0.008	-0.007	-0.010	0.952	0.440	0.355	0.947	0.983	0.990	0.452	0.999	1.000	0.052	0.008	0.004

Note. CC = complete case analysis; Mean = mean imputation; MBE = model-based estimation; DA = data augmentation; MICE = multiple imputation by chained equations; FCS-BB = fully conditional specification using Bayesian bootstrap; FCS-BB* = modified fully conditional specification using Bayesian bootstrap.

Table 4.17
Mediation Model Metrics – Mediator Categorical, Endogenous Continuous, 20% Missingness Per Variable

		Empirical Bias			*Confidence Interval Length*			*Coverage Probability*			*Rejection Rate*			*Mean Squared Error*		
Method	*Effect*	*100*	*500*	*1000*	*100*	*500*	*1000*	*100*	*500*	*1000*	*100*	*500*	*1000*	*100*	*500*	*1000*
CC	.00	0.002	0.000	0.000	0.383	0.068	0.033	1.000	1.000	1.000	0.000	0.000	0.000	0.006	0.000	0.000
CC	.14	0.114	-0.057	0.022	0.419	0.106	0.071	0.996	0.917	0.914	0.000	0.004	0.043	0.009	0.001	0.000
CC	.39	-0.016	0.020	0.013	0.703	0.270	0.188	0.901	0.924	0.936	0.007	0.655	0.987	0.033	0.005	0.002
CC	.59	0.055	0.014	0.005	1.046	0.421	0.293	0.926	0.949	0.945	0.072	0.994	1.000	0.070	0.011	0.006
Mean	.00	-0.002	0.000	0.000	0.235	0.042	0.021	1.000	1.000	1.000	0.000	0.000	0.000	0.002	0.000	0.000
Mean	.14	-0.302	-0.383	-0.328	0.261	0.068	0.045	0.982	0.784	0.791	0.000	0.005	0.057	0.004	0.000	0.000
Mean	.39	-0.406	-0.353	-0.349	0.423	0.170	0.119	0.775	0.676	0.534	0.011	0.655	0.986	0.015	0.005	0.004
Mean	.59	-0.344	-0.361	-0.370	0.632	0.260	0.181	0.755	0.483	0.234	0.107	0.992	1.000	0.041	0.020	0.019
MBE	.00	0.001	0.000	0.000	0.203	0.037	0.018	0.996	0.997	0.992	0.004	0.003	0.008	0.001	0.000	0.000
MBE	.14	-0.619	-0.664	-0.631	0.211	0.048	0.029	0.972	0.761	0.712	0.009	0.064	0.211	0.001	0.000	0.000
MBE	.39	-0.658	-0.649	-0.648	0.275	0.092	0.063	0.727	0.074	0.000	0.122	0.862	0.994	0.013	0.010	0.010
MBE	.59	-0.671	-0.674	-0.672	0.332	0.123	0.086	0.353	0.000	0.000	0.383	0.999	1.000	0.060	0.056	0.055
DA	.00	0.000	0.000	0.000	0.371	0.065	0.032	1.000	1.000	1.000	0.000	0.000	0.000	0.004	0.000	0.000
DA	.14	-0.057	-0.124	-0.059	0.408	0.098	0.064	1.000	0.953	0.916	0.000	0.004	0.052	0.006	0.000	0.000
DA	.39	-0.136	-0.078	-0.071	0.639	0.239	0.166	0.921	0.926	0.924	0.004	0.687	0.991	0.022	0.004	0.002
DA	.59	-0.048	-0.077	-0.089	0.947	0.373	0.257	0.918	0.921	0.884	0.075	0.998	1.000	0.048	0.009	0.005
MICE	.00	0.000	0.000	0.000	0.390	0.069	0.034	1.000	1.000	1.000	0.000	0.000	0.000	0.004	0.000	0.000
MICE	.14	0.032	-0.045	0.027	0.424	0.104	0.067	1.000	0.964	0.928	0.000	0.006	0.062	0.007	0.001	0.000
MICE	.39	-0.058	0.006	0.013	0.677	0.253	0.177	0.932	0.934	0.945	0.007	0.719	0.990	0.025	0.004	0.002
MICE	.59	0.038	0.017	0.004	1.004	0.397	0.274	0.937	0.944	0.929	0.091	0.997	1.000	0.058	0.010	0.005
FCS-BB	.00	0.000	0.001	0.000	0.830	0.268	0.197	0.999	1.000	1.000	0.001	0.000	0.000	0.010	0.001	0.000
FCS-BB	.14	0.511	0.946	1.460	0.869	0.313	0.250	0.999	0.996	0.992	0.002	0.009	0.023	0.016	0.003	0.002
FCS-BB	.39	0.372	0.693	0.747	1.167	0.478	0.371	0.986	0.916	0.835	0.064	0.726	0.963	0.058	0.020	0.017
FCS-BB	.59	0.468	0.492	0.485	1.521	0.645	0.507	0.966	0.850	0.777	0.307	0.996	1.000	0.134	0.045	0.037
FCS-BB*	.00	-0.003	0.000	0.000	0.488	0.111	0.069	0.998	1.000	1.000	0.002	0.000	0.000	0.003	0.000	0.000
FCS-BB*	.14	-0.032	-0.046	-0.031	0.525	0.150	0.106	0.999	0.995	0.995	0.001	0.015	0.019	0.005	0.001	0.000
FCS-BB*	.39	-0.055	-0.035	-0.033	0.809	0.319	0.253	0.971	0.984	0.995	0.080	0.730	0.971	0.027	0.004	0.002
FCS-BB*	.59	0.013	-0.013	0.000	1.148	0.497	0.402	0.953	0.979	0.993	0.323	0.996	1.000	0.073	0.010	0.005

Note. CC = complete case analysis; Mean = mean imputation; MBE = model-based estimation; DA = data augmentation; MICE = multiple imputation by chained equations; FCS-BB = fully conditional specification using Bayesian bootstrap; FCS-BB* = modified fully conditional specification using Bayesian bootstrap.

information at smaller effect sizes (see Table 4.18). Supplemental figures for the results are presented in Appendix D.

Table 4.18
Mediation Model FMI Metric – Mediator Categorical, Endogenous Continuous

		FMI – 10% Missingness			*FMI – 20% Missingness*		
Method	*Effect*	*100*	*500*	*1000*	*100*	*500*	*1000*
DA	.00	0.141	0.132	0.128	0.254	0.242	0.240
DA	.14	0.145	0.150	0.150	0.267	0.273	0.285
DA	.39	0.163	0.161	0.159	0.289	0.299	0.296
DA	.59	0.167	0.165	0.165	0.303	0.306	0.303
MICE	.00	0.143	0.135	0.135	0.269	0.258	0.257
MICE	.14	0.151	0.151	0.154	0.277	0.285	0.294
MICE	.39	0.166	0.161	0.159	0.301	0.313	0.317
MICE	.59	0.171	0.164	0.163	0.317	0.318	0.314
FCS-BB	.00	0.166	0.193	0.207	0.287	0.352	0.367
FCS-BB	.14	0.155	0.177	0.183	0.293	0.329	0.335
FCS-BB	.39	0.141	0.120	0.116	0.271	0.245	0.226
FCS-BB	.59	0.128	0.107	0.099	0.244	0.212	0.206
FCS-BB*	.00	0.158	0.147	0.140	0.298	0.271	0.266
FCS-BB*	.14	0.161	0.148	0.148	0.296	0.280	0.276
FCS-BB*	.39	0.160	0.153	0.158	0.304	0.293	0.292
FCS-BB*	.59	0.169	0.156	0.157	0.300	0.294	0.288

Note. DA = data augmentation; MICE = multiple imputation by chained equations; FCS-BB = fully conditional specification using Bayesian bootstrap; FCS-BB* = modified fully conditional specification using Bayesian bootstrap.

Results of the MC simulation for a mediation model with categorical mediator variable and categorical endogenous variable and missingness rates of 10% and 20% (for each variable) are displayed in Table 4.19 and Table 4.20, respectively. As can be seen in the tables, FCS-BB and MBE have the largest biases, whereas other methods are slightly biased in small samples and approximately unbiased in other conditions. Confidence interval lengths are widest for FCS-BB and lowest for MBE; in small samples the FCS-BB* has slightly wider intervals than other methods, whereas mean imputation has slightly

narrowest intervals. In other conditions, confidence intervals are relatively similar across methods. Similar to previous results, FCS-BB* is the only method that has nominal or above nominal coverage rates across all conditions. Coverage probabilities are poor for both mean imputation and MBE, with lower probabilities for MBE; FCS-BB also has poor coverage in medium and large samples with larger effect sizes and more missingness. Across other conditions, DA, MICE, and CC have approximately nominal or higher coverage probabilities.

Echoing previous results, all methods have lower than nominal Type I Error rates. For power, in small samples relative to other methods, MBE and FCS methods have slightly more power at medium and large effect sizes. In medium sample sizes MBE has higher power, but with 20% missingness mean imputation has comparable power at medium effect sizes and FCS-BB* has comparable power at large effect sizes. In large samples with 10% missingness, relative to other methods MBE has higher power across all conditions. Mean imputation, however, performs comparable to MBE at small effect sizes, but has lower power as effect size increases compared to other methods. At medium effect sizes, Mice, DA, and CC have higher power than FCS-BB and FCS-BB*. In large samples with 20% missingness, mean imputation outperforms all methods at small effect sizes, but power decrease compared to other methods as effect size increases. In large samples with 20% missingness at medium effect sizes, MBE and MICE perform slightly better than other methods, whereas, DA, CC, and FCS-BB* perform similarly, whereas FCS-BB performs the worst. MSEs are highest for FCS-BB across all conditions and lowest for MBE. With regards to other methods, FCS-BB* has slightly more variability in smaller samples and mean imputation has slightly lower variability; in other conditions methods have approximately

Table 4.19
Mediation Model Metrics – Mediator Categorical, Endogenous Categorical, 10% Missingness Per Variable

		Empirical Bias			*Confidence Interval Length*			*Coverage Probability*			*Rejection Rate*			*Mean Squared Error*		
Method	*Effect*	*100*	*500*	*1000*	*100*	*500*	*1000*	*100*	*500*	*1000*	*100*	*500*	*1000*	*100*	*500*	*1000*
CC	.00	0.001	0.000	0.000	0.553	0.106	0.051	1.000	1.000	1.000	0.000	0.000	0.000	0.015	0.001	0.000
CC	.14	-0.112	-0.002	-0.055	0.631	0.156	0.098	1.000	0.964	0.924	0.000	0.003	0.016	0.025	0.001	0.001
CC	.39	0.108	0.015	0.023	1.038	0.372	0.260	0.949	0.950	0.939	0.005	0.281	0.709	0.074	0.009	0.004
CC	.59	0.134	-0.004	0.020	1.516	0.574	0.406	0.945	0.940	0.948	0.035	0.712	0.973	0.171	0.022	0.011
Mean	.00	-0.001	0.000	0.000	0.489	0.123	0.075	1.000	0.999	0.996	0.000	0.001	0.004	0.013	0.001	0.000
Mean	.14	-0.473	-0.140	-0.127	0.548	0.154	0.102	0.997	0.941	0.755	0.000	0.025	0.111	0.019	0.003	0.002
Mean	.39	-0.097	-0.181	-0.140	0.844	0.315	0.221	0.925	0.761	0.641	0.004	0.346	0.559	0.052	0.014	0.011
Mean	.59	-0.106	-0.168	-0.160	1.211	0.474	0.336	0.885	0.754	0.618	0.043	0.642	0.775	0.113	0.033	0.029
MBE	.00	0.000	0.000	0.000	0.198	0.038	0.019	0.994	0.995	0.993	0.006	0.005	0.007	0.001	0.000	0.000
MBE	.14	-0.769	-0.760	-0.769	0.210	0.047	0.028	0.982	0.699	0.545	0.015	0.038	0.130	0.001	0.000	0.000
MBE	.39	-0.761	-0.769	-0.762	0.264	0.086	0.059	0.630	0.017	0.000	0.062	0.509	0.815	0.016	0.014	0.014
MBE	.59	-0.771	-0.785	-0.781	0.323	0.117	0.081	0.206	0.000	0.000	0.195	0.805	0.978	0.078	0.076	0.074
DA	.00	-0.001	0.000	0.000	0.564	0.105	0.050	1.000	1.000	1.000	0.000	0.000	0.000	0.012	0.000	0.000
DA	.14	-0.104	-0.070	-0.082	0.628	0.149	0.093	1.000	0.985	0.929	0.000	0.001	0.016	0.018	0.001	0.001
DA	.39	0.010	-0.051	-0.041	0.989	0.348	0.242	0.963	0.943	0.929	0.004	0.256	0.732	0.059	0.007	0.004
DA	.59	0.043	-0.065	-0.044	1.435	0.536	0.377	0.937	0.939	0.953	0.028	0.734	0.979	0.136	0.018	0.009
MICE	.00	-0.001	0.000	0.000	0.580	0.109	0.052	1.000	1.000	1.000	0.000	0.000	0.000	0.013	0.000	0.000
MICE	.14	-0.076	-0.021	-0.019	0.646	0.154	0.096	1.000	0.993	0.938	0.000	0.002	0.020	0.021	0.001	0.001
MICE	.39	0.080	0.008	0.017	1.019	0.358	0.249	0.968	0.948	0.939	0.008	0.292	0.759	0.067	0.008	0.004
MICE	.59	0.115	-0.004	0.021	1.487	0.550	0.390	0.942	0.941	0.954	0.032	0.757	0.978	0.160	0.020	0.010
FCS-BB	.00	-0.003	0.000	0.000	1.395	0.480	0.362	0.998	1.000	1.000	0.002	0.000	0.000	0.028	0.002	0.001
FCS-BB	.14	0.384	1.068	1.512	1.483	0.554	0.445	0.996	0.999	1.000	0.005	0.002	0.000	0.046	0.006	0.005
FCS-BB	.39	0.622	0.900	1.033	1.996	0.824	0.664	0.986	0.957	0.934	0.035	0.264	0.505	0.161	0.043	0.037
FCS-BB	.59	0.650	0.627	0.680	2.579	1.072	0.881	0.972	0.921	0.893	0.126	0.665	0.895	0.369	0.091	0.079
FCS-BB*	.00	0.000	0.000	0.000	0.766	0.172	0.108	0.998	1.000	1.000	0.002	0.000	0.000	0.008	0.000	0.000
FCS-BB*	.14	0.041	-0.034	-0.024	0.831	0.220	0.154	0.997	1.000	1.000	0.006	0.002	0.005	0.012	0.001	0.000
FCS-BB*	.39	-0.020	-0.019	-0.013	1.186	0.440	0.347	0.984	0.981	0.996	0.042	0.314	0.589	0.051	0.008	0.003
FCS-BB*	.59	0.025	-0.018	-0.011	1.674	0.673	0.541	0.950	0.976	0.994	0.164	0.737	0.946	0.158	0.020	0.010

Note. CC = complete case analysis; Mean = mean imputation; MBE = model-based estimation; DA = data augmentation; MICE = multiple imputation by chained equations; FCS-BB = fully conditional specification using Bayesian bootstrap; FCS-BB* = modified fully conditional specification using Bayesian bootstrap.

Table 4.20
Mediation Model Metrics – Mediator Categorical, Endogenous Categorical, 20% Missingness Per Variable

		Empirical Bias			*Confidence Interval Length*			*Coverage Probability*			*Rejection Rate*			*Mean Squared Error*		
Method	*Effect*	*100*	*500*	*1000*	*100*	*500*	*1000*	*100*	*500*	*1000*	*100*	*500*	*1000*	*100*	*500*	*1000*
CC	.00	-0.002	-0.001	0.000	0.751	0.133	0.065	1.000	1.000	1.000	0.000	0.000	0.000	0.024	0.001	0.000
CC	.14	0.324	-0.003	-0.035	0.835	0.193	0.115	0.998	0.980	0.940	0.000	0.002	0.008	0.039	0.002	0.001
CC	.39	0.140	0.016	-0.003	1.256	0.435	0.296	0.957	0.926	0.942	0.004	0.165	0.528	0.123	0.013	0.006
CC	.59	0.183	0.037	0.007	1.824	0.679	0.464	0.939	0.932	0.942	0.009	0.576	0.904	0.242	0.032	0.015
Mean	.00	-0.006	-0.001	-0.001	0.631	0.197	0.133	1.000	0.997	0.992	0.000	0.003	0.008	0.024	0.002	0.001
Mean	.14	-0.229	-0.163	-0.263	0.672	0.216	0.148	0.998	0.958	0.696	0.002	0.032	0.166	0.031	0.006	0.004
Mean	.39	-0.148	-0.272	-0.295	0.912	0.341	0.234	0.914	0.573	0.438	0.008	0.397	0.549	0.076	0.028	0.026
Mean	.59	-0.261	-0.318	-0.356	1.224	0.477	0.326	0.802	0.541	0.440	0.041	0.511	0.502	0.162	0.069	0.065
MBE	.00	-0.001	0.000	0.000	0.258	0.048	0.024	0.999	0.997	0.997	0.001	0.003	0.003	0.001	0.000	0.000
MBE	.14	-0.699	-0.754	-0.764	0.271	0.059	0.033	0.987	0.718	0.647	0.005	0.035	0.068	0.002	0.000	0.000
MBE	.39	-0.761	-0.770	-0.770	0.322	0.101	0.068	0.688	0.054	0.000	0.043	0.399	0.674	0.017	0.014	0.014
MBE	.59	-0.770	-0.780	-0.783	0.386	0.136	0.093	0.324	0.000	0.000	0.115	0.709	0.934	0.078	0.075	0.075
DA	.00	-0.001	-0.001	0.000	0.722	0.127	0.062	1.000	1.000	1.000	0.000	0.000	0.000	0.015	0.000	0.000
DA	.14	0.307	-0.140	-0.092	0.785	0.173	0.103	1.000	1.000	0.968	0.000	0.000	0.007	0.025	0.001	0.001
DA	.39	0.008	-0.105	-0.122	1.139	0.375	0.253	0.995	0.925	0.922	0.000	0.171	0.556	0.073	0.009	0.004
DA	.59	-0.025	-0.099	-0.124	1.602	0.580	0.394	0.950	0.916	0.911	0.004	0.593	0.935	0.140	0.021	0.011
MICE	.00	-0.002	-0.002	0.000	0.776	0.136	0.067	1.000	0.999	1.000	0.000	0.001	0.000	0.021	0.001	0.000
MICE	.14	0.531	-0.050	0.025	0.853	0.185	0.110	1.000	1.000	0.972	0.000	0.003	0.008	0.033	0.002	0.001
MICE	.39	0.144	0.020	-0.002	1.218	0.405	0.272	0.999	0.937	0.939	0.002	0.218	0.632	0.094	0.012	0.005
MICE	.59	0.114	0.031	0.002	1.738	0.627	0.426	0.959	0.931	0.949	0.009	0.652	0.950	0.187	0.026	0.012
FCS-BB	.00	-0.006	-0.007	0.000	2.634	1.078	0.892	0.999	1.000	1.000	0.001	0.000	0.000	0.100	0.011	0.005
FCS-BB	.14	2.561	3.087	4.495	2.765	1.191	1.010	0.995	0.996	0.998	0.005	0.006	0.002	0.174	0.031	0.027
FCS-BB	.39	1.527	2.285	2.580	3.436	1.531	1.237	0.976	0.919	0.853	0.033	0.199	0.359	0.493	0.206	0.194
FCS-BB	.59	1.344	1.630	1.649	4.311	1.839	1.466	0.972	0.824	0.703	0.083	0.547	0.800	1.003	0.443	0.395
FCS-BB*	.00	0.001	0.000	0.000	0.924	0.196	0.123	1.000	1.000	1.000	0.000	0.000	0.000	0.012	0.000	0.000
FCS-BB*	.14	-0.052	-0.045	-0.024	0.981	0.242	0.166	1.000	1.000	0.997	0.000	0.001	0.007	0.012	0.001	0.000
FCS-BB*	.39	-0.037	-0.035	-0.028	1.381	0.468	0.363	0.992	0.979	0.995	0.026	0.256	0.511	0.064	0.008	0.004
FCS-BB*	.59	0.042	-0.038	-0.018	1.959	0.723	0.565	0.956	0.985	0.995	0.143	0.731	0.924	0.195	0.020	0.010

Note. CC = complete case analysis; Mean = mean imputation; MBE = model-based estimation; DA = data augmentation; MICE = multiple imputation by chained equations; FCS-BB = fully conditional specification using Bayesian bootstrap; FCS-BB* = modified fully conditional specification using Bayesian bootstrap.

similar MSEs. With regards to FMI, FCS-BB tends to overestimate the FMI at smaller effect sizes; other methods have comparable estimates and overestimate the amount of FMI in conditions (see Table 4.21). Supplemental figures for the results are presented in Appendix D.

Table 4.21
Mediation Model FMI Metric – Mediator Categorical, Endogenous Categorical

		FMI – 10% Missingness			*FMI – 20% Missingness*		
Method	*Effect*	*100*	*500*	*1000*	*100*	*500*	*1000*
DA	.00	0.137	0.131	0.132	0.245	0.236	0.236
DA	.14	0.142	0.136	0.132	0.251	0.245	0.251
DA	.39	0.149	0.141	0.139	0.260	0.262	0.261
DA	.59	0.156	0.148	0.143	0.279	0.270	0.268
MICE	.00	0.143	0.139	0.137	0.264	0.254	0.253
MICE	.14	0.147	0.141	0.139	0.270	0.267	0.271
MICE	.39	0.153	0.143	0.145	0.281	0.285	0.282
MICE	.59	0.157	0.143	0.146	0.293	0.292	0.292
FCS-BB	.00	0.226	0.310	0.362	0.376	0.540	0.598
FCS-BB	.14	0.212	0.302	0.318	0.393	0.504	0.540
FCS-BB	.39	0.190	0.188	0.174	0.342	0.353	0.374
FCS-BB	.59	0.157	0.134	0.146	0.294	0.288	0.305
FCS-BB*	.00	0.162	0.149	0.138	0.292	0.262	0.259
FCS-BB*	.14	0.158	0.147	0.136	0.290	0.263	0.259
FCS-BB*	.39	0.158	0.140	0.140	0.291	0.268	0.261
FCS-BB*	.59	0.158	0.146	0.143	0.295	0.269	0.265

Note. DA = data augmentation; MICE = multiple imputation by chained equations; FCS-BB = fully conditional specification using Bayesian bootstrap; FCS-BB* = modified fully conditional specification using Bayesian bootstrap.

4.2.3. Results: Moderated mediation models. Results of the MC simulation for a moderated mediation model with continuous mediator variable and continuous endogenous variable and missingness rates of 10% and 20% are displayed in Table 4.22 and Table 4.23, respectively. As can be seen in the tables, only CC, MBE, and FCS-BB-JAV methods are relatively unbiased across all conditions, with lower biases for CC and MBE. Among other methods, mean imputation has the largest biases, whereas, FCS-BB-PI has the

lowest biases. Confidence interval lengths are narrowest for mean imputation and tend to be highest for FCS-BB methods (i.e., both PI and JAV); other methods have comparable confidence interval lengths. Similar to results on biases, for coverage probabilities only CC, MBE, and FCS-BB-JAV methods have approximately nominal or higher coverage probabilities in all conditions; MBE has slightly better coverage performance. Mean imputation has the lowest coverage probabilities across all conditions; for other methods, both DA and MICE have similar coverage probabilities, with below nominal rates at larger effect sizes. FCS-BB-PI also has lower than nominal coverage probabilities at medium and large effect sizes, however, this trend only occurs with 20% missingness and its coverage probabilities are higher than DA and MICE. All methods have below nominal Type I Error rates. In small samples at small effect sizes, MBE and the FCS methods have higher power than other methods. In medium sized samples with 20% missingness, MBE and mean imputation have slightly higher power relative to other methods, whereas, DA and MICE have slightly lower power relative to other methods; methods perform similarly across other conditions. MSEs are highest for mean imputation and generally lowest for MBE, CC, and FCS-BB-JAV methods. DA, MICE, and FCS-BB-PI have similar MSEs, which are lower than mean imputation, but larger than MBE, CC, and FCS-BB-JAV methods (especially with 20% missingness). With regards to MI methods, FMI is most consistently estimated by FCS-BB-JAV; other methods perform similarly and overestimate the true amount of missing information (see Table 4.24). Supplemental figures for the results are presented in Appendix D.

Results of the MC simulation for a moderated mediation model with continuous mediator variable and categorical endogenous variable and missingness rates of 10% and

Table 4.22
Moderated Mediation Model Metrics – Mediator Continuous, Endogenous Continuous, 10% Missingness Per Variable

		Empirical Bias			*Confidence Interval Length*			*Coverage Probability*			*Rejection Rate*			*Mean Squared Error*		
Method	*Effect*	*100*	*500*	*1000*	*100*	*500*	*1000*	*100*	*500*	*1000*	*100*	*500*	*1000*	*100*	*500*	*1000*
CC	.00	0.001	0.000	0.000	0.128	0.023	0.012	1.000	1.000	1.000	0.000	0.000	0.000	0.001	0.000	0.000
CC	.14	0.009	-0.007	0.006	0.276	0.106	0.075	0.917	0.938	0.933	0.055	0.921	1.000	0.005	0.001	0.000
CC	.39	-0.007	0.007	0.002	0.677	0.285	0.199	0.933	0.942	0.957	0.979	1.000	1.000	0.031	0.005	0.002
CC	.59	0.004	-0.003	-0.001	0.993	0.416	0.292	0.942	0.946	0.953	1.000	1.000	1.000	0.067	0.012	0.005
Mean	.00	0.001	0.000	0.000	0.104	0.019	0.010	1.000	1.000	1.000	0.000	0.000	0.000	0.001	0.000	0.000
Mean	.14	-0.134	-0.137	-0.125	0.227	0.090	0.063	0.877	0.870	0.853	0.067	0.927	1.000	0.003	0.001	0.000
Mean	.39	-0.171	-0.144	-0.146	0.554	0.240	0.168	0.792	0.660	0.456	0.976	1.000	1.000	0.036	0.012	0.010
Mean	.59	-0.177	-0.161	-0.159	0.827	0.357	0.251	0.721	0.367	0.125	0.999	1.000	1.000	0.130	0.064	0.056
MBE	.00	0.001	0.000	0.000	0.196	0.034	0.017	0.995	0.991	0.997	0.005	0.009	0.003	0.001	0.000	0.000
MBE	.14	0.013	-0.008	0.007	0.321	0.110	0.076	0.948	0.958	0.945	0.235	0.963	1.000	0.005	0.001	0.000
MBE	.39	-0.008	0.007	0.002	0.704	0.283	0.197	0.949	0.939	0.953	0.988	1.000	1.000	0.030	0.005	0.002
MBE	.59	0.006	-0.003	-0.001	1.021	0.415	0.291	0.947	0.940	0.950	1.000	1.000	1.000	0.065	0.012	0.005
DA	.00	0.001	0.000	0.000	0.125	0.023	0.012	1.000	1.000	1.000	0.000	0.000	0.000	0.001	0.000	0.000
DA	.14	-0.128	-0.114	-0.102	0.254	0.100	0.070	0.908	0.893	0.889	0.032	0.885	0.997	0.003	0.001	0.000
DA	.39	-0.131	-0.091	-0.094	0.645	0.278	0.195	0.875	0.835	0.750	0.960	1.000	1.000	0.035	0.009	0.006
DA	.59	-0.099	-0.083	-0.079	1.016	0.437	0.307	0.840	0.758	0.666	1.000	1.000	1.000	0.103	0.030	0.021
MICE	.00	0.001	0.000	0.000	0.124	0.023	0.012	1.000	1.000	1.000	0.000	0.000	0.000	0.001	0.000	0.000
MICE	.14	-0.134	-0.114	-0.104	0.255	0.100	0.070	0.908	0.900	0.889	0.027	0.886	0.998	0.003	0.001	0.000
MICE	.39	-0.131	-0.092	-0.095	0.647	0.278	0.195	0.872	0.819	0.748	0.964	1.000	1.000	0.035	0.009	0.006
MICE	.59	-0.100	-0.084	-0.079	1.017	0.436	0.307	0.848	0.748	0.669	1.000	1.000	1.000	0.104	0.030	0.020
FCS-BB-PI	.00	0.001	0.000	0.000	0.142	0.038	0.024	0.996	0.999	1.000	0.004	0.001	0.000	0.000	0.000	0.000
FCS-BB-PI	.14	-0.108	-0.098	-0.089	0.244	0.116	0.096	0.906	0.959	0.982	0.183	0.913	0.997	0.004	0.001	0.000
FCS-BB-PI	.39	-0.102	-0.084	-0.086	0.646	0.351	0.288	0.876	0.922	0.922	0.984	1.000	1.000	0.038	0.009	0.007
FCS-BB-PI	.59	-0.095	-0.076	-0.074	1.121	0.613	0.506	0.887	0.927	0.930	1.000	1.000	1.000	0.111	0.031	0.022
FCS-BB-JAV	.00	0.001	0.000	0.000	0.167	0.043	0.027	0.999	0.999	1.000	0.001	0.001	0.000	0.000	0.000	0.000
FCS-BB-JAV	.14	-0.028	-0.024	-0.022	0.281	0.130	0.106	0.927	0.981	0.992	0.183	0.916	0.998	0.004	0.001	0.000
FCS-BB-JAV	.39	-0.027	-0.008	-0.010	0.712	0.371	0.302	0.922	0.985	0.987	0.985	1.000	1.000	0.035	0.006	0.003
FCS-BB-JAV	.59	-0.018	-0.005	-0.004	1.162	0.612	0.498	0.935	0.987	0.995	1.000	1.000	1.000	0.092	0.016	0.008

Note. CC = complete case analysis; Mean = mean imputation; MBE = model-based estimation; DA = data augmentation; MICE = multiple imputation by chained equations; FCS-BB-PI = fully conditional specification using Bayesian bootstrapping and passive imputation; FCS-BB-JAV = fully conditional specification using Bayesian bootstrapping and just another variable imputation.

Table 4.23
Moderated Mediation Model Metrics – Mediator Continuous, Endogenous Continuous, 20% Missingness Per Variable

		Empirical Bias			*Confidence Interval Length*			*Coverage Probability*			*Rejection Rate*			*Mean Squared Error*		
Method	*Effect*	*100*	*500*	*1000*	*100*	*500*	*1000*	*100*	*500*	*1000*	*100*	*500*	*1000*	*100*	*500*	*1000*
CC	.00	0.000	0.000	0.000	0.157	0.027	0.013	1.000	1.000	1.000	0.000	0.000	0.000	0.001	0.000	0.000
CC	.14	-0.011	-0.009	-0.012	0.303	0.116	0.081	0.906	0.928	0.942	0.036	0.845	0.998	0.006	0.001	0.000
CC	.39	0.012	0.003	0.002	0.749	0.307	0.215	0.941	0.942	0.947	0.960	1.000	1.000	0.039	0.006	0.003
CC	.59	0.004	0.000	0.002	1.098	0.452	0.316	0.935	0.948	0.954	0.999	1.000	1.000	0.086	0.013	0.006
Mean	.00	0.000	0.000	0.000	0.102	0.018	0.009	1.000	1.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000
Mean	.14	-0.290	-0.270	-0.268	0.201	0.081	0.057	0.802	0.735	0.631	0.055	0.885	0.999	0.003	0.001	0.001
Mean	.39	-0.293	-0.284	-0.281	0.500	0.214	0.151	0.615	0.167	0.036	0.960	1.000	1.000	0.055	0.034	0.031
Mean	.59	-0.325	-0.306	-0.302	0.754	0.325	0.229	0.400	0.018	0.000	0.995	1.000	1.000	0.279	0.197	0.184
MBE	.00	0.000	0.000	0.000	0.237	0.039	0.020	0.995	0.993	0.996	0.005	0.006	0.004	0.001	0.000	0.000
MBE	.14	-0.006	-0.008	-0.012	0.357	0.118	0.081	0.938	0.961	0.954	0.183	0.962	0.998	0.005	0.001	0.000
MBE	.39	0.016	0.002	0.001	0.776	0.301	0.211	0.951	0.944	0.942	0.981	1.000	1.000	0.037	0.006	0.003
MBE	.59	0.007	0.001	0.003	1.115	0.449	0.312	0.940	0.947	0.955	1.000	1.000	1.000	0.082	0.013	0.006
DA	.00	0.000	0.000	0.000	0.147	0.026	0.013	1.000	1.000	1.000	0.000	0.000	0.000	0.001	0.000	0.000
DA	.14	-0.251	-0.223	-0.223	0.262	0.101	0.070	0.892	0.850	0.804	0.006	0.751	0.993	0.004	0.001	0.001
DA	.39	-0.192	-0.184	-0.180	0.679	0.285	0.200	0.828	0.649	0.437	0.891	1.000	1.000	0.045	0.018	0.015
DA	.59	-0.184	-0.160	-0.155	1.088	0.465	0.328	0.757	0.527	0.310	0.997	1.000	1.000	0.172	0.068	0.056
MICE	.00	0.000	0.000	0.000	0.149	0.026	0.013	1.000	1.000	1.000	0.000	0.000	0.000	0.001	0.000	0.000
MICE	.14	-0.248	-0.223	-0.223	0.263	0.101	0.070	0.896	0.855	0.797	0.016	0.742	0.995	0.004	0.001	0.001
MICE	.39	-0.196	-0.184	-0.180	0.684	0.285	0.201	0.825	0.640	0.435	0.880	1.000	1.000	0.046	0.018	0.015
MICE	.59	-0.186	-0.160	-0.155	1.094	0.466	0.329	0.763	0.540	0.307	0.997	1.000	1.000	0.171	0.068	0.055
FCS-BB-PI	.00	-0.001	0.000	0.000	0.146	0.038	0.024	0.999	1.000	1.000	0.001	0.000	0.000	0.000	0.000	0.000
FCS-BB-PI	.14	-0.151	-0.144	-0.142	0.236	0.112	0.092	0.896	0.912	0.945	0.137	0.813	0.986	0.003	0.001	0.001
FCS-BB-PI	.39	-0.147	-0.139	-0.128	0.628	0.351	0.291	0.815	0.772	0.726	0.974	1.000	1.000	0.049	0.020	0.016
FCS-BB-PI	.59	-0.125	-0.124	-0.120	1.117	0.632	0.531	0.792	0.701	0.674	1.000	1.000	1.000	0.175	0.079	0.061
FCS-BB-JAV	.00	-0.001	0.000	0.000	0.199	0.049	0.031	0.997	1.000	1.000	0.003	0.000	0.000	0.001	0.000	0.000
FCS-BB-JAV	.14	-0.057	-0.054	-0.033	0.314	0.140	0.114	0.929	0.971	0.994	0.145	0.822	0.984	0.005	0.001	0.000
FCS-BB-JAV	.39	-0.036	-0.010	-0.005	0.759	0.393	0.321	0.940	0.979	0.991	0.977	1.000	1.000	0.037	0.007	0.004
FCS-BB-JAV	.59	-0.022	-0.009	-0.003	1.213	0.637	0.522	0.933	0.975	0.993	1.000	1.000	1.000	0.102	0.018	0.009

Note. CC = complete case analysis; Mean = mean imputation; MBE = model-based estimation; DA = data augmentation; MICE = multiple imputation by chained equations; FCS-BB-PI = fully conditional specification using Bayesian bootstrapping and passive imputation; FCS-BB-JAV = fully conditional specification using Bayesian bootstrapping and just another variable imputation.

Table 4.24
Moderated Mediation Model FMI Metric – Mediator Continuous, Endogenous Continuous

		FMI – 10% Missingness			*FMI – 20% Missingness*		
Method	*Effect*	*100*	*500*	*1000*	*100*	*500*	*1000*
DA	.00	0.103	0.096	0.096	0.181	0.175	0.177
DA	.14	0.119	0.113	0.118	0.216	0.213	0.218
DA	.39	0.156	0.154	0.153	0.262	0.261	0.257
DA	.59	0.200	0.208	0.209	0.308	0.314	0.319
MICE	.00	0.101	0.096	0.093	0.189	0.176	0.173
MICE	.14	0.120	0.115	0.116	0.217	0.218	0.217
MICE	.39	0.160	0.154	0.155	0.269	0.262	0.262
MICE	.59	0.200	0.205	0.208	0.313	0.314	0.321
FCS-BB-PI	.00	0.111	0.098	0.096	0.189	0.182	0.173
FCS-BB-PI	.14	0.110	0.115	0.108	0.200	0.201	0.199
FCS-BB-PI	.39	0.157	0.155	0.157	0.257	0.254	0.257
FCS-BB-PI	.59	0.233	0.245	0.251	0.319	0.342	0.349
FCS-BB-JAV	.00	0.095	0.098	0.098	0.179	0.189	0.188
FCS-BB-JAV	.14	0.096	0.101	0.103	0.186	0.196	0.197
FCS-BB-JAV	.39	0.097	0.092	0.095	0.182	0.181	0.183
FCS-BB-JAV	.59	0.088	0.086	0.089	0.171	0.177	0.176

Note. DA = data augmentation; MICE = multiple imputation by chained equations; FCS-BB-PI = fully conditional specification using Bayesian bootstrapping and passive imputation; FCS-BB-JAV = fully conditional specification using Bayesian bootstrapping and just another variable imputation.

20% are displayed in Table 4.25 and Table 4.26, respectively. In small sample sizes, FCS-BB-JAV is the least biased method; CC and MBE are both highly biased. In small samples, DA, MICE, and FCS-BB-PI perform similarly such that these methods are slightly biased in conditions with 10% missingness; however, in 20% missingness, FCS-BB-PI tends to be less biased than DA and MICE. Across other conditions, only CC and FCS-BB-JAV are relatively unbiased, with less bias for CC; other methods underestimate the true indirect effect, with the most biased results for MBE. For confidence interval lengths, in small sample sizes confidence interval lengths are widest for FCS-BB-JAV and CC methods and lowest for MBE; across other conditions MBE has slightly narrower intervals and the FCS

Table 4.25
Moderated Mediation Model Metrics – Mediator Continuous, Endogenous Categorical, 10% Missingness Per Variable

		Empirical Bias			*Confidence Interval Length*			*Coverage Probability*			*Rejection Rate*			*Mean Squared Error*		
Method	*Effect*	*100*	*500*	*1000*	*100*	*500*	*1000*	*100*	*500*	*1000*	*100*	*500*	*1000*	*100*	*500*	*1000*
CC	.00	-0.002	0.000	0.000	0.296	0.048	0.023	1.000	1.000	0.999	0.000	0.000	0.001	0.005	0.000	0.000
CC	.14	0.078	0.021	0.003	0.500	0.179	0.122	0.945	0.932	0.938	0.005	0.358	0.755	0.018	0.002	0.001
CC	.39	0.196	0.041	0.014	1.431	0.550	0.382	0.951	0.941	0.950	0.472	1.000	1.000	0.187	0.022	0.010
CC	.59	0.172	0.022	0.021	2.530	0.963	0.676	0.953	0.940	0.950	0.850	1.000	1.000	0.543	0.066	0.033
Mean	.00	-0.001	0.000	0.000	0.251	0.042	0.021	1.000	1.000	0.999	0.000	0.000	0.001	0.003	0.000	0.000
Mean	.14	-0.076	-0.108	-0.128	0.427	0.157	0.108	0.924	0.909	0.891	0.007	0.354	0.749	0.014	0.002	0.001
Mean	.39	-0.025	-0.144	-0.167	1.185	0.469	0.326	0.899	0.819	0.733	0.460	1.000	1.000	0.115	0.024	0.017
Mean	.59	-0.067	-0.174	-0.168	2.041	0.804	0.568	0.878	0.719	0.599	0.824	1.000	1.000	0.351	0.107	0.079
MBE	.00	-0.002	0.000	0.000	0.449	0.051	0.024	0.995	0.998	0.993	0.005	0.002	0.007	0.002	0.000	0.000
MBE	.14	-0.414	-0.386	-0.387	0.542	0.124	0.080	0.909	0.853	0.711	0.037	0.488	0.798	0.008	0.002	0.001
MBE	.39	-0.378	-0.422	-0.431	0.959	0.298	0.203	0.838	0.134	0.004	0.454	0.999	1.000	0.093	0.072	0.071
MBE	.59	-0.438	-0.481	-0.481	1.379	0.432	0.297	0.543	0.001	0.000	0.718	1.000	1.000	0.462	0.461	0.455
DA	.00	-0.001	0.000	0.000	0.283	0.048	0.023	1.000	1.000	0.999	0.000	0.000	0.001	0.003	0.000	0.000
DA	.14	-0.083	-0.105	-0.122	0.465	0.168	0.114	0.956	0.924	0.912	0.001	0.275	0.702	0.012	0.002	0.001
DA	.39	-0.021	-0.113	-0.135	1.308	0.512	0.355	0.937	0.882	0.840	0.321	1.000	1.000	0.108	0.021	0.013
DA	.59	-0.093	-0.176	-0.173	2.275	0.881	0.623	0.914	0.756	0.653	0.683	1.000	1.000	0.293	0.103	0.080
MICE	.00	-0.002	0.000	0.000	0.285	0.048	0.023	1.000	1.000	1.000	0.000	0.000	0.000	0.003	0.000	0.000
MICE	.14	-0.082	-0.095	-0.113	0.466	0.168	0.115	0.955	0.921	0.912	0.001	0.298	0.709	0.012	0.002	0.001
MICE	.39	-0.014	-0.101	-0.122	1.306	0.513	0.357	0.928	0.888	0.855	0.333	0.999	1.000	0.112	0.021	0.012
MICE	.59	-0.066	-0.156	-0.151	2.295	0.890	0.629	0.923	0.801	0.712	0.709	1.000	1.000	0.314	0.092	0.067
FCS-BB-PI	.00	-0.001	0.000	0.000	0.361	0.080	0.049	0.996	1.000	1.000	0.004	0.000	0.000	0.002	0.000	0.000
FCS-BB-PI	.14	-0.076	-0.072	-0.068	0.524	0.204	0.159	0.969	0.979	0.990	0.058	0.320	0.537	0.011	0.002	0.001
FCS-BB-PI	.39	-0.041	-0.064	-0.055	1.312	0.596	0.477	0.938	0.956	0.964	0.693	1.000	1.000	0.116	0.020	0.012
FCS-BB-PI	.59	-0.043	-0.051	-0.049	2.162	0.988	0.807	0.927	0.890	0.878	0.930	1.000	1.000	0.303	0.096	0.074
FCS-BB-JAV	.00	-0.002	0.000	0.000	0.439	0.091	0.055	0.996	1.000	1.000	0.004	0.000	0.000	0.003	0.000	0.000
FCS-BB-JAV	.14	0.014	-0.015	-0.029	0.634	0.232	0.179	0.974	0.980	0.989	0.056	0.331	0.542	0.016	0.002	0.001
FCS-BB-JAV	.39	0.045	0.035	0.001	1.594	0.683	0.544	0.929	0.978	0.995	0.690	1.000	1.000	0.193	0.023	0.009
FCS-BB-JAV	.59	0.039	-0.027	-0.035	2.695	1.151	0.929	0.920	0.984	0.992	0.915	1.000	1.000	0.521	0.062	0.033

Note. CC = complete case analysis; Mean = mean imputation; MBE = model-based estimation; DA = data augmentation; MICE = multiple imputation by chained equations; FCS-BB-PI = fully conditional specification using Bayesian bootstrapping and passive imputation; FCS-BB-JAV = fully conditional specification using Bayesian bootstrapping and just another variable imputation.

Table 4.26
Moderated Mediation Model Metrics – Mediator Continuous, Endogenous Categorical, 20% Missingness Per Variable

		Empirical Bias			*Confidence Interval Length*			*Coverage Probability*			*Rejection Rate*			*Mean Squared Error*		
Method	*Effect*	*100*	*500*	*1000*	*100*	*500*	*1000*	*100*	*500*	*1000*	*100*	*500*	*1000*	*100*	*500*	*1000*
CC	.00	0.003	0.001	0.000	0.340	0.055	0.027	1.000	1.000	1.000	0.000	0.000	0.000	0.005	0.000	0.000
CC	.14	0.152	0.020	0.035	0.581	0.195	0.134	0.960	0.941	0.944	0.004	0.292	0.701	0.024	0.003	0.001
CC	.39	0.213	0.030	0.016	1.598	0.594	0.414	0.961	0.958	0.955	0.377	0.999	1.000	0.215	0.023	0.011
CC	.59	0.275	0.035	0.014	2.921	1.051	0.728	0.967	0.958	0.966	0.779	1.000	1.000	0.864	0.075	0.032
Mean	.00	0.002	0.000	0.000	0.245	0.042	0.021	1.000	1.000	1.000	0.000	0.000	0.000	0.002	0.000	0.000
Mean	.14	-0.156	-0.237	-0.222	0.422	0.149	0.103	0.905	0.836	0.807	0.005	0.288	0.649	0.013	0.002	0.001
Mean	.39	-0.169	-0.297	-0.309	1.120	0.432	0.302	0.841	0.570	0.319	0.354	0.998	1.000	0.116	0.045	0.042
Mean	.59	-0.169	-0.304	-0.323	1.933	0.746	0.516	0.803	0.412	0.126	0.734	1.000	1.000	0.397	0.223	0.222
MBE	.00	0.003	0.000	0.000	0.614	0.059	0.027	0.993	0.994	0.994	0.007	0.006	0.006	0.002	0.000	0.000
MBE	.14	-0.357	-0.382	-0.374	0.640	0.136	0.088	0.929	0.869	0.763	0.026	0.447	0.759	0.010	0.002	0.001
MBE	.39	-0.364	-0.428	-0.430	1.155	0.323	0.220	0.893	0.157	0.016	0.361	0.996	1.000	0.098	0.074	0.071
MBE	.59	-0.402	-0.477	-0.484	1.619	0.469	0.321	0.668	0.000	0.000	0.619	1.000	1.000	0.435	0.455	0.460
DA	.00	0.003	0.000	0.000	0.308	0.053	0.026	1.000	1.000	1.000	0.000	0.000	0.000	0.002	0.000	0.000
DA	.14	-0.186	-0.228	-0.218	0.495	0.169	0.115	0.981	0.901	0.884	0.000	0.154	0.563	0.011	0.002	0.001
DA	.39	-0.173	-0.258	-0.263	1.298	0.498	0.350	0.914	0.749	0.556	0.134	0.996	1.000	0.090	0.036	0.031
DA	.59	-0.211	-0.313	-0.324	2.257	0.848	0.591	0.885	0.470	0.139	0.436	1.000	1.000	0.293	0.221	0.217
MICE	.00	0.003	0.000	0.000	0.313	0.054	0.026	1.000	1.000	1.000	0.000	0.000	0.000	0.002	0.000	0.000
MICE	.14	-0.165	-0.208	-0.196	0.505	0.171	0.117	0.986	0.903	0.892	0.001	0.176	0.577	0.012	0.002	0.001
MICE	.39	-0.148	-0.233	-0.238	1.324	0.506	0.353	0.919	0.781	0.629	0.146	0.998	1.000	0.092	0.033	0.027
MICE	.59	-0.170	-0.277	-0.289	2.328	0.871	0.604	0.902	0.566	0.256	0.474	1.000	1.000	0.298	0.184	0.177
FCS-BB-PI	.00	-0.001	0.000	0.000	0.355	0.080	0.050	1.000	1.000	1.000	0.000	0.000	0.000	0.001	0.000	0.000
FCS-BB-PI	.14	-0.116	-0.101	-0.087	0.515	0.196	0.153	0.968	0.969	0.977	0.045	0.271	0.468	0.011	0.002	0.001
FCS-BB-PI	.39	-0.099	-0.091	-0.081	1.208	0.543	0.440	0.932	0.866	0.831	0.602	0.997	1.000	0.095	0.034	0.029
FCS-BB-PI	.59	-0.084	-0.071	-0.062	1.928	0.878	0.710	0.905	0.620	0.355	0.901	1.000	1.000	0.298	0.209	0.205
FCS-BB-JAV	.00	0.002	0.000	0.000	0.510	0.103	0.063	1.000	1.000	1.000	0.000	0.000	0.000	0.003	0.000	0.000
FCS-BB-JAV	.14	0.075	-0.026	-0.005	0.734	0.253	0.196	0.971	0.985	0.996	0.055	0.282	0.476	0.021	0.002	0.001
FCS-BB-JAV	.39	0.051	0.044	-0.021	1.741	0.710	0.568	0.947	0.984	0.993	0.606	0.997	1.000	0.201	0.022	0.010
FCS-BB-JAV	.59	0.049	-0.041	-0.011	2.849	1.154	0.920	0.931	0.971	0.977	0.882	1.000	1.000	0.568	0.068	0.044

Note. CC = complete case analysis; Mean = mean imputation; MBE = model-based estimation; DA = data augmentation; MICE = multiple imputation by chained equations; FCS-BB-PI = fully conditional specification using Bayesian bootstrapping and passive imputation; FCS-BB-JAV = fully conditional specification using Bayesian bootstrapping and just another variable imputation.

methods have slightly wider intervals compared to other methods. In terms of coverage probabilities, only the FCS-BB-JAV method has approximately nominal or higher coverage across all conditions. Moreover, across all conditions, MBE and mean imputation have the lowest coverage rates, with lower coverage for MBE; for other methods, CC has the highest coverage probabilities, whereas, DA has the lowest coverage probabilities.

All methods have lower than nominal Type I Error rates. In small samples, FCS methods have higher power across all conditions, whereas, DA and MICE have the lowest power. In medium sample sizes, MBE has the highest power, whereas, DA and MICE have the lowest power. In large sample sizes, with small effect sizes, FCS methods have lower power and MBE has slightly higher power than other methods; other methods perform similar with 10% missingness, but with 20% missingness, DA and MICE have slightly lower power than CC and mean imputation. For MSEs, estimates are highest for CC in small samples, whereas, highest for MBE in medium and large sized samples. Relative to other methods, DA, MICE, and FCS-BB-PI generally have the lower MSEs for small samples, whereas, FCS-BB-JAV and CC generally have lower MSEs for medium and large sized samples. With regards to MI methods and FMI estimation, all methods have comparable performance with null and small effect sizes; for larger effect sizes, all methods tend to overestimate the FMI (see Table 4.27). Supplemental figures for the results are presented in Appendix D.

Results of the MC simulation for a moderated mediation model with categorical mediator variable and continuous endogenous variable and missingness rates of 10% and 20% are displayed in Table 4.28 and Table 4.29, respectively. In small sample sizes, FCS-BB-JAV

Table 4.27
Moderated Mediation Model FMI Metric – Mediator Continuous, Endogenous Categorical

		FMI – 10% Missingness			*FMI – 20% Missingness*		
Method	*Effect*	*100*	*500*	*1000*	*100*	*500*	*1000*
DA	.00	0.108	0.094	0.095	0.178	0.171	0.173
DA	.14	0.118	0.121	0.119	0.211	0.223	0.223
DA	.39	0.175	0.168	0.167	0.278	0.280	0.283
DA	.59	0.231	0.228	0.235	0.347	0.344	0.351
MICE	.00	0.107	0.094	0.096	0.181	0.177	0.175
MICE	.14	0.125	0.119	0.121	0.220	0.227	0.228
MICE	.39	0.168	0.162	0.165	0.284	0.284	0.281
MICE	.59	0.222	0.224	0.229	0.347	0.345	0.346
FCS-BB-PI	.00	0.118	0.100	0.100	0.204	0.190	0.184
FCS-BB-PI	.14	0.126	0.120	0.122	0.224	0.219	0.221
FCS-BB-PI	.39	0.200	0.171	0.177	0.301	0.293	0.295
FCS-BB-PI	.59	0.262	0.247	0.249	0.374	0.371	0.369
FCS-BB-JAV	.00	0.110	0.103	0.105	0.205	0.194	0.195
FCS-BB-JAV	.14	0.123	0.120	0.122	0.230	0.231	0.239
FCS-BB-JAV	.39	0.174	0.167	0.168	0.299	0.297	0.300
FCS-BB-JAV	.59	0.235	0.231	0.222	0.380	0.377	0.379

Note. DA = data augmentation; MICE = multiple imputation by chained equations; FCS-BB-PI = fully conditional specification using Bayesian bootstrapping and passive imputation; FCS-BB-JAV = fully conditional specification using Bayesian bootstrapping and just another variable imputation.

is the least biased method; in medium to large sample sizes, both CC and FCS-BB-JAV are the only methods that are relatively unbiased, with less bias for CC; other methods generally demonstrate stronger bias in all conditions, with highest biases for MBE and lowest biases for MICE. In small sample sizes, confidence interval lengths are widest for FCS-BB-JAV and CC methods and lowest for mean imputation and MBE. Across other conditions, MBE has slightly narrower intervals and the FCS methods have slightly wider intervals compared to other methods; DA, MICE, and CC have comparable confidence interval lengths. With regards to coverage probabilities, only the FCS-BB-JAV method has approximately nominal or higher coverage rates across all conditions. Both mean

Table 4.28
Moderated Mediation Model Metrics – Mediator Categorical, Endogenous Continuous, 10% Missingness Per Variable

		Empirical Bias			*Confidence Interval Length*			*Coverage Probability*			*Rejection Rate*			*Mean Squared Error*		
Method	*Effect*	*100*	*500*	*1000*	*100*	*500*	*1000*	*100*	*500*	*1000*	*100*	*500*	*1000*	*100*	*500*	*1000*
CC	.00	-0.002	0.000	0.000	0.546	0.092	0.047	1.000	1.000	1.000	0.000	0.000	0.000	0.015	0.000	0.000
CC	.14	0.147	0.010	0.030	0.780	0.235	0.160	0.971	0.905	0.933	0.001	0.077	0.436	0.039	0.003	0.002
CC	.39	0.103	0.018	0.026	1.835	0.698	0.490	0.930	0.939	0.951	0.099	0.999	1.000	0.237	0.035	0.017
CC	.59	0.127	0.027	0.013	3.039	1.182	0.823	0.936	0.954	0.962	0.505	1.000	1.000	0.819	0.096	0.042
Mean	.00	-0.001	0.000	0.000	0.440	0.077	0.039	1.000	1.000	1.000	0.000	0.000	0.000	0.010	0.000	0.000
Mean	.14	-0.095	-0.196	-0.185	0.620	0.192	0.130	0.946	0.861	0.854	0.001	0.076	0.403	0.025	0.003	0.001
Mean	.39	-0.172	-0.215	-0.216	1.399	0.548	0.383	0.851	0.748	0.662	0.091	0.995	1.000	0.145	0.039	0.028
Mean	.59	-0.218	-0.252	-0.267	2.179	0.882	0.611	0.785	0.590	0.323	0.439	1.000	1.000	0.460	0.179	0.165
MBE	.00	-0.001	0.000	0.000	0.542	0.076	0.037	0.993	0.994	0.995	0.007	0.006	0.005	0.004	0.000	0.000
MBE	.14	-0.417	-0.494	-0.483	0.693	0.141	0.090	0.896	0.835	0.708	0.026	0.336	0.754	0.012	0.002	0.002
MBE	.39	-0.452	-0.491	-0.488	1.285	0.381	0.260	0.896	0.286	0.040	0.362	0.999	1.000	0.139	0.098	0.092
MBE	.59	-0.459	-0.496	-0.502	1.970	0.617	0.423	0.760	0.068	0.002	0.743	1.000	1.000	0.602	0.502	0.499
DA	.00	0.000	0.000	0.000	0.531	0.091	0.046	1.000	1.000	1.000	0.000	0.000	0.000	0.010	0.000	0.000
DA	.14	-0.045	-0.134	-0.117	0.724	0.217	0.148	0.992	0.899	0.909	0.000	0.046	0.344	0.028	0.003	0.001
DA	.39	-0.109	-0.141	-0.133	1.630	0.635	0.447	0.905	0.875	0.847	0.039	0.995	1.000	0.138	0.032	0.019
DA	.59	-0.100	-0.149	-0.161	2.726	1.072	0.747	0.892	0.851	0.764	0.294	1.000	1.000	0.454	0.109	0.079
MICE	.00	0.000	0.000	0.000	0.534	0.092	0.046	1.000	1.000	1.000	0.000	0.000	0.000	0.010	0.000	0.000
MICE	.14	-0.020	-0.117	-0.087	0.736	0.219	0.150	0.995	0.899	0.909	0.000	0.045	0.370	0.030	0.003	0.001
MICE	.39	-0.087	-0.114	-0.109	1.651	0.644	0.453	0.905	0.895	0.870	0.043	0.996	1.000	0.144	0.031	0.017
MICE	.59	-0.060	-0.116	-0.126	2.780	1.091	0.759	0.900	0.884	0.838	0.340	1.000	1.000	0.478	0.098	0.062
FCS-BB-PI	.00	0.000	-0.000	0.000	0.688	0.155	0.097	0.998	1.000	1.000	0.002	0.000	0.000	0.007	0.000	0.000
FCS-BB-PI	.14	-0.091	-0.084	-0.073	0.862	0.277	0.211	0.989	0.966	0.992	0.018	0.095	0.257	0.023	0.003	0.001
FCS-BB-PI	.39	-0.114	-0.101	-0.087	1.678	0.753	0.611	0.934	0.966	0.978	0.360	0.998	1.000	0.148	0.032	0.018
FCS-BB-PI	.59	-0.077	-0.074	-0.075	2.744	1.266	1.018	0.931	0.954	0.966	0.786	1.000	1.000	0.487	0.099	0.064
FCS-BB-JAV	.00	0.001	-0.000	0.000	0.825	0.179	0.111	0.995	1.000	1.000	0.005	0.000	0.000	0.010	0.000	0.000
FCS-BB-JAV	.14	0.015	-0.014	-0.024	1.035	0.319	0.243	0.991	0.982	0.996	0.011	0.091	0.264	0.034	0.003	0.002
FCS-BB-JAV	.39	0.042	0.011	0.020	2.076	0.884	0.714	0.949	0.981	0.994	0.351	0.996	1.000	0.217	0.036	0.018
FCS-BB-JAV	.59	0.049	0.034	0.017	3.446	1.512	1.208	0.941	0.977	0.995	0.786	1.000	1.000	0.864	0.105	0.045

Note. CC = complete case analysis; Mean = mean imputation; MBE = model-based estimation; DA = data augmentation; MICE = multiple imputation by chained equations; FCS-BB-PI = fully conditional specification using Bayesian bootstrapping and passive imputation; FCS-BB-JAV = fully conditional specification using Bayesian bootstrapping and just another variable imputation.

Table 4.29
Moderated Mediation Model Metrics – Mediator Categorical, Endogenous Continuous, 20% Missingness Per Variable

		Empirical Bias			*Confidence Interval Length*			*Coverage Probability*			*Rejection Rate*			*Mean Squared Error*		
Method	*Effect*	*100*	*500*	*1000*	*100*	*500*	*1000*	*100*	*500*	*1000*	*100*	*500*	*1000*	*100*	*500*	*1000*
CC	.00	0.002	0.000	0.000	0.677	0.110	0.054	1.000	1.000	1.000	0.000	0.000	0.000	0.022	0.000	0.000
CC	.14	0.128	0.035	0.018	0.933	0.260	0.174	0.984	0.896	0.929	0.001	0.059	0.294	0.053	0.005	0.002
CC	.39	0.163	0.035	0.004	2.092	0.769	0.523	0.935	0.941	0.932	0.067	0.989	1.000	0.311	0.038	0.019
CC	.59	0.147	0.033	0.016	3.376	1.286	0.892	0.942	0.953	0.956	0.355	1.000	1.000	0.874	0.111	0.054
Mean	.00	0.000	0.000	0.000	0.447	0.078	0.038	1.000	1.000	1.000	0.000	0.000	0.000	0.009	0.000	0.000
Mean	.14	-0.315	-0.351	-0.349	0.583	0.173	0.116	0.941	0.767	0.748	0.001	0.050	0.268	0.020	0.003	0.002
Mean	.39	-0.336	-0.388	-0.394	1.238	0.475	0.327	0.766	0.473	0.221	0.072	0.974	1.000	0.147	0.071	0.065
Mean	.59	-0.382	-0.428	-0.442	1.884	0.741	0.510	0.638	0.182	0.028	0.295	1.000	1.000	0.543	0.396	0.399
MBE	.00	0.002	0.000	0.000	0.668	0.088	0.042	0.998	0.997	0.989	0.002	0.003	0.011	0.006	0.000	0.000
MBE	.14	-0.440	-0.483	-0.496	0.840	0.156	0.096	0.925	0.850	0.739	0.017	0.279	0.661	0.016	0.003	0.002
MBE	.39	-0.450	-0.497	-0.504	1.493	0.409	0.272	0.905	0.327	0.053	0.295	0.993	1.000	0.154	0.100	0.099
MBE	.59	-0.463	-0.502	-0.509	2.256	0.659	0.449	0.820	0.086	0.003	0.625	0.999	1.000	0.619	0.514	0.515
DA	.00	0.003	0.000	0.000	0.606	0.105	0.051	1.000	1.000	1.000	0.000	0.000	0.000	0.009	0.000	0.000
DA	.14	-0.209	-0.234	-0.249	0.791	0.220	0.147	1.000	0.872	0.882	0.000	0.024	0.154	0.023	0.003	0.001
DA	.39	-0.206	-0.275	-0.279	1.676	0.620	0.427	0.902	0.772	0.616	0.009	0.969	1.000	0.137	0.045	0.038
DA	.59	-0.245	-0.293	-0.302	2.650	1.028	0.715	0.871	0.626	0.346	0.095	1.000	1.000	0.420	0.213	0.201
MICE	.00	0.002	-0.001	0.000	0.626	0.108	0.053	1.000	1.000	1.000	0.000	0.000	0.000	0.009	0.000	0.000
MICE	.14	-0.192	-0.201	-0.215	0.813	0.227	0.151	1.000	0.884	0.894	0.000	0.020	0.167	0.025	0.003	0.001
MICE	.39	-0.150	-0.226	-0.234	1.744	0.644	0.443	0.912	0.824	0.713	0.015	0.977	1.000	0.152	0.039	0.031
MICE	.59	-0.186	-0.235	-0.246	2.765	1.077	0.748	0.884	0.745	0.522	0.130	1.000	1.000	0.418	0.164	0.145
FCS-BB-PI	.00	0.002	-0.000	-0.000	0.691	0.159	0.099	0.998	1.000	1.000	0.002	0.000	0.000	0.006	0.000	0.000
FCS-BB-PI	.14	-0.184	-0.174	-0.161	0.851	0.269	0.203	0.994	0.960	0.988	0.013	0.069	0.149	0.019	0.003	0.002
FCS-BB-PI	.39	-0.144	-0.131	-0.122	1.611	0.699	0.558	0.935	0.912	0.896	0.310	0.983	1.000	0.150	0.042	0.033
FCS-BB-PI	.59	-0.154	-0.148	-0.138	2.473	1.146	0.922	0.912	0.834	0.770	0.726	1.000	1.000	0.441	0.174	0.156
FCS-BB-JAV	.00	0.003	-0.000	0.000	0.978	0.209	0.129	0.997	1.000	1.000	0.003	0.000	0.000	0.013	0.000	0.000
FCS-BB-JAV	.14	0.023	0.024	0.021	1.241	0.358	0.266	0.989	0.981	0.996	0.016	0.070	0.151	0.043	0.004	0.002
FCS-BB-JAV	.39	0.051	0.035	0.014	2.391	0.963	0.762	0.949	0.984	0.995	0.315	0.987	1.000	0.321	0.038	0.020
FCS-BB-JAV	.59	0.054	0.047	0.024	3.783	1.630	1.297	0.952	0.982	0.992	0.733	1.000	1.000	0.911	0.118	0.057

Note. CC = complete case analysis; Mean = mean imputation; MBE = model-based estimation; DA = data augmentation; MICE = multiple imputation by chained equations; FCS-BB-PI = fully conditional specification using Bayesian bootstrapping and passive imputation; FCS-BB-JAV = fully conditional specification using Bayesian bootstrapping and just another variable imputation.

imputation and MBE have poor coverage probabilities, with poorer performance in larger samples with larger effect sizes. For DA, MICE, and CC methods, CC has the best coverage performance, whereas, DA has the worst coverage performance across conditions.

All methods have lower than nominal Type I Error rates. In small samples, MBE and FCS methods have highest power across all conditions, with slightly higher power for FCS methods at larger effect sizes. On the contrary, DA and MICE have the lowest power in small sample sizes. In medium sample sizes, MBE has the highest power, whereas, other methods have comparable performance. In large sample sizes, MBE also has higher power than other methods; FCS methods have slightly lower power than other methods in at small effects with 10% missingness; with 20% missingness, CC and mean imputation have higher power than DA, MICE and FCS methods, which perform similarly. In terms of MSEs, estimates are highest for FCS-BB-JAV and CC in small samples, whereas, highest for MBE and mean imputation in medium and large sized samples; other methods perform comparable. With regards to MI methods, the FCS-BB-JAV method accurately estimates the fraction of missing information; other methods tend to overestimate the FMI (see Table 4.30). Supplemental figures for the results are presented in Appendix D.

Results of the final MC simulation for a moderated mediation model with categorical mediator variable and categorical endogenous variable and missingness rates of 10% and 20% are displayed in Table 4.31 and Table 4.32, respectively. In small sample sizes, DA, MICE, and FCS-BB (both PI and JAV) are the least biased methods, with lower biases for FCS-BB-JAV; other methods exhibit larger biases, with higher biases for MBE and mean imputation. In medium to large sample sizes, CC and FCS-BB-JAV are the least biased methods, with a tendency for CC to have lower biases. Across all conditions, MBE has the

Table 4.30
Moderated Mediation Model FMI Metric – Mediator Categorical, Endogenous Continuous

		FMI – 10% Missingness			*FMI – 20% Missingness*		
Method	*Effect*	*100*	*500*	*1000*	*100*	*500*	*1000*
DA	.00	0.099	0.093	0.093	0.170	0.166	0.165
DA	.14	0.108	0.108	0.113	0.190	0.195	0.202
DA	.39	0.136	0.136	0.140	0.227	0.237	0.234
DA	.59	0.167	0.163	0.168	0.261	0.270	0.274
MICE	.00	0.099	0.098	0.094	0.176	0.173	0.173
MICE	.14	0.110	0.107	0.110	0.190	0.205	0.210
MICE	.39	0.132	0.128	0.131	0.226	0.236	0.236
MICE	.59	0.156	0.152	0.152	0.252	0.263	0.269
FCS-BB-PI	.00	0.118	0.106	0.103	0.213	0.190	0.184
FCS-BB-PI	.14	0.121	0.111	0.109	0.218	0.202	0.200
FCS-BB-PI	.39	0.137	0.131	0.131	0.242	0.230	0.232
FCS-BB-PI	.59	0.158	0.148	0.146	0.259	0.258	0.255
FCS-BB-JAV	.00	0.104	0.094	0.091	0.199	0.180	0.172
FCS-BB-JAV	.14	0.103	0.092	0.094	0.198	0.185	0.184
FCS-BB-JAV	.39	0.110	0.095	0.094	0.200	0.194	0.188
FCS-BB-JAV	.59	0.105	0.097	0.097	0.200	0.191	0.187

Note. DA = data augmentation; MICE = multiple imputation by chained equations; FCS-BB-PI = fully conditional specification using Bayesian bootstrapping and passive imputation; FCS-BB-JAV = fully conditional specification using Bayesian bootstrapping and just another variable imputation.

narrowest confidence intervals, whereas FCS-BB-JAV has the widest, especially in small sample sizes. With regards to other methods, most differences arise in small sample sizes with larger effect sizes, with CC yielding the widest intervals and mean imputation yielding the narrowest intervals. For coverage probabilities, only the FCS-BB methods have approximately nominal or higher coverage rates across all conditions. In small samples mean imputation and MBE have poor coverage probabilities as effect sizes increase; in conditions with 10% missingness, mean imputation has better coverage than MBE, however, performance is similar for the two methods in conditions with 20% missingness. In small samples, other methods have approximately nominal or higher coverage

Table 4.31
Moderated Mediation Model Metrics – Mediator Categorical, Endogenous Categorical, 10% Missingness Per Variable

		Empirical Bias			*Confidence Interval Length*			*Coverage Probability*			*Rejection Rate*			*Mean Squared Error*		
Method	*Effect*	*100*	*500*	*1000*	*100*	*500*	*1000*	*100*	*500*	*1000*	*100*	*500*	*1000*	*100*	*500*	*1000*
CC	.00	-0.003	-0.003	0.000	1.189	0.182	0.089	1.000	1.000	1.000	0.000	0.000	0.000	0.070	0.001	0.000
CC	.14	0.278	0.073	0.006	1.527	0.400	0.260	0.989	0.943	0.916	0.000	0.025	0.093	0.154	0.011	0.005
CC	.39	0.318	0.051	0.003	3.375	1.120	0.761	0.955	0.943	0.949	0.011	0.663	0.945	0.954	0.091	0.036
CC	.59	0.297	0.042	0.017	5.387	1.883	1.286	0.943	0.954	0.945	0.065	0.940	1.000	2.478	0.244	0.107
Mean	.00	-0.008	-0.004	-0.001	1.038	0.199	0.116	1.000	1.000	1.000	0.000	0.000	0.000	0.060	0.002	0.001
Mean	.14	-0.040	-0.138	-0.164	1.304	0.365	0.239	0.985	0.858	0.770	0.000	0.032	0.182	0.111	0.013	0.008
Mean	.39	-0.058	-0.236	-0.301	2.676	0.912	0.613	0.902	0.755	0.575	0.014	0.506	0.679	0.588	0.114	0.094
Mean	.59	-0.141	-0.333	-0.367	3.990	1.412	0.958	0.846	0.597	0.382	0.064	0.733	0.956	1.463	0.436	0.388
MBE	.00	-0.001	-0.001	0.000	0.797	0.097	0.046	0.998	0.997	0.997	0.002	0.003	0.003	0.008	0.000	0.000
MBE	.14	-0.626	-0.660	-0.681	0.949	0.162	0.096	0.969	0.758	0.544	0.005	0.129	0.299	0.019	0.004	0.003
MBE	.39	-0.609	-0.677	-0.689	1.664	0.388	0.253	0.844	0.105	0.002	0.096	0.727	0.944	0.233	0.179	0.180
MBE	.59	-0.646	-0.698	-0.703	2.737	0.607	0.401	0.681	0.007	0.000	0.201	0.913	0.996	1.004	0.965	0.967
DA	.00	-0.004	-0.002	0.000	1.133	0.179	0.088	1.000	1.000	1.000	0.000	0.000	0.000	0.044	0.001	0.000
DA	.14	0.038	-0.083	-0.137	1.420	0.373	0.242	1.000	0.951	0.918	0.000	0.009	0.063	0.093	0.008	0.004
DA	.39	0.041	-0.125	-0.165	3.009	1.017	0.689	0.954	0.913	0.896	0.001	0.562	0.922	0.506	0.067	0.035
DA	.59	0.001	-0.158	-0.171	4.716	1.677	1.150	0.931	0.897	0.853	0.013	0.896	0.998	1.277	0.199	0.128
MICE	.00	-0.004	-0.002	0.000	1.158	0.183	0.090	1.000	1.000	1.000	0.000	0.000	0.000	0.048	0.001	0.000
MICE	.14	0.080	-0.057	-0.111	1.445	0.378	0.246	1.000	0.951	0.919	0.000	0.015	0.055	0.097	0.009	0.004
MICE	.39	0.083	-0.090	-0.132	3.066	1.041	0.703	0.962	0.923	0.909	0.001	0.586	0.922	0.570	0.069	0.034
MICE	.59	0.050	-0.115	-0.131	4.819	1.726	1.177	0.933	0.927	0.890	0.019	0.905	0.997	1.439	0.196	0.111
FCS-BB-PI	.00	-0.003	-0.002	0.000	1.574	0.311	0.190	0.999	1.000	1.000	0.001	0.000	0.000	0.033	0.001	0.000
FCS-BB-PI	.14	-0.040	-0.051	-0.049	1.847	0.498	0.361	0.995	0.992	0.984	0.006	0.027	0.035	0.078	0.007	0.003
FCS-BB-PI	.39	-0.061	-0.065	-0.072	3.416	1.219	0.943	0.955	0.968	0.985	0.129	0.618	0.850	0.559	0.068	0.033
FCS-BB-PI	.59	-0.081	-0.074	-0.081	5.159	1.988	1.560	0.931	0.970	0.978	0.363	0.914	0.995	1.511	0.192	0.109
FCS-BB-JAV	.00	-0.002	-0.002	0.000	1.976	0.359	0.217	0.998	1.000	1.000	0.002	0.000	0.000	0.055	0.001	0.000
FCS-BB-JAV	.14	-0.051	-0.047	-0.039	2.307	0.581	0.415	0.994	0.994	0.991	0.008	0.028	0.035	0.125	0.009	0.004
FCS-BB-JAV	.39	0.054	0.047	-0.013	4.422	1.444	1.108	0.954	0.983	0.996	0.140	0.620	0.846	1.023	0.090	0.035
FCS-BB-JAV	.59	0.064	0.055	0.023	6.837	2.417	1.876	0.935	0.982	0.995	0.364	0.917	0.996	2.812	0.255	0.110

Note. CC = complete case analysis; Mean = mean imputation; MBE = model-based estimation; DA = data augmentation; MICE = multiple imputation by chained equations; FCS-BB-PI = fully conditional specification using Bayesian bootstrapping and passive imputation; FCS-BB-JAV = fully conditional specification using Bayesian bootstrapping and just another variable imputation.

Table 4.32
Moderated Mediation Model Metrics – Mediator Categorical, Endogenous Categorical, 20% Missingness Per Variable

		Empirical Bias			*Confidence Interval Length*			*Coverage Probability*			*Rejection Rate*			*Mean Squared Error*		
Method	*Effect*	*100*	*500*	*1000*	*100*	*500*	*1000*	*100*	*500*	*1000*	*100*	*500*	*1000*	*100*	*500*	*1000*
CC	.00	-0.012	-0.001	-0.001	1.413	0.222	0.110	1.000	1.000	1.000	0.000	0.000	0.000	0.088	0.002	0.000
CC	.14	0.249	0.113	0.015	1.913	0.445	0.286	0.995	0.957	0.917	0.000	0.016	0.078	0.249	0.014	0.006
CC	.39	0.385	0.033	0.021	3.767	1.221	0.832	0.951	0.953	0.945	0.007	0.539	0.906	1.350	0.095	0.049
CC	.59	0.336	0.041	0.014	6.201	2.048	1.398	0.942	0.944	0.950	0.028	0.892	0.997	3.913	0.271	0.127
Mean	.00	-0.011	-0.005	0.000	1.177	0.276	0.178	1.000	1.000	0.995	0.000	0.000	0.005	0.068	0.005	0.002
Mean	.14	-0.195	-0.211	-0.284	1.473	0.398	0.266	0.987	0.770	0.559	0.000	0.064	0.247	0.162	0.023	0.017
Mean	.39	-0.175	-0.446	-0.487	2.547	0.841	0.570	0.837	0.517	0.380	0.007	0.379	0.418	0.632	0.201	0.196
Mean	.59	-0.322	-0.489	-0.573	3.676	1.281	0.833	0.750	0.398	0.229	0.038	0.487	0.601	1.520	0.784	0.850
MBE	.00	-0.006	-0.001	0.000	0.996	0.116	0.054	0.997	0.996	0.998	0.003	0.004	0.002	0.010	0.000	0.000
MBE	.14	-0.649	-0.653	-0.675	1.217	0.182	0.106	0.976	0.776	0.567	0.008	0.103	0.255	0.027	0.004	0.003
MBE	.39	-0.606	-0.682	-0.686	1.987	0.426	0.276	0.865	0.139	0.008	0.062	0.625	0.899	0.243	0.182	0.179
MBE	.59	-0.635	-0.695	-0.701	2.959	0.660	0.439	0.771	0.009	0.000	0.129	0.883	0.992	1.048	0.962	0.965
DA	.00	-0.003	-0.001	0.000	1.283	0.206	0.104	1.000	1.000	1.000	0.000	0.000	0.000	0.041	0.001	0.000
DA	.14	-0.089	-0.205	-0.273	1.616	0.379	0.244	1.000	0.955	0.907	0.000	0.004	0.026	0.102	0.007	0.004
DA	.39	-0.112	-0.292	-0.291	3.012	0.987	0.674	0.949	0.859	0.785	0.000	0.323	0.798	0.408	0.074	0.053
DA	.59	-0.204	-0.321	-0.336	4.595	1.590	1.083	0.916	0.778	0.586	0.000	0.746	0.992	1.027	0.307	0.268
MICE	.00	-0.003	0.000	0.000	1.337	0.214	0.107	1.000	1.000	1.000	0.000	0.000	0.000	0.048	0.001	0.000
MICE	.14	-0.027	-0.148	-0.219	1.673	0.395	0.253	1.000	0.958	0.913	0.000	0.005	0.034	0.119	0.008	0.004
MICE	.39	-0.009	-0.229	-0.228	3.186	1.035	0.705	0.947	0.896	0.842	0.000	0.382	0.828	0.498	0.071	0.045
MICE	.59	-0.113	-0.249	-0.264	4.899	1.691	1.153	0.927	0.836	0.747	0.002	0.786	0.990	1.205	0.256	0.197
FCS-BB-PI	.00	-0.003	0.000	0.000	1.495	0.305	0.189	0.998	1.000	1.000	0.002	0.000	0.000	0.028	0.001	0.000
FCS-BB-PI	.14	-0.160	-0.151	-0.145	1.778	0.468	0.338	0.991	0.995	0.982	0.009	0.025	0.038	0.081	0.006	0.004
FCS-BB-PI	.39	-0.092	-0.085	-0.084	3.021	1.091	0.854	0.943	0.950	0.966	0.115	0.548	0.813	0.422	0.069	0.046
FCS-BB-PI	.59	-0.132	-0.125	-0.124	4.467	1.732	1.366	0.928	0.906	0.902	0.283	0.908	0.991	1.102	0.246	0.199
FCS-BB-JAV	.00	-0.012	-0.001	0.000	2.341	0.414	0.252	0.999	1.000	1.000	0.001	0.000	0.000	0.066	0.001	0.000
FCS-BB-JAV	.14	0.024	-0.039	-0.099	2.804	0.640	0.455	0.992	0.995	0.991	0.010	0.023	0.031	0.195	0.011	0.005
FCS-BB-JAV	.39	0.047	0.036	0.015	5.029	1.557	1.190	0.963	0.988	0.992	0.120	0.545	0.794	1.420	0.094	0.046
FCS-BB-JAV	.59	0.067	0.055	0.021	7.745	2.559	1.988	0.934	0.982	0.994	0.279	0.887	0.991	3.598	0.262	0.124

Note. CC = complete case analysis; Mean = mean imputation; MBE = model-based estimation; DA = data augmentation; MICE = multiple imputation by chained equations; FCS-BB-PI = fully conditional specification using Bayesian bootstrapping and passive imputation; FCS-BB-JAV = fully conditional specification using Bayesian bootstrapping and just another variable imputation.

probabilities. In medium and large samples, both mean imputation and MBE have considerably low coverage as effect sizes increases; among DA, MICE, and CC, CC generally has the highest coverage probabilities, whereas, DA has the lowest coverage probabilities.

All methods have lower than nominal Type I Error rates. In small samples, MBE and FCS methods have highest power across all conditions, with higher power for FCS methods at larger effect sizes. In medium sample sizes, MBE has the highest power; CC and FCS methods have higher power for larger effect sizes than other methods. In large samples, both MBE and mean imputation have higher power than other methods at small effects; however, as effect sizes increase, mean imputation has considerably lower power than other methods. Among other methods, with 10% missingness, CC has slightly higher power, whereas, FCS methods have slightly lower power; with 20% missingness, CC has higher power, whereas, DA, MICE, and FCS methods perform similarly. In terms of MSEs, estimates are highest for FCS-BB-JAV and CC in small samples, whereas, highest for MBE in medium and large sized samples. Mean imputation also has larger MSEs compared to other methods, especially with more missingness and larger effect sizes; other methods perform comparable. With regards to MI methods, the FCS-BB-JAV method accurately estimates the fraction of missing information; other methods tend to overestimate the FMI at larger effect sizes (see Table 4.33). Supplemental figures for the results are presented in Appendix D.

Table 4.33
Moderated Mediation Model FMI Metric – Mediator Categorical, Endogenous Categorical

		FMI – 10% Missingness			*FMI – 20% Missingness*		
Method	*Effect*	*100*	*500*	*1000*	*100*	*500*	*1000*
DA	.00	0.099	0.092	0.092	0.171	0.165	0.168
DA	.14	0.109	0.110	0.110	0.191	0.194	0.204
DA	.39	0.137	0.128	0.131	0.225	0.231	0.230
DA	.59	0.155	0.152	0.151	0.248	0.251	0.250
MICE	.00	0.099	0.095	0.093	0.173	0.173	0.174
MICE	.14	0.109	0.111	0.115	0.189	0.203	0.214
MICE	.39	0.131	0.131	0.128	0.227	0.238	0.233
MICE	.59	0.148	0.148	0.141	0.246	0.255	0.256
FCS-BB-PI	.00	0.131	0.105	0.099	0.221	0.190	0.184
FCS-BB-PI	.14	0.131	0.107	0.111	0.236	0.203	0.203
FCS-BB-PI	.39	0.148	0.133	0.130	0.250	0.231	0.225
FCS-BB-PI	.59	0.163	0.145	0.140	0.268	0.243	0.247
FCS-BB-JAV	.00	0.110	0.093	0.087	0.208	0.181	0.176
FCS-BB-JAV	.14	0.112	0.096	0.092	0.211	0.181	0.180
FCS-BB-JAV	.39	0.116	0.097	0.092	0.217	0.187	0.184
FCS-BB-JAV	.59	0.121	0.101	0.094	0.220	0.190	0.183

Note. DA = data augmentation; MICE = multiple imputation by chained equations; FCS-BB-PI = fully conditional specification using Bayesian bootstrapping and passive imputation; FCS-BB-JAV = fully conditional specification using Bayesian bootstrapping and just another variable imputation.

CHAPTER 5

DISCUSSION

5.1. Study 1

The purpose of this study was to examine empirical performance of Bayesian bootstrapping (BB) in estimating and testing indirect-type effects that occur in mediation and moderated mediation models. Monte Carlo (MC) simulations were conducted to determine the relative performance of BB to commonly used methods in the literature such as the delta methods (first- and second-order) and the bias-corrected (BC) bootstrap. Four different mediator/endogenous variable combinations (i.e., continuous/continuous, continuous/categorical, categorical/continuous, categorical/categorical) were examined for both mediation and moderated mediation models. Methods were compared based on empirical biases, confidence interval lengths, coverage probabilities, rejection rates (e.g., Type I Error, power), and mean squared errors.

Results from our studied conditions highlight several important points about using BB for testing indirect-type effects. Most importantly, the indirect effect estimator (e.g., mean or median) has a significant influence on the empirical bias of the indirect effect estimate. Specifically, for continuous mediators and endogenous variables, the mean estimator tended to have lower bias than the median estimator in both mediation and moderated mediation models. For mediation models with at least one categorical response variable (i.e., mediator and/or endogenous variable), in general the median estimator had lower bias in small sample sizes for non-null effects and in larger samples (i.e., $n = 500$ and 1,000) with larger effects (i.e., effects = .39 and .59), whereas the mean estimator had

lower bias in larger samples with small effects. Although similar effects were found in moderated mediation models, two interesting differences appeared in which the median estimator had lowest bias: (1) for continuous mediators and categorical endogenous variables, the median estimator had lower bias than the mean estimator across all non-null effects, and (2) for categorical mediators and endogenous variables, the median estimator had lower bias than the mean estimator in medium sample sizes with small effects.

In terms of empirical biases, the differences in performance of the mean and median estimators may be attributable to a combination of two factors. First, these differences in bias performance highlight the fact that the posterior distribution of the indirect effect is skewed, especially when one or more response variables are categorical. For skewed posterior distributions, researchers have found that different Bayesian estimators combined with different prior distributions can impact empirical performance (e.g., Bayes & Branco, 2007; Browne & Draper, 2006). Although no previous research to our knowledge has systematically compared different posterior estimators of the unconditional indirect effect from Bayesian mediation models, Wang and Preacher (2015) found that for continuous response variables, the mean estimator yielded more accurate results than a median estimator for conditional indirect effects from Bayesian moderated mediation models. Results from the current study echo findings from the Wang and Preacher (2015) study for Bayesian moderated mediation models. When at least one response variable is categorical, however, our results indicate that BB estimator should be selected based on sample size and (expected) effect size.

With regards to the choice of prior distribution for BB, we used the improper prior specified in Rubin (1981). However, as Hastie et al. (2009) note, this choice of improper

(and non-informative) prior distribution can be viewed as a "poor man's Bayesian model". In essence, this choice of prior distribution has an effect of smoothing the bootstrap distribution. Future research should investigate other prior distributions to determine the effect of prior distribution on the empirical performance of the BB for indirect effects analysis.

Second, the differences in bias of the Bayesian estimators may be influenced by the two-stage BB sampling scheme. Specifically, we used the same number of samples (i.e., 1,000) for stage one (number of bootstrap samples) and stage two (sample size draws) sampling. As such, this sampling scheme corresponds to drawing samples at stage two that are 10 times, two times, and equal to the sample sizes of small ($n = 100$), medium ($n = 500$), and large ($n = 1{,}000$) samples, respectively, examined in the present study. Based on the current results, it appears that the sampling scheme is reasonable for testing indirect-type effects in larger sample sizes with larger effects, but for smaller samples and effect sizes, more samples may need to be drawn to see if empirical bias decreases for the BB.

Relative to frequentist methods (i.e., delta methods and BC bootstrap), for continuous response variables the BB with mean estimator had comparable bias across all conditions. On the contrary, when at least one response variable was categorical, the results are more complicated. Generally, the BB (depending on the estimator) tended to have comparable or slightly higher biases than the frequentist methods, except for in small samples with small effects where the BB methods tended to have comparable or slightly lower biases (especially with conditional indirect effects). Given that the BC bootstrap is considered a gold standard method for indirect effect analysis, these results show promise for the use of the BB in certain conditions. Despite these somewhat mixed results for

empirical biases, the current findings highlight an important finding for indirect effect analysis in small sample sizes. That is, regardless of the method, indirect effect (both unconditional and conditional) estimates are often biased. This bias, however, disappears as sample size increases, across all effect sizes, and corroborates the need for larger sample sizes for indirect effects analysis.

In terms of confidence interval lengths, with continuous responses in both mediation and moderated mediation models the BB had comparable performance to the delta methods and BC bootstrap in small samples but slightly wider intervals in larger sample sizes. Across conditions with at least one categorical response, the BC bootstrap tended to have slightly wider intervals in small samples relative other methods, whereas, the BB methods had slightly wider intervals than the delta methods; in larger samples, the BB methods tended to have slightly wider intervals than both BC bootstrap and delta methods. Relative to other methods, the BB methods generally had wider intervals at larger effect sizes. Although previous research has shown that the BB often leads to similar and sometimes narrower confidence intervals than nonparametric bootstrap methods (Taddy et al., 2015), our results only partial support this finding. Specifically, we found comparable performance between the BB and nonparametric bootstrap at smaller effect sizes, however, at larger effect sizes, results varied depending on the response variable type (i.e., continuous or categorical).

Coverage probabilities for the BB methods tended to be higher than compared to other methods, which may reflect the consequence of BB methods yielding wider confidence intervals. In fact, aside from one condition (i.e., large effects in small samples with moderated mediation models and continuous mediator and categorical endogenous

variables), BB methods were the only methods that had approximately nominal (e.g., ≥ .93) coverage probabilities or higher across all examined conditions. Moreover, in the condition where coverage rates dropped below nominal levels, the BB methods still had slightly higher coverage than the BC bootstrap. Overall, these findings demonstrate that BB methods yielded higher coverage rates than frequentist methods and support prior simulation results for Bayesian mediation analysis (Yang & MacKinnon, 2009). For moderated mediation results, however, prior research has demonstrated comparable coverage performance of Bayesian methods to bootstrapping methods, whereas, our results demonstrated consistently higher coverage for BB methods. As Yang and MacKinnon (2009) suggest, simulations conducted in a frequentist setting tend to favor frequentist approaches to estimation. If the data were generated from a Bayesian model, the coverage of the credible intervals would more closely match nominal coverage levels.

For null effect sizes, the BB and other methods produced Type I Error rates that were well below nominal coverage rates (i.e., .05). These results are consistent with findings using Bayesian methods with non-informative priors and frequentist methods (Preacher et al., 2007; Yuan & MacKinnon, 2009; Wang & Preacher, 2015) for indirect effects analysis. With regards to non-null effects, however, power levels changed as a function of sample size and effect size. Specifically, in small samples with smaller effects, the bootstrap methods yielded higher power than the delta methods, with the highest power among the BC bootstrap. Given that the delta methods are based on asymptotic variances, these findings are not surprising and support previous research (Preacher et al., 2007; Yuan & MacKinnon, 2009; Wang & Preacher, 2015). In addition, the findings that the

power of Bayesian methods is slightly lower than BC bootstrap in small samples for some conditions supports results in Wang and Preacher (2015).

For larger sample sizes, the BC bootstrap had higher power for smaller effect sizes, whereas, the BB methods performed similar to delta methods in medium sample sizes; in large samples with small effects, surprisingly the BB methods consistently had lower power than other methods. These findings are somewhat inconsistent with previous research, which has demonstrated that in larger sample sizes the Bayesian methods tend to outperform frequentist methods (Yuan & MacKinnon, 2009; Wang & Preacher, 2015). One explanation for the better performance of the BB methods in smaller samples may be due to the resampling scheme described earlier. If a larger number of samples are drawn at stage two, the Power of the BB methods may become more consistent with the BC bootstrap at larger sample sizes.

Across most conditions, MSEs for the BB methods were relatively similar to other methods. As can be expected, because MSE is a function of bias and variance of an estimator, the largest differences in MSEs were observed in conditions that had demonstrated highest variability in bias across methods. In particular, when at least one response variable was categorical, the BB methods yielded highest MSEs in larger samples. Among the BB methods, however, the estimator with the lowest bias yielded smallest MSE. The conditions in which MSEs were similar supports previous research (Yuan & MacKinnon, 2009; Wang & Preacher, 2015). One explanation for the differences in MSEs in other conditions may be due to simulation error of BB methods. In practical applications, bootstrap testing of indirect effects is often implemented with significantly more samples (i.e., bootstrap samples >> 1,000) than investigated in the current study. As such, if larger

bootstrap samples are drawn, the variability in estimates may diminish (thus reducing simulation error and in turn bias and MSE).

5.2. Study 2

Study 1 demonstrated that the BB method for indirect effects analysis has comparable performance to other popular methods such as the delta and BC bootstrap methods. As such, the purpose of this study was to examine the performance of the BB in a missing data context for indirect effects analysis. Specifically, we examined the empirical performance of a novel MI imputation algorithm (FCS-BB) that uses BB to simulate an indirect effect's posterior distribution and a unified imputation framework based on gradient boosting that can model both linear and nonlinear effects. Monte Carlo (MC) simulations were conducted to determine the relative performance of the FCS-BB to several commonly used methods in the literature such as complete case analysis (CC), mean imputation, the model-based estimation with BC bootstrapping (MBE), data augmentation (CA), and multiple imputation by chained equations (MICE). Four different mediator/endogenous variable combinations (i.e., continuous/continuous, continuous/categorical, categorical/continuous, categorical/categorical) were examined for both mediation and moderated mediation models. Methods were compared based on empirical biases, confidence interval lengths, coverage probabilities, rejection rates (e.g., Type I Error, power), mean squared errors (MSEs), and fraction of missing information (FMI) in samples with 10% and 20% on select variables (i.e., all variables in mediation models, only response variables in moderated mediation models).

Several important points can be made about using FCS-BB for indirect effects analysis based on the results from our MC simulations. Most importantly, in the case of MI

using linear boosters, prediction error and/or parameter uncertainty must be incorporated into the imputation scheme. For linear boosters with continuous response variables, we incorporated prediction error by adding a normally distributed random error to predicted values based on the residual variance of the observed data. This method uses a similar technique to incorporate prediction error as stochastic regression imputation (Little & Rubin, 2002). On the contrary, for linear boosters for categorical response variables, originally we did not incorporate prediction error or parameter uncertainty into the imputations, which resulted in poor performance across all conditions examined. However, once we modified the gradient boosting imputation scheme to incorporate parameter uncertainty using techniques similar to Bayesian logistic regression, we found that our results improved substantially across all conditions.

In models with continuous response variables, although MBE had the best overall performance, FCS methods (i.e., FCS-BB and FCS-BB-JAV) had comparable performance across all metrics except confidence interval length and power; MBE had narrower confidence intervals and higher power in smaller sample sizes with smaller effect sizes. These findings echo previous results found in Enders et al. (2013) for Bayesian missing data models versus frequentist models based on nonparameteric bootstrapping. In addition, our findings support previous research demonstrating similar empirical performance for DA and MICE in the case of imputing continuous response variables (Karangwa, 2013; Kropko et al., 2014; Raghunathan et al., 2001).

Furthermore, FCS methods generally performed better (especially in moderated mediation models) than other MI methods (i.e., DA and MICE) and CC in terms of coverage probabilities and power (in smaller samples), but had wider confidence intervals. It is

important to note that in the case of continuous response variables and linear boosters, FCS-BB bears close resemblance to MICE and DA in terms of imputation models. More specifically, all three methods implement a type of iterative stochastic linear regression model to impute missing values. As such, these findings highlight the added benefit of including resampling strategies for MI schemes, which have been demonstrated previously (van Buuren, 2012).

For conditions in which at least one response variable was categorical, MI methods (i.e., DA, MICE, FCS-BB*) tended to outperform non-MI methods in mediation models, whereas in moderated mediation models, FCS-BB-JAV and CC analysis performed better than other methods (including DA and MICE). Surprisingly, although MBE tended to have highest power, it was generally the most biased method and had the worst confidence interval coverage, demonstrating higher bias and worse coverage probabilities than mean imputation in most conditions. It is important to note that the underlying estimation algorithm for MBE changes when at least one categorical response variable is present. That is, as opposed to using a full information maximum likelihood approach to estimate parameters in the presence of missing data with all continuous response variables, a limited information weighted least squares approach is used instead. Although the latter estimation algorithm is supposed to produce consistent estimates when data are MAR (Asparouhov & Muthén, 2010), the findings from of our study do not provide empirical support.

Among the results of MI methods for mediation models with at least one categorical response variable, FCS-BB* tended to have the least biased estimates, highest coverage probabilities, highest power in small samples, but also widest confidence intervals and

slightly higher MSEs. With regards to other MI methods, MICE tended to outperform DA across all conditions, especially in terms of empirical bias. These findings are not surprisingly given the implementations of DA and MICE in the current study. Specifically, our implementation of DA to impute categorical variables relied on thresholding imputed categorical values, whereas, our implementation of MICE used a Bayesian logistic regression model to impute categorical variables. The latter model is specifically designed to handle categorical response variables, whereas, the former model is specifically designed to handle normally distributed continuous response variables. Although previous research (Finch, 2010; Kropko et al., 2014; Lee & Carlin, 2010) has been inconclusive in determining whether DA or MICE has better empirical performance with categorical (i.e., binary) response variables, our results provide empirical support for using MICE as opposed to DA in such cases.

Among findings of MI methods for moderated mediation models with at least one categorical response variable, FCS-BB-JAV resulted in the best performance across almost all metrics. Specifically, FCS-BB-JAV tended to have the least biased estimates, highest coverage probabilities, highest power, but widest confidence intervals. On the contrary, across some conditions in small samples, FCS-BB-JAV had the highest MSEs. Given that the bias estimates were relatively low and MSE is a function of bias and variance, it appears that the high MSEs are primarily a function of the variance of the imputation estimates. Inherently, tree models yield low bias, but high variance estimates (Hastie et al., 2009). Despite the fact that boosting is a method that was designed to reduce the variance of tree models, it appears that in small samples with mixed variable types and interaction effects, tree-based models tend to yield highly variable estimates in missing data contexts. Put

another way, compared to other conditions examined, in these conditions there appears to be a higher bias/variance trade-off for using tree boosted imputation models.

Interestingly, CC analysis outperformed both DA and MICE across most conditions. Neither DA nor MICE are designed to explicitly impute interactive effects (both use PI), therefore, this finding underscores the negative effects that misspecified imputation models can have on regression coefficients. Comparing the FCS-BB methods, estimates from JAV were less biased and had higher coverage probabilities than PI. Therefore, the poorer performance among PI methods (i.e., DA, MICE, and FCS-BB-PI) relative to JAV highlights the importance of using imputation models that explicitly incorporate the same effects present in underlying substantive models. In other words, our findings support Allison's (2003) suggestion that when interactions are present in a model with missing data, these effects should also be included into an imputation model.

In addition, these findings support previous claims by von Hippel (2009) that JAV is superior to PI in both linear and logistic (or equivalently probit) regression models. Prior work by Seaman et al. (2012) demonstrated that both JAV and PI resulted in substantial bias for imputing quadratic covariates with MCAR and MAR missingness mechanisms; however, our results demonstrated that when combined with tree-based imputation models for categorical variables, the JAV approach led to confidence proper imputations across most conditions. In terms of our imputation methodology, combining MI with bootstrapping has previously demonstrated good performance for indirect effects analysis in mediation models with continuous variables (Wang & Wang, 2014). Collectively, our results extend these findings and provide empirical support for combining MI with Bayesian bootstrapping for indirect effects analysis in both mediation and moderated

mediation analysis and with both continuous and categorical response variables. However, our results also underscore the importance of selecting an appropriate imputation model that contains the similar effects as the underlying substantive model.

5.3. Study Strengths

There are several noteworthy strengths of the present study. First, this is the first comprehensive simulation study to our knowledge that systematically compared different estimation methods and missing data methods in mediation and moderated mediation models with both continuous and categorical variables. Second, we demonstrated the potential benefits of using BB for Bayesian inference in indirect effects analysis. Specifically, BB can be used to generate posterior inferences of indirect effects without having to specify a fully Bayesian model (Rubin, 1981). Lastly, we introduced a unified MI framework based on gradient boosted models for imputing linear and nonlinear effects and demonstrated the empirical performance of this framework across different conditions. A strength of gradient boosted models is that these models can be applied to continuous, categorical (binary, ordered categorical, unordered categorical), count, and censored data by using the same fundamental algorithm but changing the loss function to be optimized (Friedman, 2001).

5.4. Study Limitations and Future Research

As with any research, our study is not without limitations. Even though we examined many conditions in our simulation studies, we did not examine different combinations of effect sizes (e.g., small/large, null/large, etc.) as done in other studies (Enders et al., 2014, Koopman et al., 2015). In examining different combinations of effect sizes, we would obtain more accurate empirical estimates, especially empirical Type I Error

rates, that may reflect metrics which generalize better to practical settings. In addition, we did not examine multiple moderator values for conditional indirect effects. It is likely that empirical estimates of conditional indirect effects, such as the ones examined in this study, would change as a function of different moderator values. We also did not compare the potential effects of varying the sampling scheme (e.g., vary stage 1 and/or stage 2 samples) for Bayesian bootstrapping. Furthermore, because our results are based on using linear models, these cannot be generalized to settings in which underlying linear model assumptions are violated (e.g., multilevel data with correlated errors).

In addition to addressing the issues above for future research, there are several other important research avenues to explore as well. First, with the recent popularity of Bayesian methods for indirect effects analysis, a comprehensive comparison of the Bayesian bootstrap with a fully Bayesian model for indirect effects analysis would provide further insight into the relative performance of these Bayesian methods. Second, although increasing the number of bootstrap samples (at stage 1 and stage 2) may alleviate the issue, future research should attempt to identify the underlying differences in posterior estimators for indirect effects that we found in Study 1 using Bayesian bootstrapping. Third, an interesting avenue for future research would be to examine the performance of using gradient boosted imputation schemes to impute non-normal data, count data, and multi-categorical data. Lastly, MI schemes using gradient boosted imputers should attempt to incorporate and optimize automatic hyperparameter tuning into the imputation pipeline using methods such as cross-validation with grid research.

5.5. Conclusion

In summary, our findings demonstrate that BB is a useful resampling technique that can be used to generate posterior inferences for indirect effects analysis. With the appropriate posterior estimator, the BB yields comparable performance to different delta method approximations (i.e., first- and second-order) and the BC bootstrap across different models and combinations of variable types. Moreover, these initial findings suggest that the BB demonstrates similar empirical performance to fully Bayesian models for indirect effects analysis, but without the need to explicitly specify prior distributions for all model parameters. In practice, these latter findings imply that the BB can used by researchers in situations where Bayesian inference is the goal, but fully Bayesian models are complicated (e.g., with categorical response variables). For example, in research involving rare events, such as substance abuse or domestic violence, the response variables of interest may be binary (e.g., meets clinical diagnosis for drug abuse/does not meet clinical diagnosis for drug abuse, perpetrated violence in last year/did not perpetrate violence in last year). In these applications, the BB can be used to generate posterior inferences for the indirect effect(s) of interest.

With missing data, our MI algorithm that uses BB and gradient boosted imputation models tends to perform as well as or better than commonly used missing data techniques for indirect effects analysis. In particular, our MI algorithm demonstrates properties of confidence proper imputation procedures (Rubin, 1996), that is, estimates which are relatively unbiased and obtain approximately nominal coverage probabilities. Furthermore, our findings demonstrate the benefits of the JAV method for handling interactions with missing data and corroborate the need for imputation models to at least

contain the same linear and nonlinear effects that are present in underlying substantive models.

REFERENCES

Abrahantes, J. C., Sotto, C., Molenberghs, G., Vromman, G., & Bierinckx, B. (2011). A comparison of various software tools for dealing with missing data via imputation. *Journal of Statistical Computation and Simulation, 81*, 1653-1657. doi: 10.1080/00949655.2010.498788

Aldrich, J. (1997). R. A. Fisher and the making of maximum likelihood. *Statistical Science, 12*, 162-176. doi: 10.1214/ss/1030037906

Allison, P. D. (2003). Missing data techniques for structural equation modeling. *Journal of Abnormal Psychology, 112*, 545-557. doi: 10.1037/0021-843X.112.4.545

Arbuckle, J. L. (1996). Full information estimation in the presence of incomplete data. In G. A. Marcoulides & R. E. Schumackers (Eds.), *Advanced Structural Equation Modeling* (pp. 243-277). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Arnold, B. C., Castillo, E., & Sarabia, J. M. (1999). *Conditional specification of statistical models*. New York, NY: Springer-Verlag.

Arnold, B. C., & Press, S. J. (1989). Compatible conditional distributions. *Journal of the American Statistical Association, 84*, 152-156. doi: 10.1080/01621459.1989.10478750

Asparouhov, T., & Muthén, B. (2010). *Weighted least squares estimation with missing data*. Retrieved from https://www.statmodel.com/download/GstrucMissingRevision.pdf

Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: What is it and how does it work? *International Journal of Methods in Psychiatric Research, 20*, 40-49. doi: 10.1002/mpr.329

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*, 1173-1182. doi: 10.1037/0022-3514.51.6.1173

Bartlett, J. W., Seaman, S. R., White, I. R., & Carpenter, J. R. (2014). Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. *Statistical Methods in Medical Research,* 0, 1-26. doi: 10.1177/0962280214521348

Basu, D. (1977). On the elimination of nuisance parameters. *Journal of the American Statistical Association, 77*, 355-366. doi: 10.1080/01621459.1977.10481002

Bayes, C. L. & Branco, M. D. (2007). Bayesian inference for the skewness parameter of the scalar skew-normal distribution. *Brazilian Journal of Probability and Statistics, 21*(2), 141-163.

Ben-Israel, A., & Greville, T. N. E. (1974). *Generalized inverses: Theory and applications*. New York, NY: Wiley.

Bentler, P. M., & Freeman, E. H. (1983). Tests for stability in linear structural equation systems. *Psychometrika, 48*(1), 143-145.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, B, 36*, 192-236.

Bickel, P. J., & Freedman, D. A. (1981). Some asymptotic theory for the bootstrap. *Annals of Statistics, 9*(6), 1196-1217.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York, NY: Springer.

Bollen, K. A. (1987). Total, direct, and indirect effects in structural equation models. *Sociological Methodology, 17*, 37-69. doi: 10.2307/271028

Bollen, K. A. (1989). *Structural equations with latent variables*. New York, NY: John Wiley.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Belmont, CA: Wadsworth.

Browne, W. J., & Draper, D. (2006). A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis*, *1*(3), 473-514.

Bühlmann, P., & Hothorn, T. (2007). Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science, 22*, 477-505. doi: 10.1214/07-STS242

Burgette, L. F., & Reiter, J. P. (2010). Multiple imputation for missing data via sequential regression trees. *American Journal of Epidemiology, 172*, 1070-1076. doi: 10.1093/aje/kwq260

Casella, G., & Berger, R. L. (2002). *Statistical inference* (2nd ed.). Australia: Thomson Learning.

Chen, H. Y. (2010). Compatibility of conditionally specified models. *Statistics and Probability Letters, 80*, 670-677. doi: 10.1016/j.spl.2009.12.025

Chen, T., Guestrin, C. (preprint 2016). XGBoost: A scalable tree boosting system *arXiv*:1603.02754

Cheung, M. W. L. (2007). Comparison of approaches to constructing confidence intervals for mediating effects using structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal, 14*, 227-246. doi: 10.1080/10705510709336745

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, M. P. (1997). The Bayesian bootstrap and multiple imputation for unequal probability sample designs. *American Statistical Association Proceedings of the Survey Research Methods Section*, 653-638.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analyses for the behavioral sciences* (3rd ed.). Hillsdale, NJ: Lawrence Erlbaum

Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods, 6*, 330-351. doi: 10.1037/1082-989X.6.4.330

Cook, R. J., Zheng, L., & Yi, G. Y. (2004). Marginal analysis of incomplete longitudinal binary data: A cautionary note on LOCF imputation. *Biometrics, 60*, 820-828. doi: 10.1111/j.0006-341X.2004.00234.x

Cox, D. R., & Hinkley, D. V. (1974). *Theoretical statistics*. Boca Raton, FL: Chapman & Hall.

Davis, C. S. (2002). *Statistical methods for the analysis of repeated measurements*. New York, NY: Springer-Verlag.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, B, 39*, 1-38.

Dettling, M., & Bühlmann, P. (2002). Boosting for tumor classification with gene

expression data. *Bioinformatics, 19*, 1061–1069. doi: 10.1093/bioinformatics/btf867

Doove, L. L., van Buuren, S., & Dusseldorp, E. (2014). Recursive partitioning for missing data imputation in the presence of interaction effects. *Computational Statistics and Data Analysis, 72*, 92-104. doi: 10.1016/j.csda.2013.10.025

Efron, B. (1981). Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods. *Biometrika, 68*, 589-599. doi: 10.1093/biomet/68.3.589

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association, 82*, 171-185. doi:10.2307/2289144

Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap*. Boca Raton, FL: Chapman & Hall/CRC.

Enders, C. K. (2001). A primer on maximum likelihood algorithms available for use with missing data. *Structural Equation Modeling, 8*, 128-141. doi: 10.1207/S15328007SEM0801_7

Enders. C. K. (2010). *Applied missing data analysis*. New York, NY: Guilford Press.

Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling, 8*, 430-457. doi: 10.1207/S15328007SEM0803_5

Enders, C. K., Baraldi, A. N., & Cham, H. (2014). Estimating interaction effects with incomplete predictor variables. *Psychological Methods, 19*, 39-55. doi: 10.1037/a0035314

Enders, C. K., Fairchild, A. J., & MacKinnon, D. P. (2013). A Bayesian approach for

estimating mediation effects with missing data. *Multivariate Behavioral Research, 48*, 340-369. doi: 10.1080/00273171.2013.784862

Feelders, A. (1999). Handling missing data in trees: Surrogate splits or statistical imputation*? Principles of Data Mining and Knowledge Discovery, 1704*, 329-334.

Finch, W. H. (2010). Imputation methods for missing categorical questionnaire data: A comparison of approaches. *Journal of Data Science, 8*(3), 361-378.

Freund, Y, & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences, 55*, 119-139. doi: 10.1006/jcss.1997.1504

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*(5), 1189-1232.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis, 38*, 367–378. doi: 10.1016/S0167-9473(01)00065-2

Friedman, J. H., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting. *The Annals of Statistics, 28*(2), 337-407.

Fox, J. (1980). Effects analysis in structural equation models: Extensions and simplified methods of computation. *Sociological Methods and Research, 9*, 3-28. doi: 10.1177/004912418000900101

Gelman, A. (2004). Parameterization and Bayesian modeling. *Journal of the American Statistical Association, 99*, 537-545. doi: 10.1198/016214504000000458

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). Boca Raton, FL: CRC Press.

Gelman, A., & Speed, T. P. (1993). Characterizing a joint probability distribution by

conditionals. *Journal of the Royal Statistical Society, B, 55*, 185-188. doi: 10.2307/2346074

Gentle, J. E. (1998). *Numerical linear algebra for applications in statistics*. Berlin: Spring-Verlag.

Geweke, J., Durham, G., & Hu, H. (preprint 2013). Bayesian inference for logistic regression models using sequential posterior simulation. *arXiv*:1304.4333

Goodnight, J. H. (1979). A tutorial on the sweep operator. *The American Statistician, 33*(3), 149-158.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology, 60*, 549-576. doi: 10.1146/annurev.psych.58.110405.085530

Graybill, F. A. (1976). *Theory and application of the linear model*. North Scituate, MA: Duxbury Press.

Groenewald, P. C. N., & Mokgatlhe, L. (2005). Bayesian computation for logistic regression. *Computational Statistics & Data Analysis, 48*, 857-868. doi: 10.1016/j.csda.2004.04.009

Harville, D. A. (1997). *Matrix algebra from a statistician's perspective*. New York, NY: Springer-Verlag.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *Elements of statistical learning* (2nd ed.). New York, NY: Springer.

Horton, N. J., & Kleinman, K. P. (2007). Much ado nothing: A comparison of missing data methods and software to fit incomplete data regression models. *The American Statistician, 61*, 79-90. doi: 10.1198/000313007X172556

Hughes, R. A., White, I. R., Seaman, S. R., Carpenter, J. R., Tilling, K., & Sterne, J. A. C. (2014). Joint modelling rationale for chained equations. *BMC Research Methodology, 14*, 1-10. doi: 10.1186/1471-2288-14-28

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: With applications in R*. New York, NY: Springer.

Jansen, I., Beunckens, C., Molenberghs, G., Verbeke, G., & Mallinckrodt, C. (2006). Analyzing incomplete discrete longitudinal clinical trial data. *Statistical Science, 21*, 52-69. doi: 10.1214/088342305000000322

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford, Englad: Claredon Press.

Johnson, R. A., & Wichern, D. W. (2002). *Applied multivariate statistical analysis* (6th ed.). Englewood Cliffs, NJ: Prentice Hall.

Karangwa, D. K. (2013). Using Markov Chain Monte Carlo method to make inferences on items of data contaminated by missing values. *American Journal of Theoretical and Applied Statistics, 2*, 48-53. doi: 10.11648/j.ajtas.20130203.12

Karangwa, I., Kotze, D., & Blignaut, R. (in press). Multiple imputation of unordered categorical missing data: A comparison of the multivariate normal imputation and multiple imputation by chained equations. *British Journal of Political Science*.

Kline, R. B. (2010). *Principles and practice of structural equation modeling* (3rd ed.). New York, NY: Guilford Press.

Koehler, E., Brown, E., & Haneuse, S. J. P. A. (2009). On the assessment of Monte Carlo error in simulation-based statistical analyses. *The American Statistician, 63*, 155-162. doi: 10.1198/tast.2009.0030

Koopman, J., Howe, M., Hollenbeck, J. R., & Sin, H. (2015). Small sample mediation

testing: Misplaced Confidence in bootstrapped confidence intervals. *Journal of Applied Psychology, 100*, 194-202. doi: 10.1037/a0036635

Kropko, J., Goodrich, B., Gelman, A., & Hill, J. (2014). Multiple imputation for continuous and categorical data: Comparing joint multivariate normal and conditional approaches. *Political Analysis, 22*, 497-519. doi: 10.1093/pan/mpu007

Lee, K. J., & Carlin, J. B. (2010). Multiple imputation for missing data: Fully conditional specification versus multivariate normal imputation. *American Journal of Epidemiology, 171*, 624-632. doi: 10.1093/aje/kwp425

Lee, S. (2007). *Structural equation modeling: A Bayesian approach*. West Sussex, England: John Wiley & Sons, Inc.

Lehmann, E. L. (1999). *Elements of large-sample theory.* New York, NY: Springer-Verlag.

Lehmann, E. L., & Casella, G. (1998). *Theory of point estimation* (2nd ed.). New York, NY: Springer-Verlag.

Li, J., Meng, X., & Rubin, D. B. (1991). Significance levels from repeated p-values with multiply-imputed data. *Statistica Sinica, 1*, 65-92

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Hoboken, NJ: John Wiley & Sons, Inc.

Liu, J., Gelman, A., Hill, J., Su, Y., & Kropko, J. (2014). On the stationary distribution of iterative imputations. *Biometrika, 101*, 155-173. doi: 10.1093/biomet/ast044

Lo, A. Y. (1987). A large sample study of the Bayesian bootstrap, *Annals of Statistics, 15*(1), 360-375.

MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis*. New York,

NY: Lawrence Erlbaum Associates.

MacKinnon, D. P., Lockwood, C. M., Brown, C. H., Wang, W., & Hoffman, J. M. (2007). The intermediate endpoint effect in logistic and probit regression. *Clinical Trials, 4*, 499-513. doi: 10.1177/1740774507083434

MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., & Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods, 7*, 83-104. doi:10.1037//1082-989X.7.1.83

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*, 99-128. doi: 10.1207/s15327906mbr3901_4

MacKinnon, D. P., Warsi, G., & Dwyer, J. H. (1995). A simulation study of mediated effect measures. *Multivariate Behavioral Research, 30*, 1-22. doi: 10.1207/s15327906mbr3001_3

Madsen, H., & Thyregod, P. (2011). *Introduction to general and generalized linear models*. Boca Raton, FL: Chapman & Hall/CRC.

Magnus, J. R., & Neudecker, H. (1999). *Matrix differential calculus with applications in statistics and econometrics* (revised ed.). New York, NY: John Wiley & Sons.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). London: Chapman & Hall/CRC.

Meng, X., & Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika, 80*, 267-278. doi: 10.1093/biomet/80.2.267

Mitchell, T. (1997). *Machine learning*. New York, NY: McGraw-Hill.

Muthén, B. & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods, 17*, 313-335. doi: 10.1037/a0026802.

Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics, 7*, 1-21. doi:10.3389/fnbot.2013.00021

Peterson, K. B., & Pedersen, M. S. (2012). *The matrix cookbook*. Copenhagen, Denmark: Technical University of Denmark.

Preacher, K. J., & Hayes, A. F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods, Instruments, & Computers*, *36*(4), 717-731.

Preacher, K. J., Rucker, D. D., & Hayes, A. F. (2007). Addressing moderated mediation hypotheses: Theory, methods and prescriptions. *Multivariate Behavioral Research, 42*, 185-227. doi: 10.1080/00273170701341316

Press, S. J. (1972). *Applied multivariate analysis*. New York, NY: Holt, Rinehart, and Winston.

Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J., & Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology, 27*(1), 88-95.

Rao, C. R. (1976). *Linear statistical inference and its applications* (2nd ed.). New York, NY: John Wiley & Sons.

Rosseel, Y. (2012). An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1-36.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*, 581-592. doi:

10.1093/biomet/63.3.581

Rubin, D. B. (1981). The Bayesian bootstrap. *The Annals of Statistics, 9*(1), 130-134.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York, NY: John Wiley & Sons.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association, 91*, 473-489. doi: 10.2307/2291635

Savalei, V. (2010). Expected versus observed information in SEM with incomplete normal and nonnormal data. *Psychological methods, 15*, 352-367. doi: 10.1037/a0020143

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. New York, NY: Chapman & Hall.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the start of the art. *Psychological Methods, 7*, 147-177. doi: 10.1037//1082-989X.7.2.147

Seaman, S. R., Bartlett, J. W., & White, I. R. (2012). Multiple imputation of missing covariates with non-linear effects and interactions: An evaluation of statistical methods. *BMC Medical Research Methodology, 12*, 1-13. doi: 10.1186/1471-2288-12-46

Searle, S. R. (1982). *Matrix algebra useful for statistics*. New York, NY: Wiley.

Shah, A. D., Bartlett, J. W., Carpenter, J., Nicolas, O., & Hemingway, H. (2014). Comparison of random forest and parametric imputation models for imputing missing data using MICE: A CALIBER study. *American Journal of Epidemiology, 179*, 764-774. doi: 10.1093/aje/kwt312

Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural

equation models. *Sociological Methodology, 13*, 290-312. doi:10.2307/270723.

Sobel, M. E. (1986). Some new results on indirect effects and their standard errors in covariance structure. *Sociological Methodology* , *16*, 159-186. doi: 10.2307/270922.

Song, X., & Lee, S. (2012). A tutorial on the Bayesian approach for analyzing structural equation models. *Journal of Mathematical Psychology, 56*, 135-148. doi: 10.1016/j.jmp.2012.02.001

Stekhoven, D. J., & Bühlmann, P. (2012). MissForest - nonparametric missing value imputation for mixed-type data. *Bioinformatics, 28*, 112-118. doi: 10.1093/bioinformatics/btr597

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods, 14*, 323-348. doi: 10.1037/a0016973.

Sutton, C. D. (2005). Classification and regression trees, bagging, and boosting. *Handbook of Statistics, 24*, 303-329. doi: 10.1016/S0169-7161(04)24011-1

Taddy, M., Chen, C., Yu, J., & Wyle, M. (preprint 2015). Bayesian and empirical Bayesian forests. *arXiv:1502.02312*

Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association, 82*, 528-540.

Timm, N. H. (2002). *Applied multivariate analysis*. New York, NY: Springer-Verlag.

Twala, B. E. T. H., Jones. M. C., & Hand, D. J. (2008). Good methods for coping with missing data in decision trees. *Pattern Recognition Letters, 29*(7), 950-956.

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. (2012). *Flexible Imputation of Missing Data*. Boca Raton, FL: Chapman & Hall/CRC.

van Buuren, S., Brand, J. P. L., Groothuis-Oudshoorn, K., & Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation, 76*(12), 1049-1064

van Buuren, S, & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software, 45*, 1-67. doi: 10.18637/jss.v045.i03

von Hippel, P. T. (2009). How to impute squares, interactions, and other transformed variables. *Sociological Methodology, 39*, 265-291.doi: 10.1111/j.1467-9531.2009.01215.x

Wand, M. P. (2002). Vector differential calculus in statistics. *The American Statistician, 56*, 55-62. doi: 10.1198/000313002753631376

Wang, C. Y., & Feng, Z. (2010). Boosting with missing predictors. *Biostatistics, 11*, 195-212. doi: 10.1093/biostatistics/kxp052

Wang, L., & Preacher, K. J. (2015). Moderated mediation analysis using Bayesian Methods. *Structural Equation Modeling: A Multidisciplinary Journal, 22*, 249-263. doi: 10.1080/10705511.2014.935256

Wang, N. & Robins, J. M. (1998). Large-sample theory for parametric multiple imputation procedures. *Biometrika, 85*, 935-948. doi: 10.1093/biomet/85.4.935

Wu, W., & Fia, F. (2013). A new procedure to test mediation with missing data

through nonparametric bootstrapping and multiple imputation. *Multivariate Behavioral Research, 48*, 663-691. doi: 10.1080/00273171.2013.816235

Yuan, K. H. (2009). Normal distribution based pseudo ML for missing data: With applications to mean and covariance structure analysis. *Journal of Multivariate Analysis, 100*, 1900-1918. doi: 10.1016/j.jmva.2009.05.001

Yuan, K. H., & Bentler, P. M. (2000). Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. *Sociological Methodology, 30*, 167-202. doi: 10.1111/0081-1750.00078

Yuan, Y., & MacKinnon, D. P. (2009). Bayesian mediation analysis. *Psychological Methods, 14*, 301-322. doi:10.1037/a0016972

Zhang, P. (2013). Multiple imputation: Theory and method. *International Statistical Review, 71*, 581-592. doi: 10.1111/j.1751-5823.2003.tb00213.x

APPENDIX A

PROOFS

Proof of Proposition 1.1.

To find the first two moments of U, we recall from elementary calculus the identity given by the derivative of a univariate logarithmic function,

$$\frac{\partial}{\partial\theta}\log\big(f(y|\theta,\phi)\big) = \frac{1}{f(y|\theta,\phi)}\frac{\partial}{\partial\theta}f(y|\theta,\phi), \tag{A.1}$$

where log denotes the natural logarithm, or base e. Taking the expectations of both sides of (A.1), $E(U)$ is expressed as

$$\begin{aligned} E(U) &= \int \frac{\partial}{\partial\theta}\log\big(f(y|\theta,\phi)\big)\, f(y|\theta,\phi)dy \\ &= \int \frac{\partial}{\partial\theta} f(y|\theta,\phi)\, dy. \end{aligned} \tag{A.2}$$

Assuming certain regularity conditions hold (see Casella & Berger, 2001 p. 516), the integral on the right-hand side of (A.2) can be calculated first

$$\begin{aligned} \int \frac{\partial}{\partial\theta} f(y|\theta,\phi)\, dy &= \frac{\partial}{\partial\theta}\underbrace{\int f(y|\theta,\phi)\, dy}_{\text{Proper density}} \\ &= \frac{\partial}{\partial\theta}(1) \\ &= 0. \end{aligned}$$

Therefore, $E(U) = 0$. The variance of U, also called the information, can be calculated using the general variance equation

$$Var(U) = E(U^2) - [E(U)]^2. \tag{A.3}$$

Since, $E(U) = 0, [E(U)]^2 = 0$ (again assuming regularity conditions hold). To calculate $E(U^2)$, we differentiate both sides of (A.2)

$$\begin{aligned} \frac{\partial}{\partial\theta}E(U) &= \frac{\partial}{\partial\theta}\left[\int \frac{\partial}{\partial\theta}\log\big(f(y|\theta,\phi)\big)\, f(y|\theta,\phi)dy\right] \\ &= \frac{\partial}{\partial\theta}\left[\int \frac{\partial}{\partial\theta} f(y|\theta,\phi)\, dy\right]. \end{aligned} \tag{A.4}$$

Regularity conditions allow the order of differentiation and integration to be interchanged, so the right-hand size of (A.4) becomes

$$\begin{aligned} &\int \frac{\partial}{\partial\theta}\left[\frac{\partial}{\partial\theta}\log\big(f(y|\theta,\phi)\big)\, f(y|\theta,\phi)\right] dy \\ &= \int \left\{\frac{\partial^2}{\partial\theta^2}\log\big(f(y|\theta,\phi)\big)\, f(y|\theta,\phi)\right. \\ &\left. + \frac{\partial}{\partial\theta}\log\big(f(y|\theta,\phi)\big)\frac{\partial}{\partial\theta} f(y|\theta,\phi)\right\} dy. \end{aligned} \tag{A.5}$$

Using the identity

$$\frac{\partial}{\partial\theta} f(y|\theta,\phi) = f(y|\theta,\phi)\frac{\partial}{\partial\theta}\log\big(f(y|\theta,\phi)\big),$$

The second term of the right-hand side of (A.5) simplifies to

$$\frac{\partial}{\partial\theta}\log\big(f(y|\theta,\phi)\big)\, f(y|\theta,\phi)\frac{\partial}{\partial\theta}\log\big(f(y|\theta,\phi)\big) = \left(\frac{\partial}{\partial\theta}\log\big(f(y|\theta,\phi)\big)\right)^2 f(y|\theta,\phi).$$

Hence, (A.5) becomes

$$\begin{aligned} &\int \frac{\partial^2}{\partial\theta^2}\log\big(f(y|\theta,\phi)\big)\, f(y|\theta,\phi)\, dy \\ &+ \int \left(\frac{\partial}{\partial\theta}\log\big(f(y|\theta,\phi)\big)\right)^2 f(y|\theta,\phi)\, dy = 0, \end{aligned} \tag{A.6}$$

or to simplify notation,

$$E\left[\frac{\partial^2}{\partial\theta^2}\log\big(f(y|\theta,\phi)\big)\right]+E\left[\left(\frac{\partial}{\partial\theta}\log\big(f(y|\theta,\phi)\big)\right)^2\right]=0 \qquad \text{(A.7)}$$

Rewriting (A.6) or (A.7) in terms of the score function,

$$E(U')+E(U^2)=0$$

where U' denotes the partial derivative of U with respect to θ. From (A.3), we see that

$$Var(U)=E(U^2)-[E(U)]^2$$

$$=E(U^2)$$

or equivalently

$$Var(U)=-E(U'),$$

which completes the proof. ■

Proof of Lemma 2.1.

Recall, $\boldsymbol{\Sigma}$ is defined as

$$\boldsymbol{\Sigma}=\begin{bmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}\\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{bmatrix}. \qquad \text{(A.8)}$$

To prove this Lemma with respect to the determinant of $\boldsymbol{\Sigma}$, multiply the second matrix column of $\boldsymbol{\Sigma}$ by $\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$ and subtract this result from the first matrix column,

$$\left|\begin{bmatrix}\boldsymbol{\Sigma}_{11}-\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{12}\\ \mathbf{0} & \boldsymbol{\Sigma}_{22}\end{bmatrix}\right|=|\boldsymbol{\Sigma}_{22}||\boldsymbol{\Sigma}_{11}-\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}|.$$

The result is similar for the other determinant. With regards to the inverse of $\boldsymbol{\Sigma}$, given that $\boldsymbol{\Sigma}^{-1}$ is the inverse of $\boldsymbol{\Sigma}$, $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}$ (or equivalently $\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}$) is a partitioned identity matrix,

$$\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}=\begin{bmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}\\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{bmatrix}\begin{bmatrix}\boldsymbol{\Sigma}^{11} & \boldsymbol{\Sigma}^{12}\\ \boldsymbol{\Sigma}^{21} & \boldsymbol{\Sigma}^{22}\end{bmatrix}$$

$$=\begin{bmatrix}\boldsymbol{\Sigma}_{11}\boldsymbol{\Sigma}^{11}+\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}^{21} & \boldsymbol{\Sigma}_{11}\boldsymbol{\Sigma}^{12}+\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}^{22}\\ \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}^{11}+\boldsymbol{\Sigma}_{22}\boldsymbol{\Sigma}^{21} & \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}^{12}+\boldsymbol{\Sigma}_{22}\boldsymbol{\Sigma}^{22}\end{bmatrix}$$

$$= \begin{bmatrix} \mathbf{I}_{p_1} & \mathbf{0} \\ \mathbf{0}' & \mathbf{I}_{p_2} \end{bmatrix},$$

where $\mathbf{I}_{p_1}$ is a $p_1 \times p_1$ identity matrix, $\mathbf{I}_{p_2}$ is a $p_2 \times p_2$ identity matrix, $\mathbf{0}$ is a $p_1 \times p_2$ zero matrix, and $\mathbf{0}'$ is a $p_2 \times p_1$ zero matrix. Solving for $\boldsymbol{\Sigma}_{11}$ and $\boldsymbol{\Sigma}_{22}$

$$\boldsymbol{\Sigma}_{11}\boldsymbol{\Sigma}^{11} + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}^{21} = \mathbf{I}_{p_1}$$

$$\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{11}\boldsymbol{\Sigma}^{11} + \boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}^{21} = \boldsymbol{\Sigma}_{11}^{-1}$$

$$\boldsymbol{\Sigma}^{11} = \boldsymbol{\Sigma}_{11}^{-1} - \boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}^{21} \qquad \text{(A.9)}$$

and

$$\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}^{12} + \boldsymbol{\Sigma}_{22}\boldsymbol{\Sigma}^{22} = \mathbf{I}_{p_2}$$

$$\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}^{12} + \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{22}\boldsymbol{\Sigma}^{22} = \boldsymbol{\Sigma}_{22}^{-1}$$

$$\boldsymbol{\Sigma}^{22} = \boldsymbol{\Sigma}_{22}^{-1} - \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}^{12} \qquad \text{(A.10)}$$

Now, solving for $\boldsymbol{\Sigma}^{12}$ and $\boldsymbol{\Sigma}^{21}$

$$\boldsymbol{\Sigma}_{11}\boldsymbol{\Sigma}^{12} + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}^{22} = \mathbf{0}$$

$$\boldsymbol{\Sigma}^{12} = -\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}^{22} \qquad \text{(A.11)}$$

and

$$\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}^{11} + \boldsymbol{\Sigma}_{22}\boldsymbol{\Sigma}^{21} = \mathbf{0}$$

$$\boldsymbol{\Sigma}^{21} = -\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}^{11}. \qquad \text{(A.12)}$$

Substituting (A.12) into (A.9),

$$\boldsymbol{\Sigma}^{11} = \boldsymbol{\Sigma}_{11}^{-1} + \boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}^{11}$$

$$(\mathbf{I} - \boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21})\boldsymbol{\Sigma}^{11} = \boldsymbol{\Sigma}_{11}^{-1}$$

$$(\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21})\boldsymbol{\Sigma}^{11} = \mathbf{I}$$

$$\boldsymbol{\Sigma}^{11} = (\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21})^{-1}.$$

Similarly, substituting (A.11) into (A.10),

$$\boldsymbol{\Sigma}^{22} = \boldsymbol{\Sigma}_{22}^{-1} + \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}^{22}$$

$$(\mathbf{I} - \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})\boldsymbol{\Sigma}^{22} = \boldsymbol{\Sigma}_{22}^{-1}$$

$$(\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})\boldsymbol{\Sigma}^{22} = \mathbf{I}$$

$$\boldsymbol{\Sigma}^{22} = (\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})^{-1}.$$

Finally, solving for $\boldsymbol{\Sigma}^{12}$ and $\boldsymbol{\Sigma}^{21}$,

$$\boldsymbol{\Sigma}^{12} = -\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}(\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})^{-1},$$

and

$$\boldsymbol{\Sigma}^{21} = -\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}(\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21})^{-1}.$$

This completes the proof. ■

Proof of Lemma 2.3

The joint distribution of $\mathbf{x}_1$ and $\mathbf{x}_2$ is given by

$$f(\mathbf{x}_1, \mathbf{x}_2) = (2\pi)^{-p/2} \begin{vmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{vmatrix}^{-1/2}$$
$$\times \exp\left(-\frac{1}{2} \left(\begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} - \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix} \right)' \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}^{-1} \left(\begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} - \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix} \right) \right). \quad \text{(A.13)}$$

Rewriting $\boldsymbol{\Sigma}^{-1}$ as

$$\begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}^{-1} = \begin{bmatrix} \boldsymbol{\Sigma}^{11} & \boldsymbol{\Sigma}^{12} \\ \boldsymbol{\Sigma}^{21} & \boldsymbol{\Sigma}^{22} \end{bmatrix}$$

the quadratic form in (A.13) can be expanded as

$$Q(\mathbf{x}_1, \mathbf{x}_2) = (\mathbf{x}_1 - \boldsymbol{\mu}_1)'\boldsymbol{\Sigma}^{11}(\mathbf{x}_1 - \boldsymbol{\mu}_1)$$
$$+ 2(\mathbf{x}_1 - \boldsymbol{\mu}_1)'\boldsymbol{\Sigma}^{12}(\mathbf{x}_2 - \boldsymbol{\mu}_2) \quad \text{(A.14)}$$
$$+ (\mathbf{x}_2 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{22}(\mathbf{x}_2 - \boldsymbol{\mu}_2).$$

$Q(\mathbf{x}_1, \mathbf{x}_2)$ in (A.14) can be further factored as

$$Q(\mathbf{x}_1, \mathbf{x}_2) = Q_1(\mathbf{x}_1) + Q_2(\mathbf{x}_1, \mathbf{x}_2),$$

where $Q_1(\mathbf{x}_1)$ does not depend on $\mathbf{x}_2$ and $Q_2(\mathbf{x}_1, \mathbf{x}_2)$ contains terms that either depend solely on $\mathbf{x}_2$ or jointly on $\mathbf{x}_1$ and $\mathbf{x}_2$. Following this factoring,

$$Q_1(\mathbf{x}_1) = (\mathbf{x}_1 - \boldsymbol{\mu}_1)'\boldsymbol{\Sigma}^{11}(\mathbf{x}_1 - \boldsymbol{\mu}_1) + \boldsymbol{\mu}_2'\boldsymbol{\Sigma}^{22}\boldsymbol{\mu}_2 - 2(\mathbf{x}_1 - \boldsymbol{\mu}_1)'\boldsymbol{\Sigma}^{12}\boldsymbol{\mu}_2$$

and

$$Q_2(\mathbf{x}_1, \mathbf{x}_2) = \mathbf{x}_2'\boldsymbol{\Sigma}^{22}\mathbf{x}_2 - 2(\boldsymbol{\mu}_2'\boldsymbol{\Sigma}^{22}\mathbf{x}_2 + \boldsymbol{\mu}_1'\boldsymbol{\Sigma}^{12}\mathbf{x}_2 - \mathbf{x}_1'\boldsymbol{\Sigma}^{12}\mathbf{x}_2)$$

If $f(\mathbf{x}_1)$ denotes the marginal distribution of $\mathbf{x}_1$, then

$$\begin{aligned} f(\mathbf{x}_1) &= \int f(\mathbf{x}_1, \mathbf{x}_2)\mathrm{d}\mathbf{x}_2 \\ &= (2\pi)^{-p/2}|\boldsymbol{\Sigma}|^{-1/2}\exp\left\{-\frac{1}{2}Q_1(\mathbf{x}_1)\right\}\int \exp\left\{-\frac{1}{2}Q_2(\mathbf{x}_1, \mathbf{x}_2)d\mathbf{x}_2\right\} \end{aligned} \tag{A.15}$$

To evaluate the integral on the right-hand side of (A.15), complete the square on $\mathbf{x}_2$. Let

$$\boldsymbol{\lambda} = \boldsymbol{\mu}_2 - (\boldsymbol{\Sigma}^{22})^{-1}\boldsymbol{\Sigma}^{21}(\mathbf{x}_1 - \boldsymbol{\mu}_1),$$

then $\int \exp\left\{-\frac{1}{2}[(\mathbf{x}_2 - \boldsymbol{\lambda})'\boldsymbol{\Sigma}^{22}(\mathbf{x}_2 - \boldsymbol{\lambda}) - \boldsymbol{\lambda}'\boldsymbol{\Sigma}^{22}\boldsymbol{\lambda}]d\mathbf{x}_2\right\}$ can be evaluated as

$$\exp\left\{-\frac{1}{2}\boldsymbol{\lambda}'\boldsymbol{\Sigma}^{22}\boldsymbol{\lambda}\right\}\int \exp\left\{-\frac{1}{2}[(\mathbf{x}_2 - \boldsymbol{\lambda})'\boldsymbol{\Sigma}^{22}(\mathbf{x}_2 - \boldsymbol{\lambda})]d\mathbf{x}_2\right\} = |\boldsymbol{\Sigma}^{22}|^{-1/2}\exp\left\{-\frac{1}{2}\boldsymbol{\lambda}'\boldsymbol{\Sigma}^{22}\boldsymbol{\lambda}\right\}.$$

Combing the results, we obtain the marginal distribution of $\mathbf{x}_1$

$$\begin{aligned} f(\mathbf{x}_1) &= (2\pi)^{-p/2}|\boldsymbol{\Sigma}|^{-1/2}|\boldsymbol{\Sigma}^{22}|^{-1/2}\exp\left\{-\frac{1}{2}Q_1(\mathbf{x}_1)\right\}\exp\left\{-\frac{1}{2}\boldsymbol{\lambda}'\boldsymbol{\Sigma}^{22}\boldsymbol{\lambda}\right\} \\ &= (2\pi)^{-p/2}|\boldsymbol{\Sigma}|^{-1/2}|\boldsymbol{\Sigma}^{22}|^{-1/2}\exp\left\{-\frac{1}{2}(\mathbf{x}_1 - \boldsymbol{\mu}_1)'\boldsymbol{\Sigma}_{\mathbf{11}}^{-1}(\mathbf{x}_1 - \boldsymbol{\mu}_1)\right\} \\ &= (2\pi)^{-p/2}|\boldsymbol{\Sigma}_{11}|^{-1/2}\exp\left\{-\frac{1}{2}(\mathbf{x}_1 - \boldsymbol{\mu}_1)'\boldsymbol{\Sigma}_{\mathbf{11}}^{-1}(\mathbf{x}_1 - \boldsymbol{\mu}_1)\right\}, \end{aligned}$$

which is a multivariate normal distribution with mean $\boldsymbol{\mu}_1$ and covariance matrix $\boldsymbol{\Sigma}_{11}$. This completes the proof. ■

Proof of Theorem 2.1.

Expanding the quadratic form, $Q(\mathbf{x}_1, \mathbf{x}_1)$, in the joint distribution of $\mathbf{x}_1$ and $\mathbf{x}_2$

$$\begin{aligned} Q(\mathbf{x}_1, \mathbf{x}_2) = &(\mathbf{x}_1 - \boldsymbol{\mu}_1)'\boldsymbol{\Sigma}^{11}(\mathbf{x}_1 - \boldsymbol{\mu}_1) \\ &+ 2(\mathbf{x}_1 - \boldsymbol{\mu}_1)'\boldsymbol{\Sigma}^{12}(\mathbf{x}_2 - \boldsymbol{\mu}_2) \\ &+ (\mathbf{x}_2 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{22}(\mathbf{x}_2 - \boldsymbol{\mu}_2), \end{aligned}$$

where $\boldsymbol{\Sigma}^{11}$ is

$$\begin{aligned} \boldsymbol{\Sigma}^{11} &= (\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21})^{-1} \\ &= \boldsymbol{\Sigma}_{11}^{-1} + \boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}(\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}, \end{aligned}$$

$\boldsymbol{\Sigma}^{22}$ is

$$\boldsymbol{\Sigma}^{22} = (\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})^{-1}$$

$\boldsymbol{\Sigma}^{12}$ is

$$\boldsymbol{\Sigma}^{12} = -\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}(\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})^{-1},$$

and $\boldsymbol{\Sigma}^{21}$ is

$$\boldsymbol{\Sigma}^{21} = -\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}(\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21})^{-1}.$$

Substituting these expressions into $Q(\mathbf{x}_1, \mathbf{x}_1)$,

$$\begin{aligned} Q(\mathbf{x}_1, \mathbf{x}_1) = &(\mathbf{x}_1 - \boldsymbol{\mu}_1)'[\boldsymbol{\Sigma}_{11}^{-1} + \boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}(\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}](\mathbf{x}_1 - \boldsymbol{\mu}_1) \\ &- 2(\mathbf{x}_1 - \boldsymbol{\mu}_1)'[\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}(\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})^{-1}](\mathbf{x}_1 - \boldsymbol{\mu}_1) \\ &+ (\mathbf{x}_2 - \boldsymbol{\mu}_2)'[(\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})^{-1}](\mathbf{x}_2 - \boldsymbol{\mu}_2). \end{aligned}$$

Rearranging terms,

$$Q(\mathbf{x}_1, \mathbf{x}_1) = (\mathbf{x}_1 - \boldsymbol{\mu}_1)'\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{x}_1 - \boldsymbol{\mu}_1)$$
$$+ (\mathbf{x}_1 - \boldsymbol{\mu}_1)'\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}(\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{x}_1 - \boldsymbol{\mu}_1)$$
$$- 2(\mathbf{x}_1 - \boldsymbol{\mu}_1)'[\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}(\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})^{-1}](\mathbf{x}_1 - \boldsymbol{\mu}_1)$$
$$+ (\mathbf{x}_2 - \boldsymbol{\mu}_2)'[(\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})^{-1}](\mathbf{x}_2 - \boldsymbol{\mu}_2).$$

Now, on the right-hand side of $Q(\mathbf{x}_1, \mathbf{x}_1)$ we recognize that the last three terms are the expansion of a matrix quadratic form. Therefore, we can rewrite $Q(\mathbf{x}_1, \mathbf{x}_1)$ as,

$$Q(\mathbf{x}_1, \mathbf{x}_2) = (\mathbf{x}_1 - \boldsymbol{\mu}_1)'\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{x}_1 - \boldsymbol{\mu}_1) + (\mathbf{u} - \mathbf{v})'\mathbf{A}(\mathbf{u} - \mathbf{v}),$$

where

$$\mathbf{u} = (\mathbf{x}_2 - \boldsymbol{\mu}_2)$$
$$\mathbf{v} = \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{x}_1 - \boldsymbol{\mu}_1)$$
$$\mathbf{A} = (\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})^{-1}.$$

To continue with simplification, define

$$\boldsymbol{\lambda} = \boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{x}_1 - \boldsymbol{\mu}_1).$$

and

$$\boldsymbol{\Omega} = \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}.$$

Substituting terms back into the joint distribution,

$$f(\mathbf{x}_1, \mathbf{x}_2) = \frac{1}{(2\pi)^{p/2}\begin{vmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{vmatrix}^{1/2}} \exp\left(-\frac{1}{2}Q(\mathbf{x}_1, \mathbf{x}_2)\right)$$

$$= \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}_{11}|^{1/2}|\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}|^{1/2}} \exp\left(-\frac{1}{2}Q(\mathbf{x}_1, \mathbf{x}_2)\right)$$

$$= \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}_{11}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}_1 - \boldsymbol{\mu}_1)'\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{x}_1 - \boldsymbol{\mu}_1)\right)$$

$$\times \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Omega}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}_2 - \boldsymbol{\lambda})'\boldsymbol{\Omega}^{-1}(\mathbf{x}_2 - \boldsymbol{\lambda})\right).$$

Importantly, here we can recognize that the first term in the product of is the marginal distribution of $\mathbf{x}_1$. Now, using the definition of conditional probability, we can derive the conditional distribution of $\mathbf{x}_2$ given $\mathbf{x}_1$,

$$f(\mathbf{x}_2|\mathbf{x}_1) = \frac{f(\mathbf{x}_2, \mathbf{x}_1)}{f(\mathbf{x}_1)}$$

$$= \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Omega}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}_2 - \boldsymbol{\lambda})'\boldsymbol{\Omega}^{-1}(\mathbf{x}_2 - \boldsymbol{\lambda})\right)$$

with mean vector $\boldsymbol{\lambda}$ and covariance matrix $\boldsymbol{\Omega}$. This completes the proof. ■

APPENDIX B

VARIANCE ESTIMATES USING MULTIVARIATE DELTA METHOD

For the purpose of the present study, only the variance estimates of the simple mediation model presented in Figure (1.3) and the moderated mediation model presented in Figure (1.4) (Model 5) will be derived. From Equation (1.53), the variance of an indirect-type effect can be approximated using the multivariate delta method given by

$$Var\left[g\left(\widehat{\boldsymbol{\theta}}_n\right) - g(\boldsymbol{\theta})\right] = \underbrace{\mathbf{d}'\widehat{\boldsymbol{\Sigma}}\left(\widehat{\boldsymbol{\theta}}\right)\mathbf{d}}_{\text{first-order}} + \underbrace{\frac{1}{2}\text{tr}\left[\left(\mathbf{H}\widehat{\boldsymbol{\Sigma}}\left(\widehat{\boldsymbol{\theta}}\right)\right)^2\right]}_{\text{second-order}}. \tag{B.1}$$

Recall, for a simple mediation model, the (unconditional) indirect effect is given by

$$g(\boldsymbol{\beta}|\mathbf{v}) = \beta_{M\cdot X}\beta_{Y\cdot M}, \tag{B.2}$$

Here, $\boldsymbol{\theta} = [\beta_{M\cdot X}, \beta_{Y\cdot M}]'$, $\widehat{\boldsymbol{\theta}} = \left[\hat{\beta}_{M\cdot X}, \hat{\beta}_{Y\cdot M}\right]'$, and

$$\widehat{\boldsymbol{\Sigma}}\left(\widehat{\boldsymbol{\theta}}\right) = \begin{bmatrix} Var\left(\hat{\beta}_{M\cdot X}\right) & 0 \\ 0 & Var\left(\hat{\beta}_{Y\cdot M}\right) \end{bmatrix}.$$

The gradient vector $\mathbf{d}$ of (B.2) given by

$$\mathbf{d} = \begin{bmatrix} \beta_{Y\cdot M} \\ \beta_{M\cdot X} \end{bmatrix}$$

and the Hessian matrix $\mathbf{H}$ of (B.2) is given by

$$\mathbf{H} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

Therefore, applying (B.1), the variance estimate under the multivariate delta method for a simple mediation model is

$$Var\left[g\left(\widehat{\boldsymbol{\beta}}|\mathbf{v}\right)\right] = \underbrace{\hat{\beta}_{Y\cdot M}^2 Var\left(\hat{\beta}_{M\cdot X}\right) + \beta_{M\cdot X}^2 Var\left(\hat{\beta}_{Y\cdot M}\right)}_{\text{first-order}} + \underbrace{\hat{\beta}_{Y\cdot M}^2\beta_{M\cdot X}^2}_{\text{second-order}}.$$

Similarly, for the moderated mediation model, the (conditional) indirect effect is given by

$$g(\boldsymbol{\beta}|\mathbf{v}) = (\beta_{M\cdot X} + \beta_{M\cdot XW}W)(\beta_{Y\cdot M} + \beta_{Y\cdot MW}W). \quad \text{(B.3)}$$

In (B.3), $\boldsymbol{\theta} = [\beta_{M\cdot X}, \beta_{M\cdot XW}, \beta_{Y\cdot M}, \beta_{Y\cdot MW}]'$, $\widehat{\boldsymbol{\theta}} = \left[\hat{\beta}_{M\cdot X}, \hat{\beta}_{M\cdot XW}, \hat{\beta}_{Y\cdot M}, \hat{\beta}_{Y\cdot MW}\right]'$, and

$$\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\theta}}) = \begin{bmatrix} Var(\hat{\beta}_{M\cdot X}) & Cov(\hat{\beta}_{M\cdot X}, \hat{\beta}_{M\cdot XW}) & 0 & 0 \\ Cov(\hat{\beta}_{M\cdot X}, \hat{\beta}_{M\cdot XW}) & Var(\hat{\beta}_{M\cdot XW}) & 0 & 0 \\ 0 & 0 & Var(\hat{\beta}_{Y\cdot M}) & Cov(\hat{\beta}_{Y\cdot M}, \hat{\beta}_{Y\cdot MW}) \\ 0 & 0 & Cov(\hat{\beta}_{Y\cdot M}, \hat{\beta}_{Y\cdot MW}) & Var(\hat{\beta}_{Y\cdot MW}) \end{bmatrix}$$

The gradient vector $\mathbf{d}$ of (B.3) given by

$$\mathbf{d} = \begin{bmatrix} \beta_{Y\cdot M} + \beta_{Y\cdot MW}W \\ \beta_{Y\cdot M}W + \beta_{Y\cdot MW}W^2 \\ \beta_{M\cdot X} + \beta_{M\cdot XW}W \\ \beta_{M\cdot X}W + \beta_{M\cdot XW}W^2 \end{bmatrix}$$

and the Hessian matrix $\mathbf{H}$ of (B.3) is given by

$$\mathbf{H} = \begin{bmatrix} 0 & 0 & 1 & W \\ 0 & 0 & W & W^2 \\ 1 & W & 0 & 0 \\ W & W^2 & 0 & 0 \end{bmatrix}.$$

Applying (B.1) and simplifying, the variance estimate under the multivariate delta method for the moderated mediation model is

$$Var[g(\widehat{\boldsymbol{\beta}}|\mathbf{v})] = \underbrace{\begin{matrix} (\hat{\beta}_{Y\cdot M} + \hat{\beta}_{Y\cdot MW})^2(Var(\hat{\beta}_{M\cdot X}) + 2Cov(\hat{\beta}_{M\cdot X}, \hat{\beta}_{M\cdot XW})W + Var(\hat{\beta}_{M\cdot XW})W^2) \\ +(\hat{\beta}_{M\cdot X} + \hat{\beta}_{M\cdot XW})^2(Var(\hat{\beta}_{Y\cdot M}) + 2Cov(\hat{\beta}_{Y\cdot M}, \hat{\beta}_{Y\cdot MW})W + Var(\hat{\beta}_{Y\cdot MW})W^2) \end{matrix}}_{\text{first-order}}$$

$$\underbrace{\begin{matrix} +(Var(\hat{\beta}_{M\cdot X}) + 2Cov(\hat{\beta}_{M\cdot X}, \hat{\beta}_{M\cdot XW})W + Var(\hat{\beta}_{M\cdot XW})W^2) \\ \times (Var(\hat{\beta}_{Y\cdot M}) + 2Cov(\hat{\beta}_{Y\cdot M}, \hat{\beta}_{Y\cdot MW})W + Var(\hat{\beta}_{Y\cdot MW})W^2) \end{matrix}}_{\text{second-order}}.$$

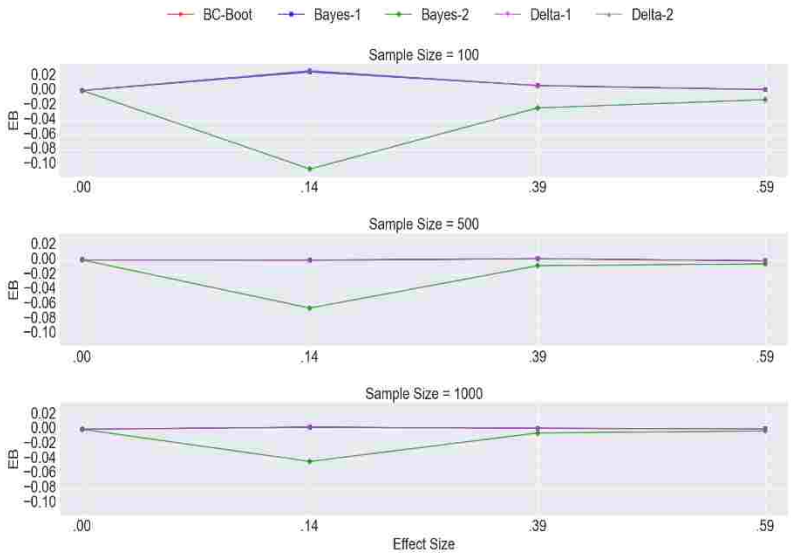
APPENDIX C

SUPPLEMENTAL FIGURES FROM SIMULATION 1

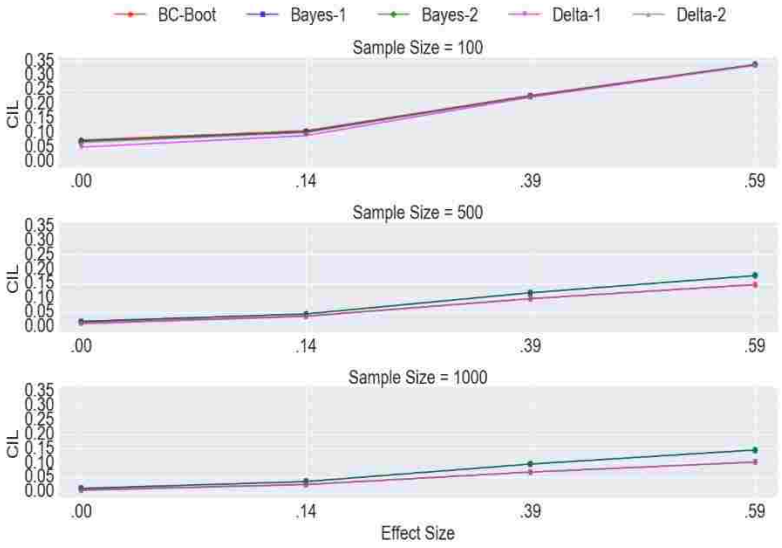
This Appendix presents the supplemental figures from Simulation Study 1 described in Chapter 4. For each of the models below, the acronyms for the metrics are: EB = empirical bias, CIL = confidence interval length, CP = coverage probability, RR = rejection rate, and MSE = mean squared error. Similarly, the methods are abbreviated as: BC-Boot = bias-corrected bootstrap, Bayes-1 = Bayesian bootstrap with mean estimator, Bayes-2 = Bayesian bootstrap with median estimator, Delta-1 = first-order delta method, and Delta-2 = second-order delta method.

1. Model: Mediation, Mediator: Continuous, Endogenous: Continuous

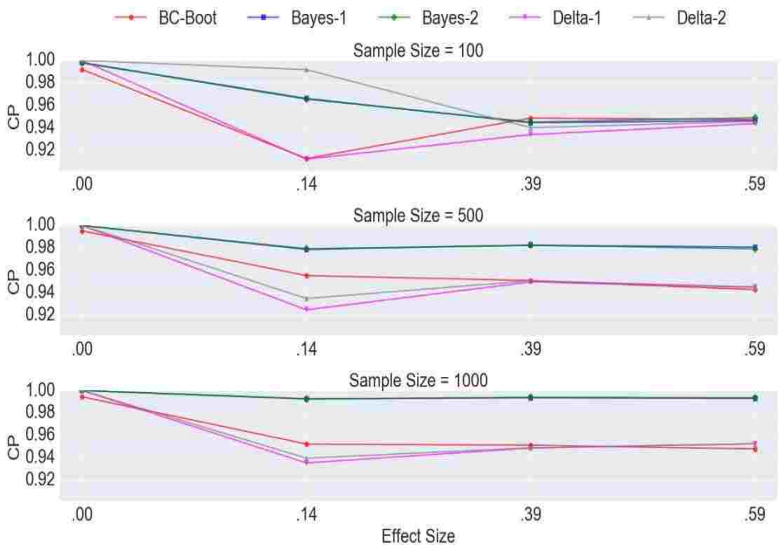
Empirical bias:



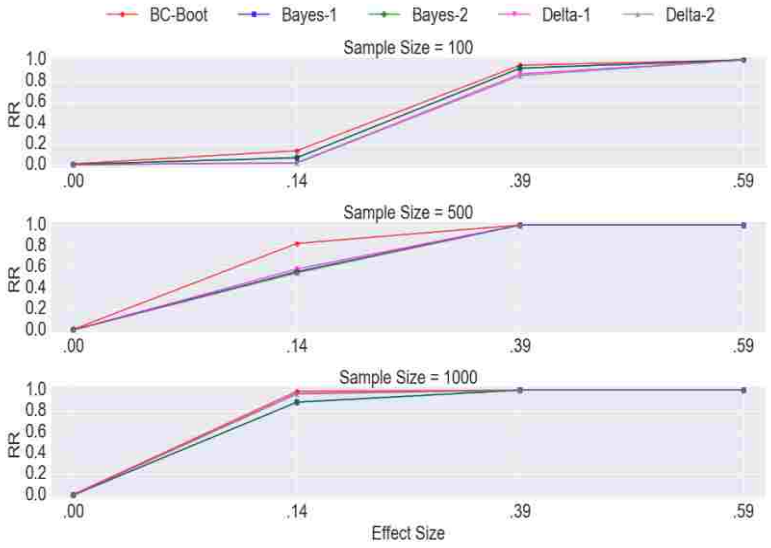
Confidence interval length:



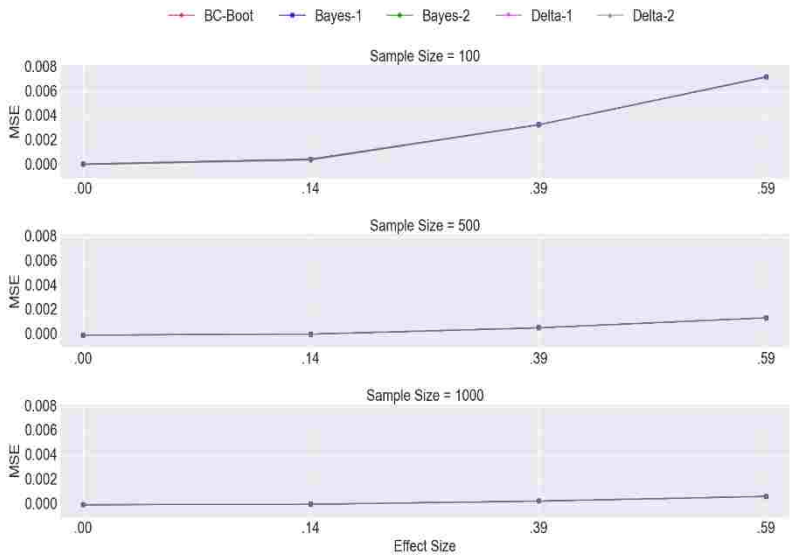
Coverage probability:



Rejection rate:

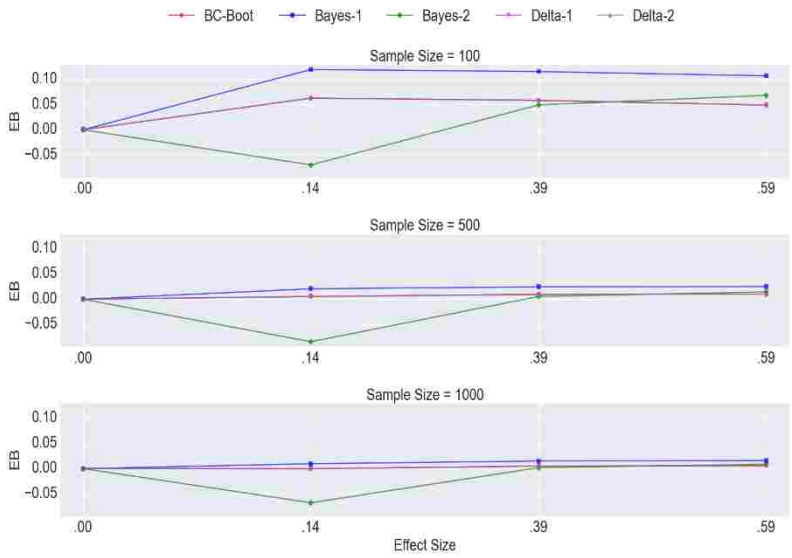


Mean squared error:

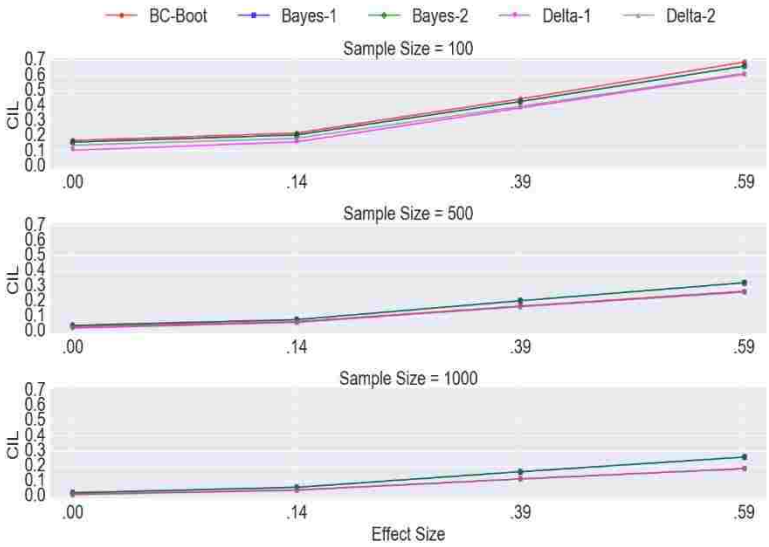


2. Model: Mediation, Mediator: Continuous, Endogenous: Categorical

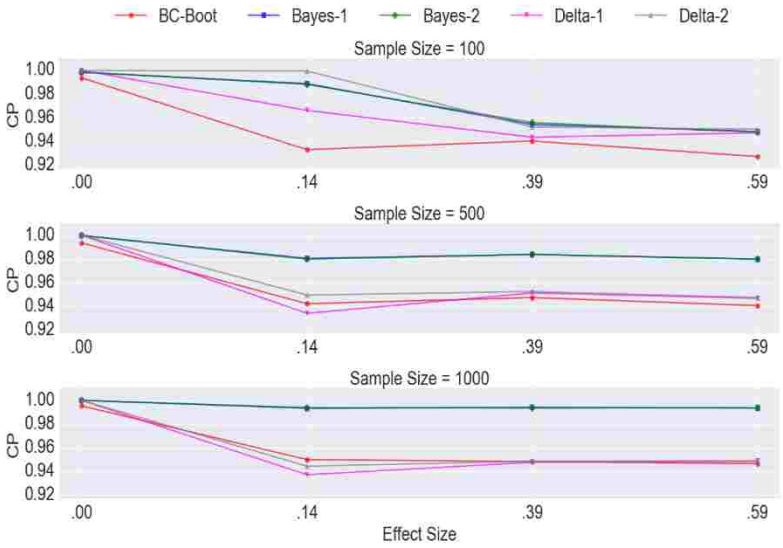
Empirical bias:



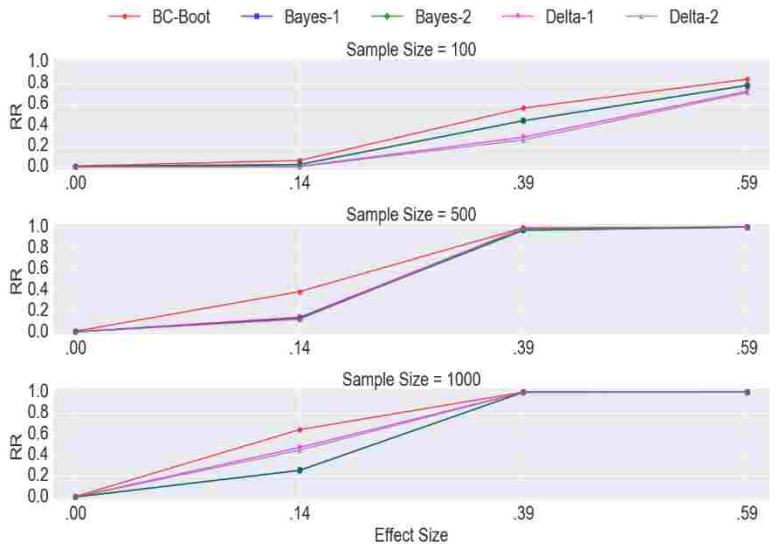
Confidence interval length:



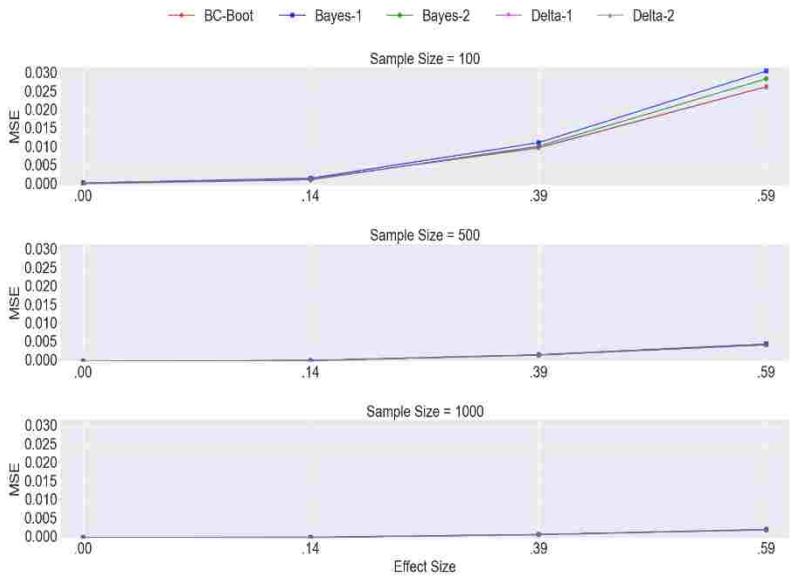
Coverage probability:



Rejection rate:

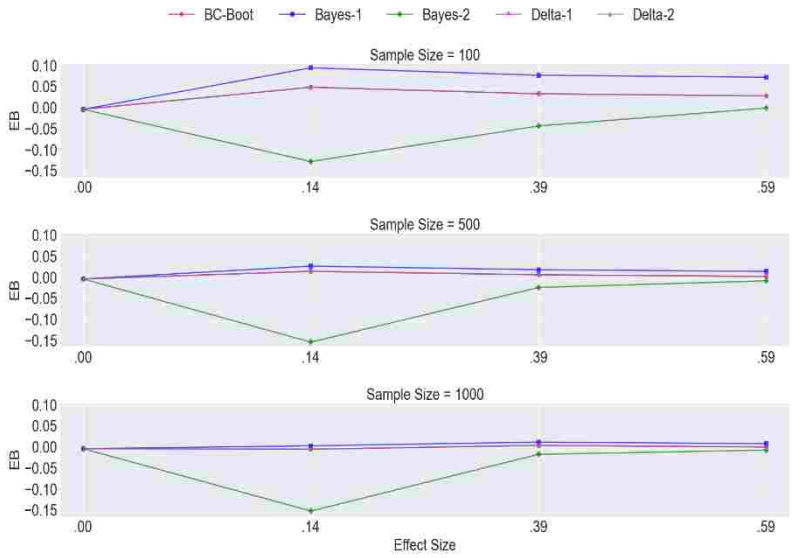


Mean squared error:

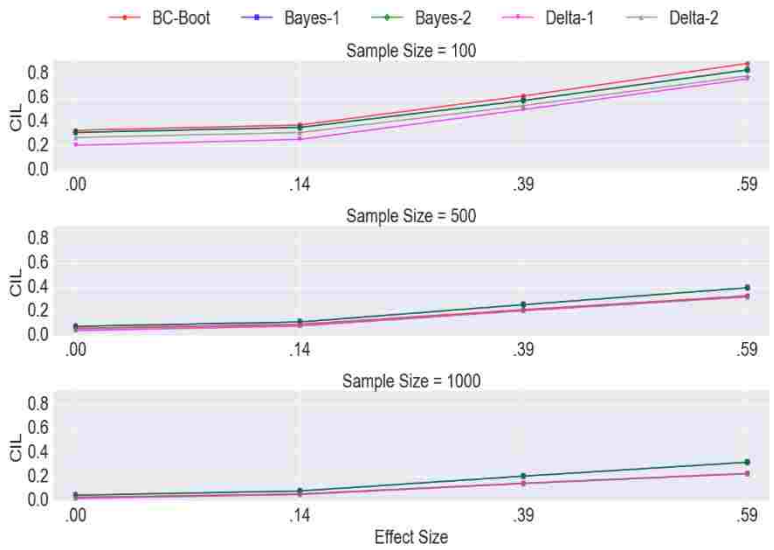


3. Model: Mediation, Mediator: Categorical, Endogenous: Continuous

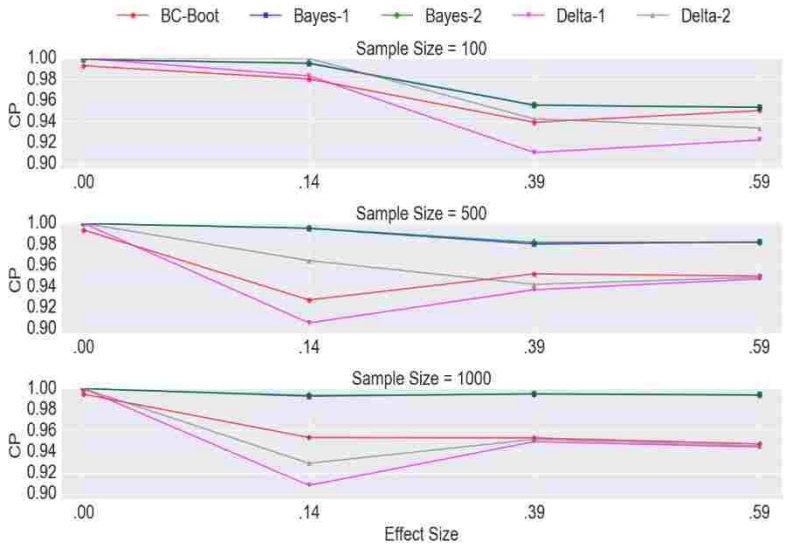
Empirical bias:



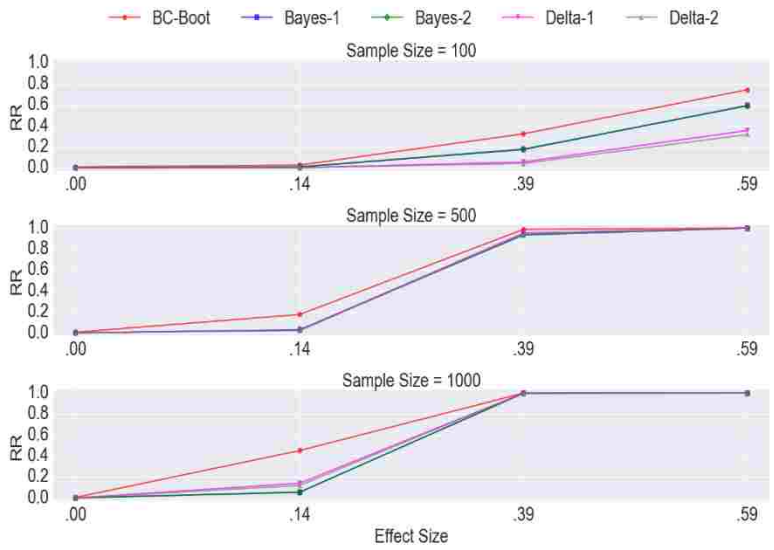
Confidence interval length:



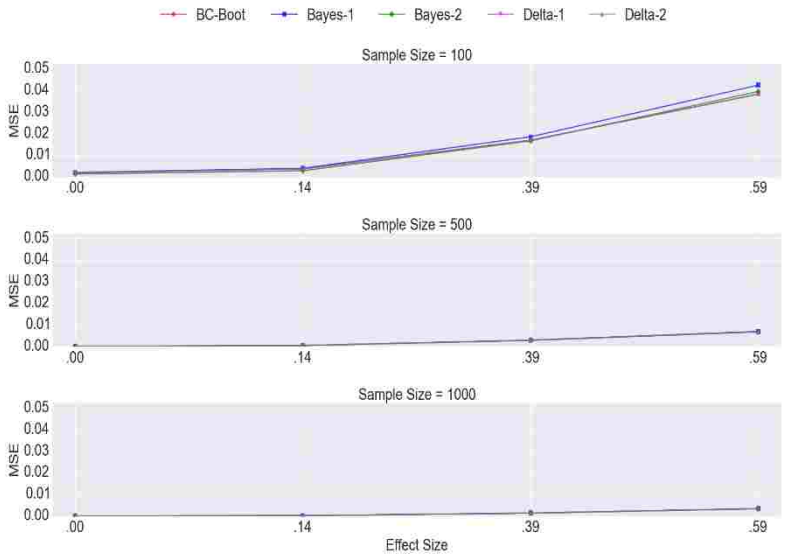
Coverage probability:



Rejection rate:

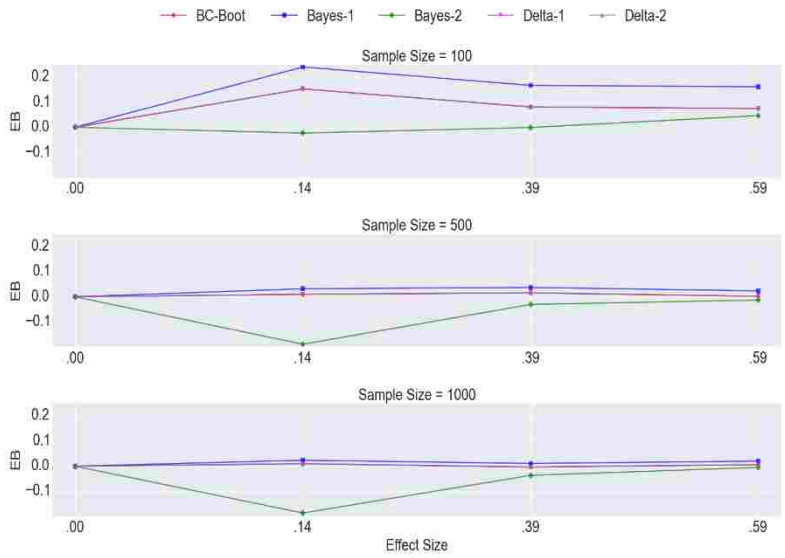


Mean squared error:

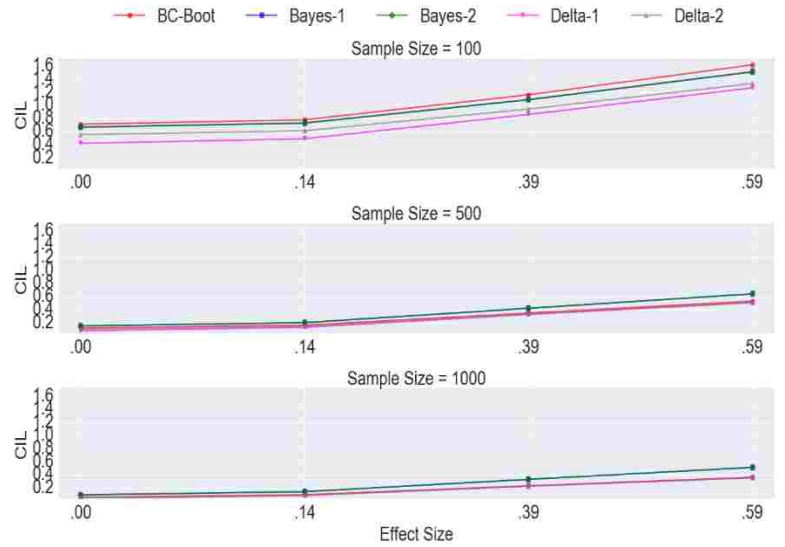


4. Model: Mediation, Mediator: Categorical, Endogenous: Categorical

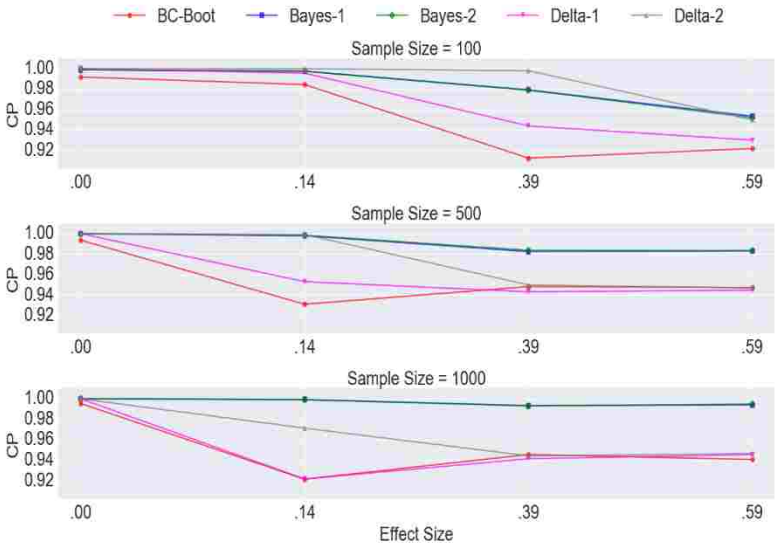
Empirical bias:



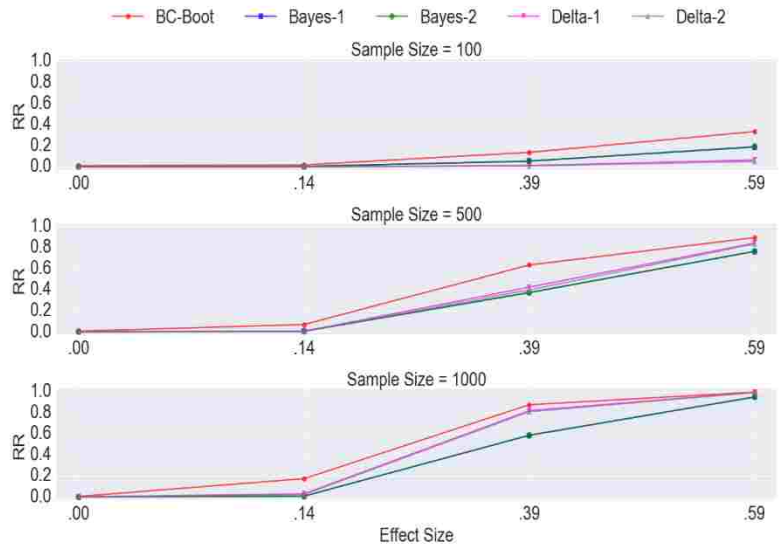
Confidence interval length:



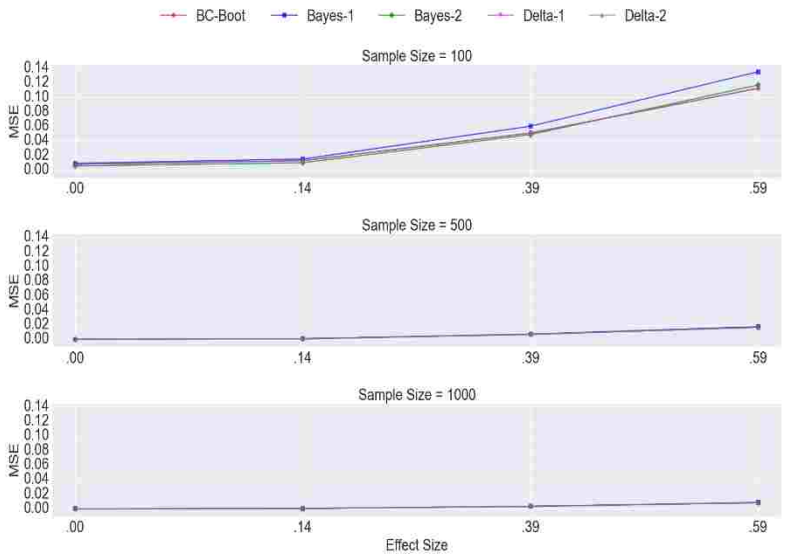
Coverage probability:



Rejection rate:

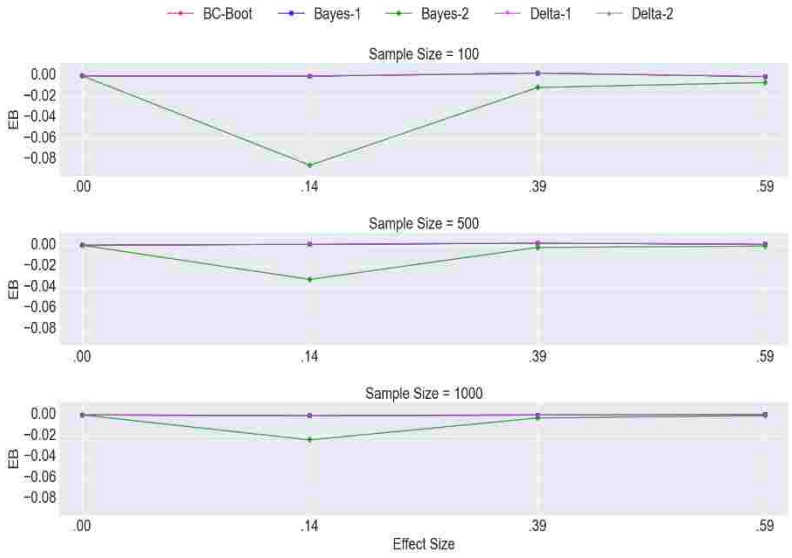


Mean squared error:

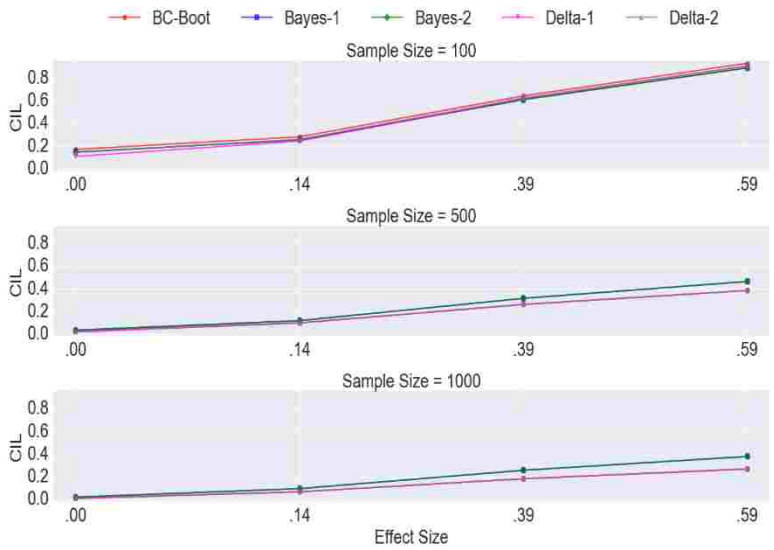


5. Model: Moderated Mediation, Mediator: Continuous, Endogenous: Continuous

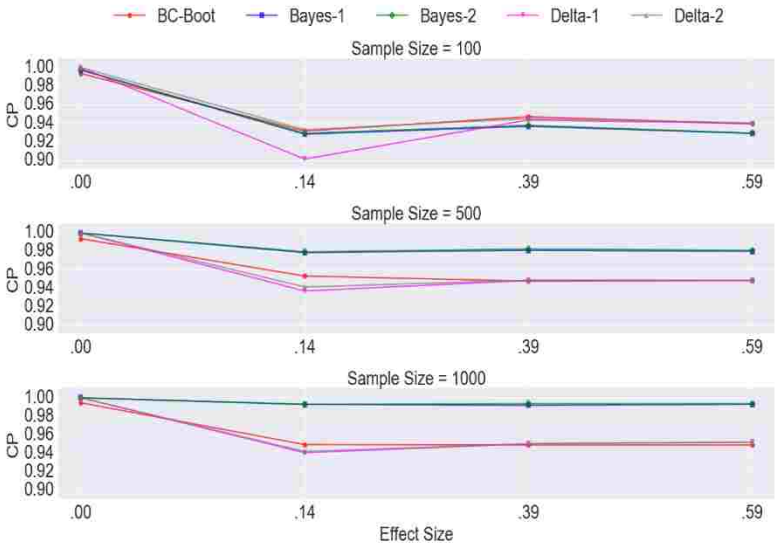
Empirical bias:



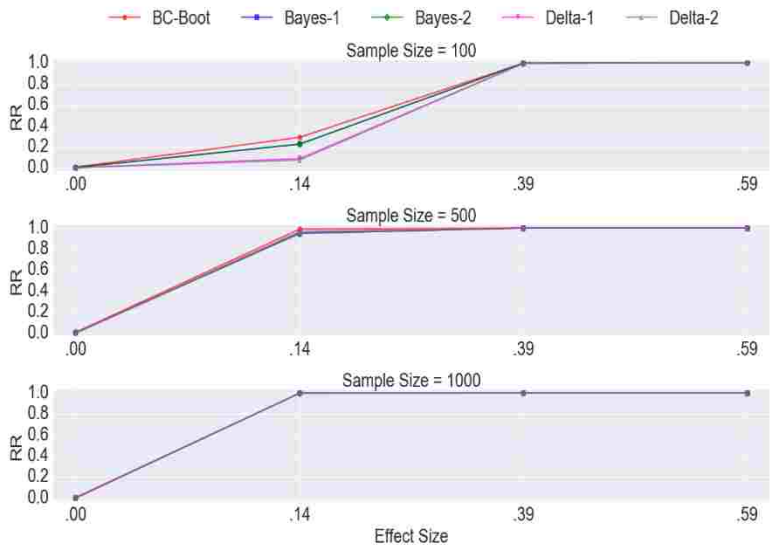
Confidence interval length:



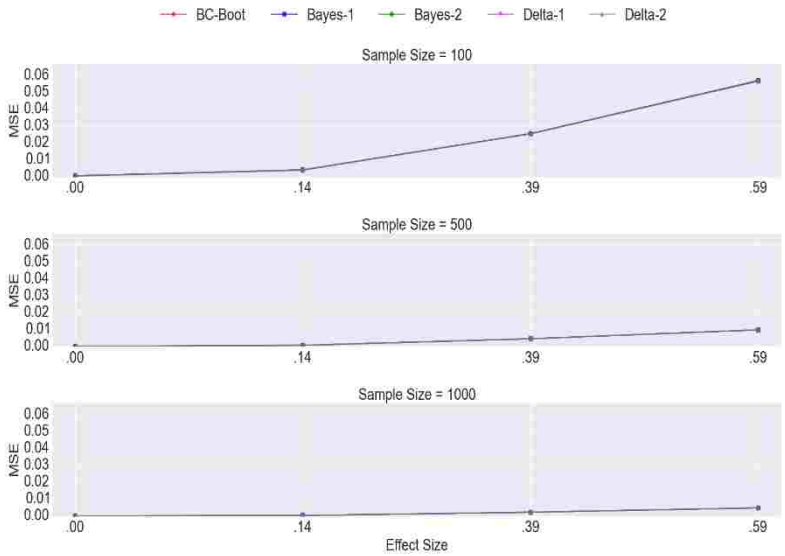
Coverage probability:



Rejection rate:

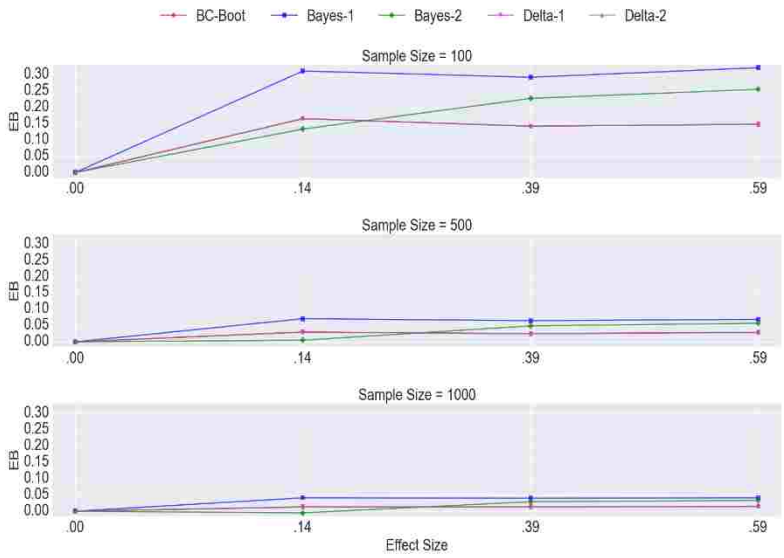


Mean squared error:

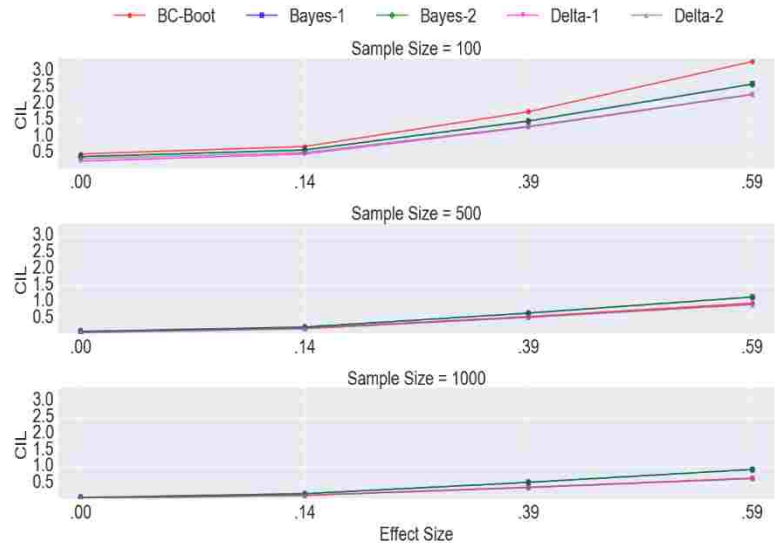


6. Model: Moderated Mediation, Mediator: Continuous, Endogenous: Categorical

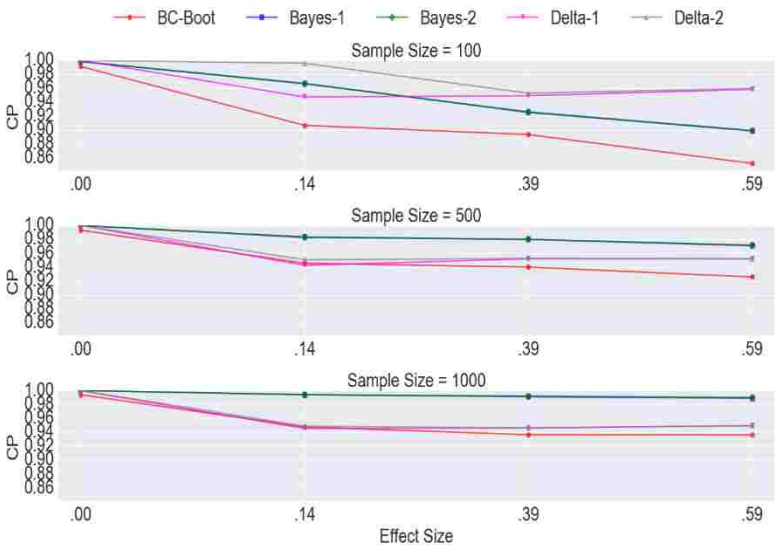
Empirical bias:



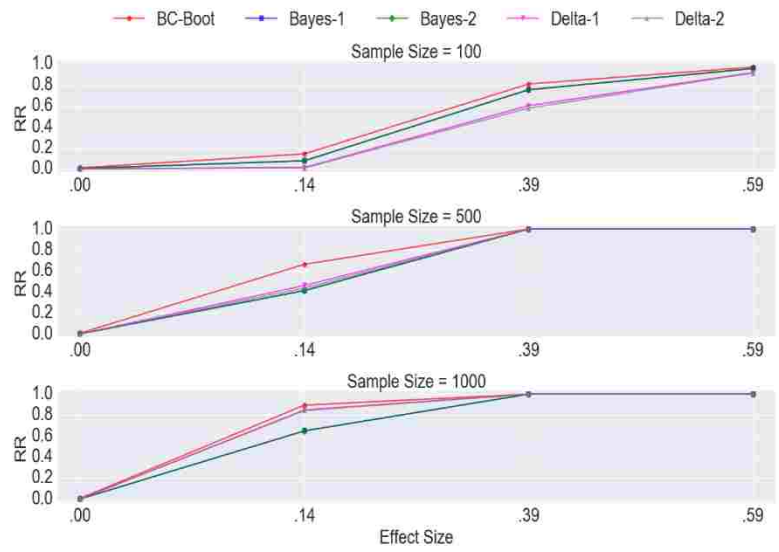
Confidence interval length:



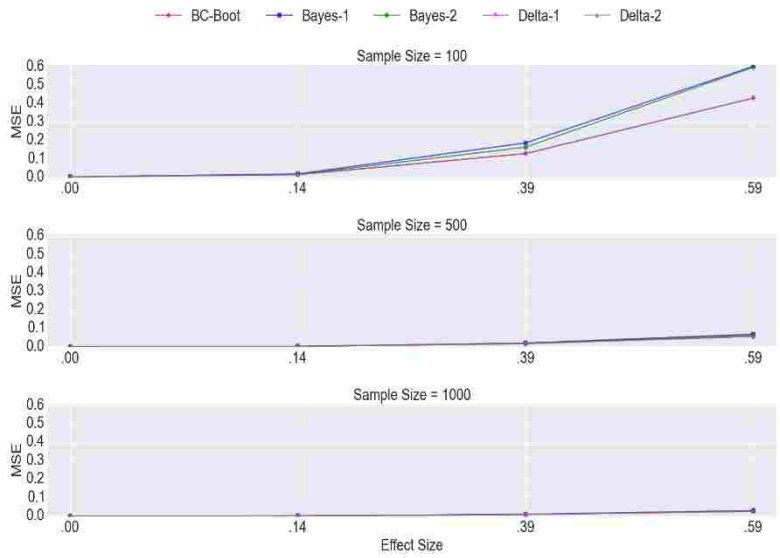
Coverage probability:



Rejection rate:

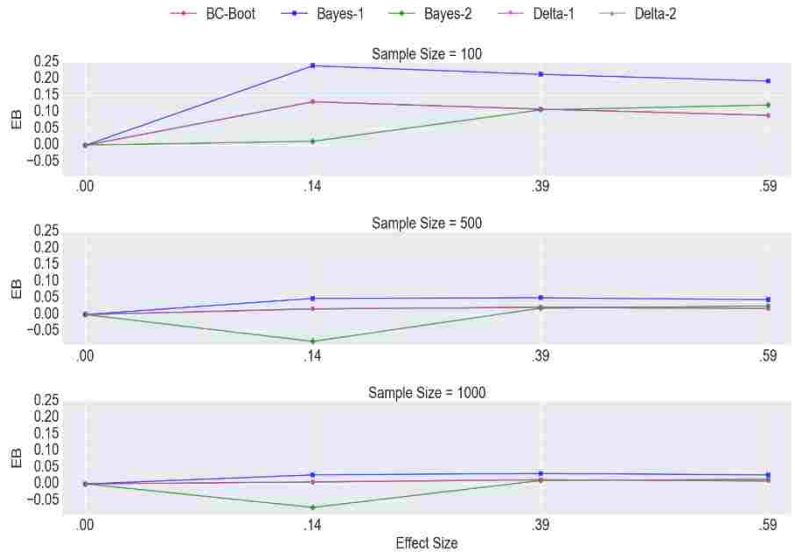


Mean squared error:

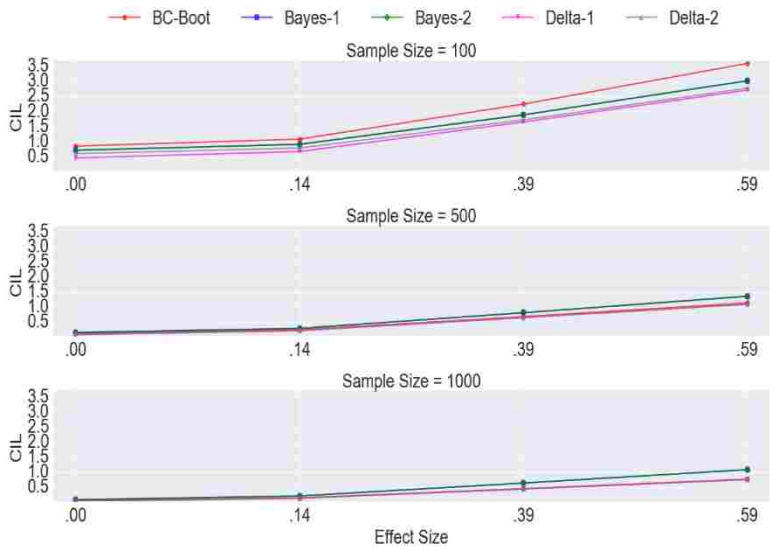


7. Model: Moderated Mediation, Mediator: Categorical, Endogenous: Continuous

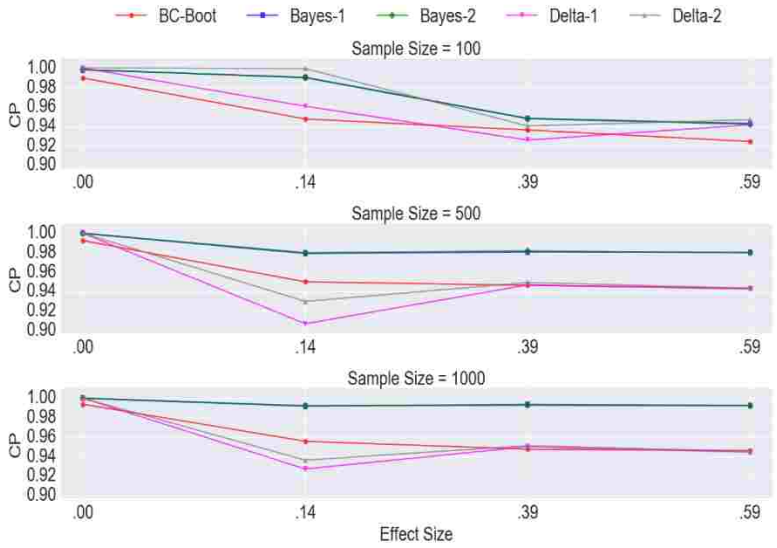
Empirical bias:



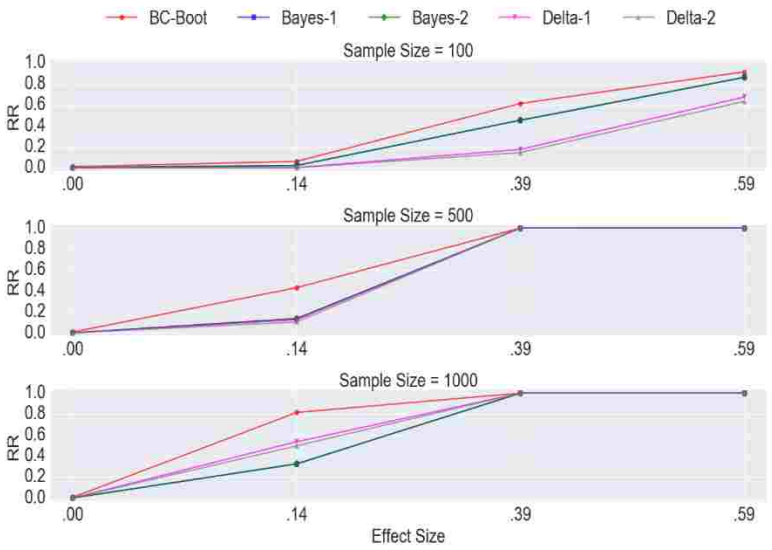
Confidence interval length:



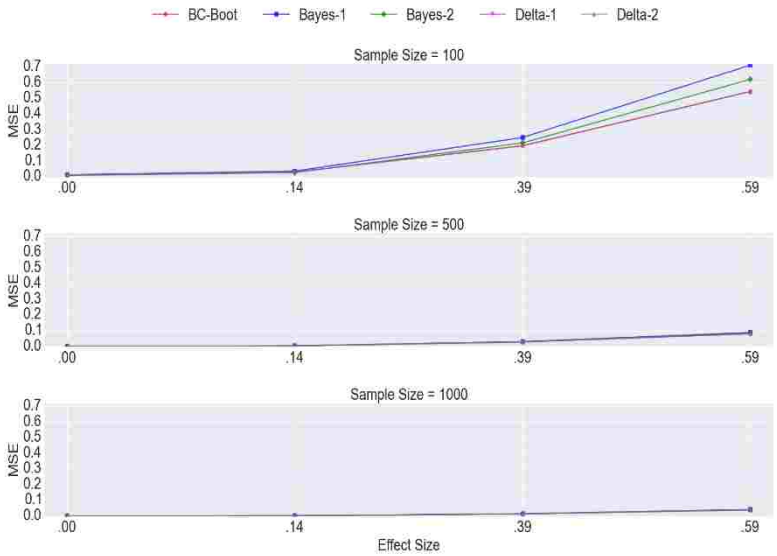
Coverage probability:



Rejection rate:

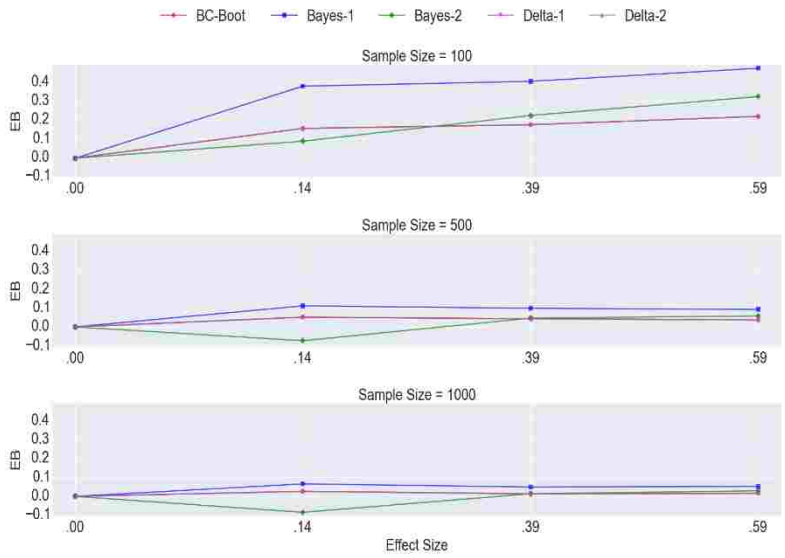


Mean squared error:

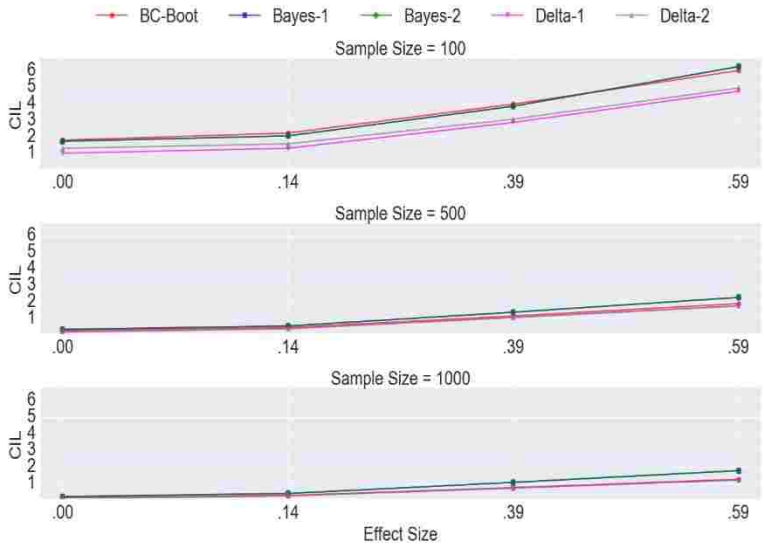


8. Model: Moderated Mediation, Mediator: Categorical, Endogenous: Categorical

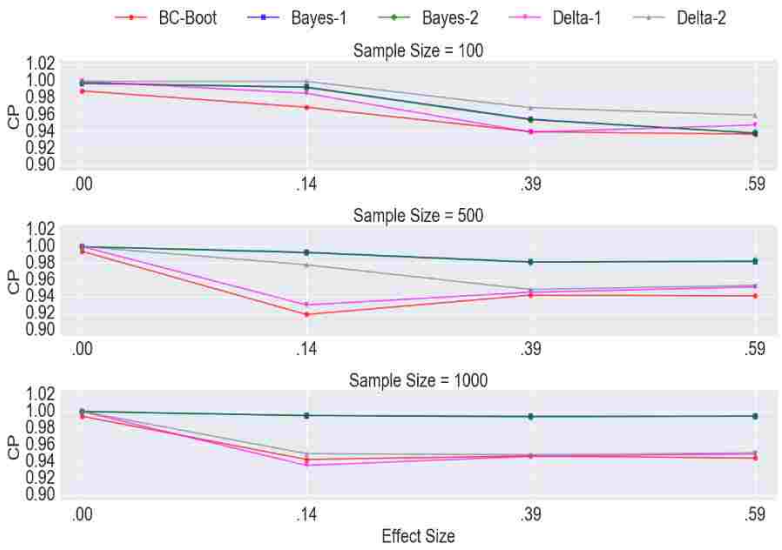
Empirical bias:



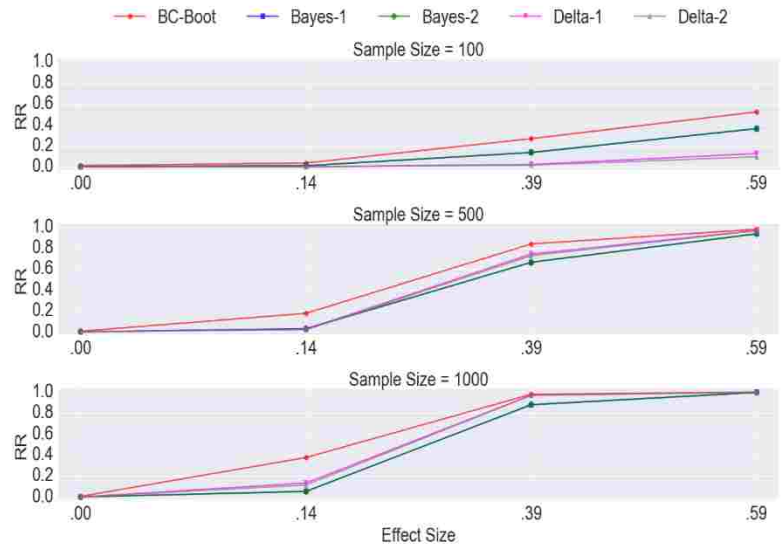
Confidence interval length:



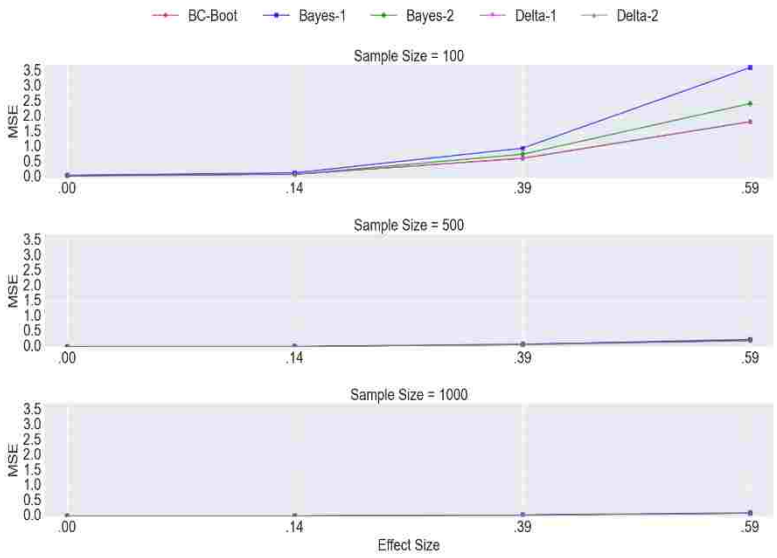
Coverage probability:



Rejection rate:



Mean squared error:



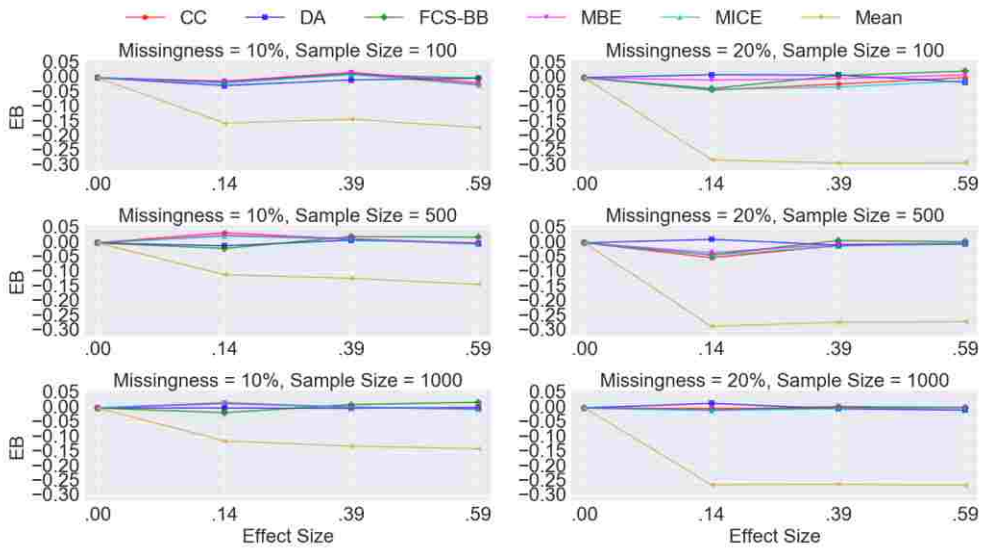
APPENDIX D

SUPPLEMENTAL FIGURES FROM SIMULATION 2

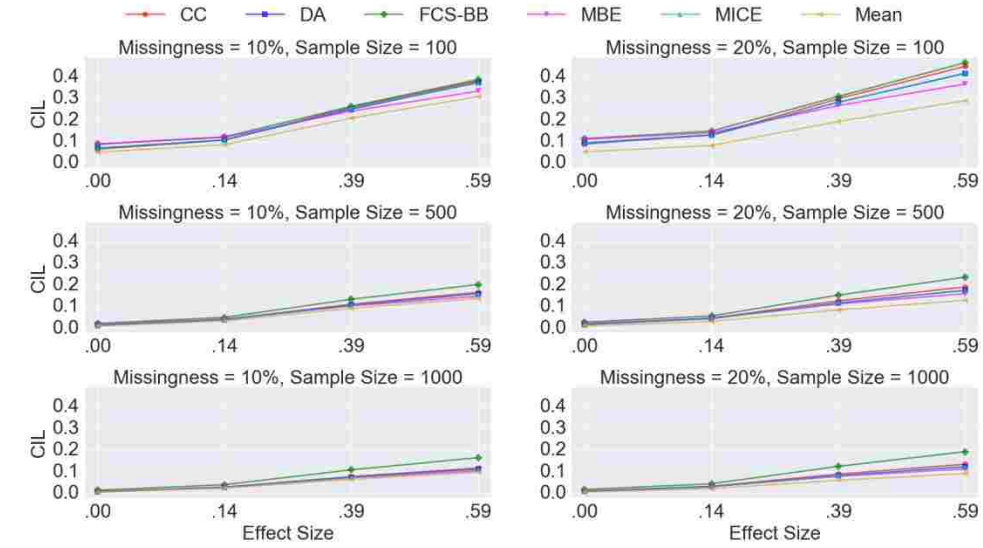
This Appendix presents the supplemental figures from Simulation Study 2 described in Chapter 4. For each of the models below, the acronyms for the metrics are: EB = empirical bias, CIL = confidence interval length, CP = coverage probability, RR = rejection rate, MSE = mean squared error, FMI = fraction of missing information. The missing data methods are abbreviated as: CC = complete case analysis, DA = data augmentation, FCS-BB = fully conditional specification with Bayesian bootstrapping, FCS-BB* = modified fully conditional specification with Bayesian bootstrapping, FCS-BB-PI = fully conditional specification with Bayesian bootstrapping – passive imputation, FCS-BB-JAV = fully conditional specification with Bayesian bootstrapping – just another variable, MBE = model-based estimation, MICE = multiple imputation by chained equations, Mean = mean imputation.

1. Model: Mediation, Mediator: Continuous, Endogenous: Continuous

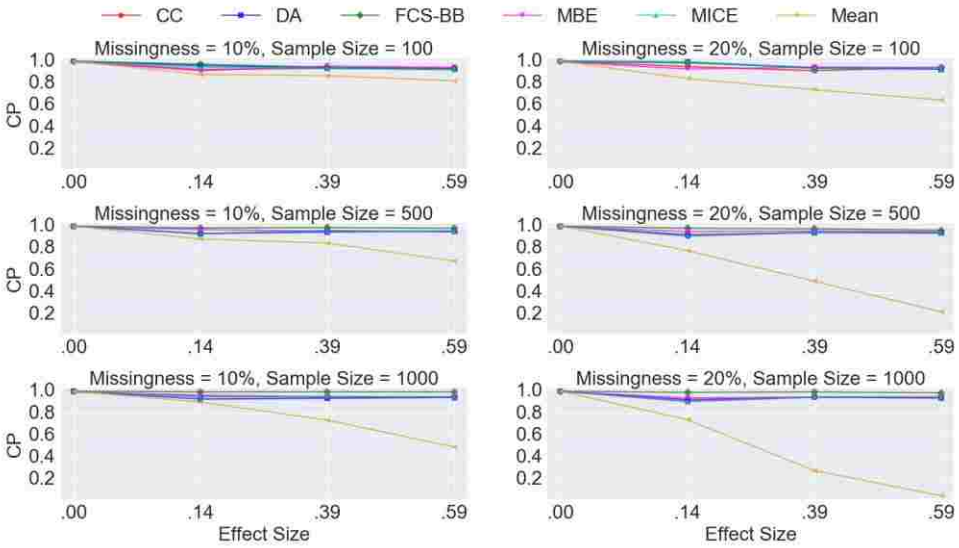
Empirical bias:



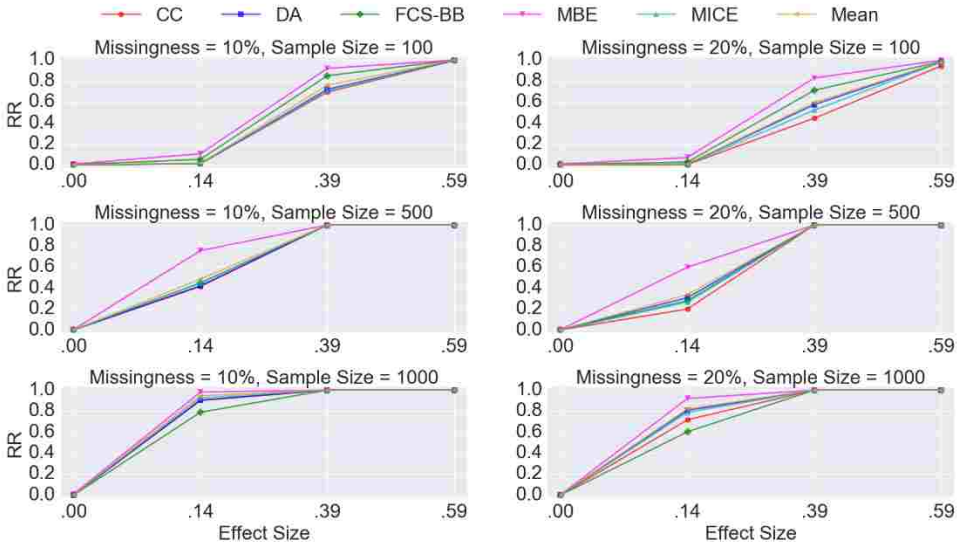
Confidence interval length:



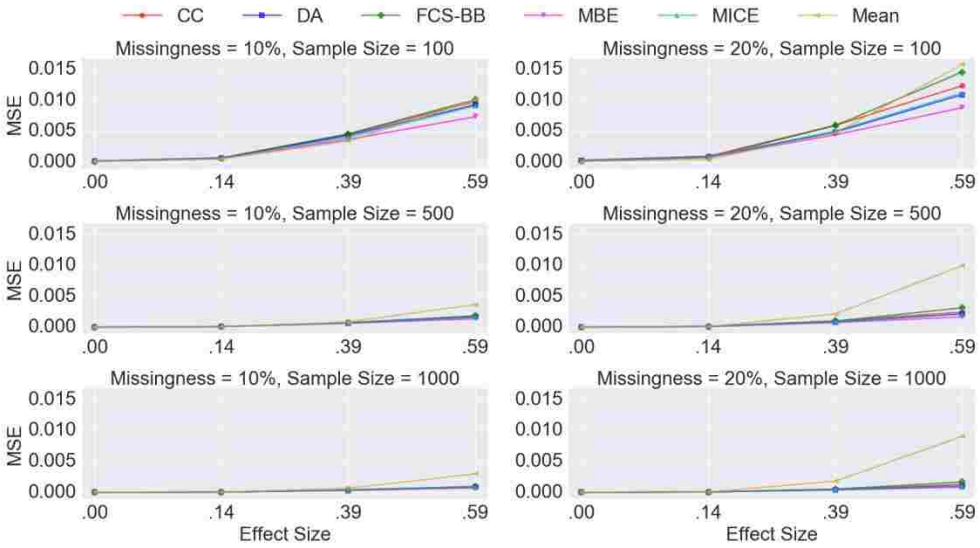
Coverage probability:



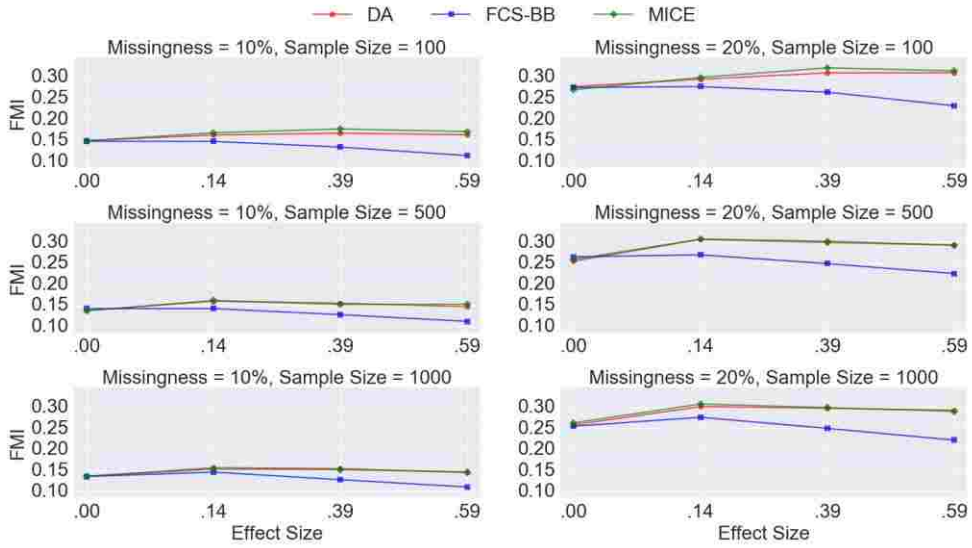
Rejection rate:



Mean squared error:

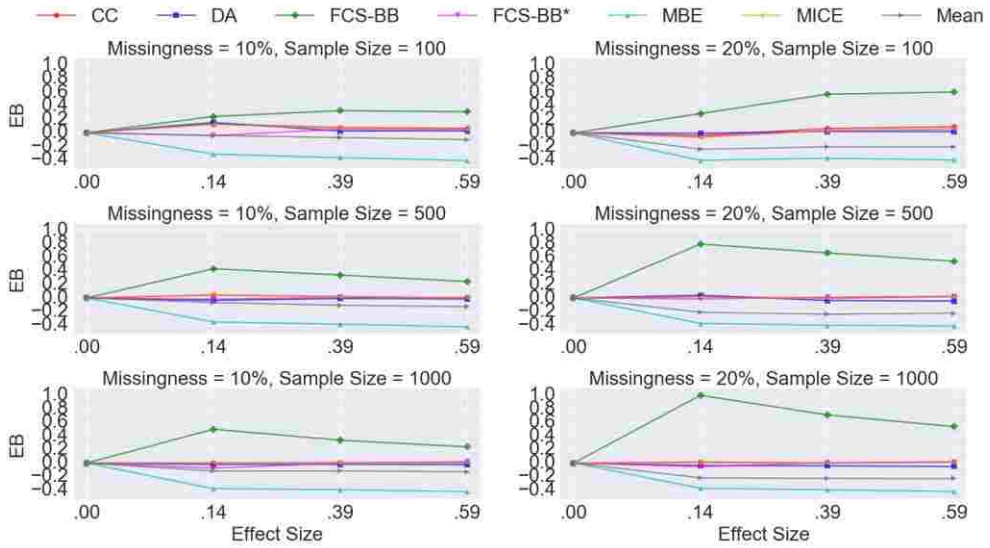


Fraction of missing information:

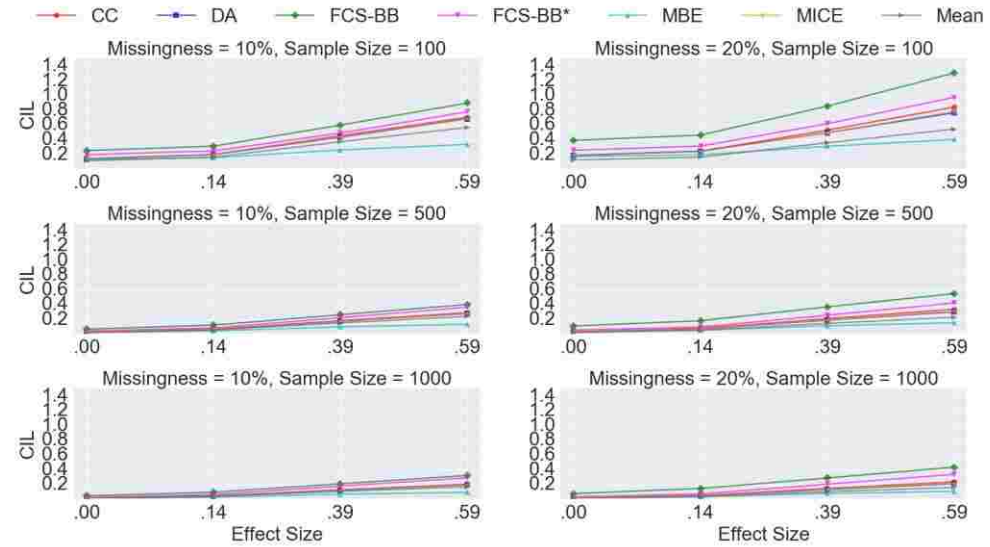


2. Model: Mediation, Mediator: Continuous, Endogenous: Categorical

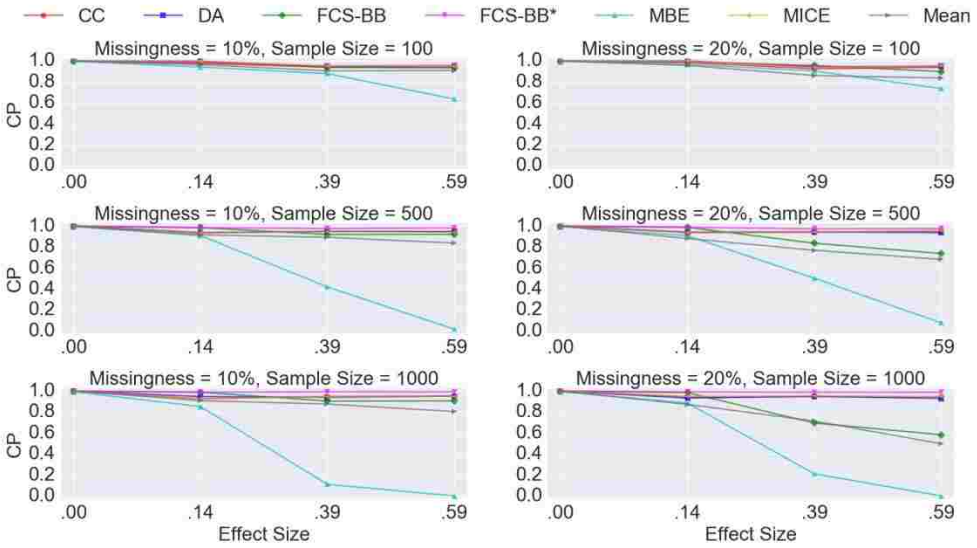
Empirical bias:



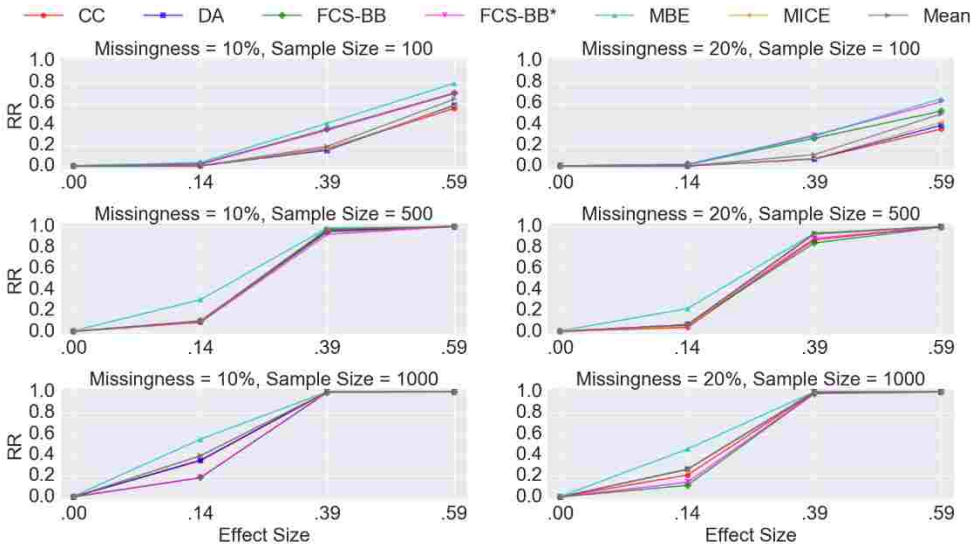
Confidence interval length:



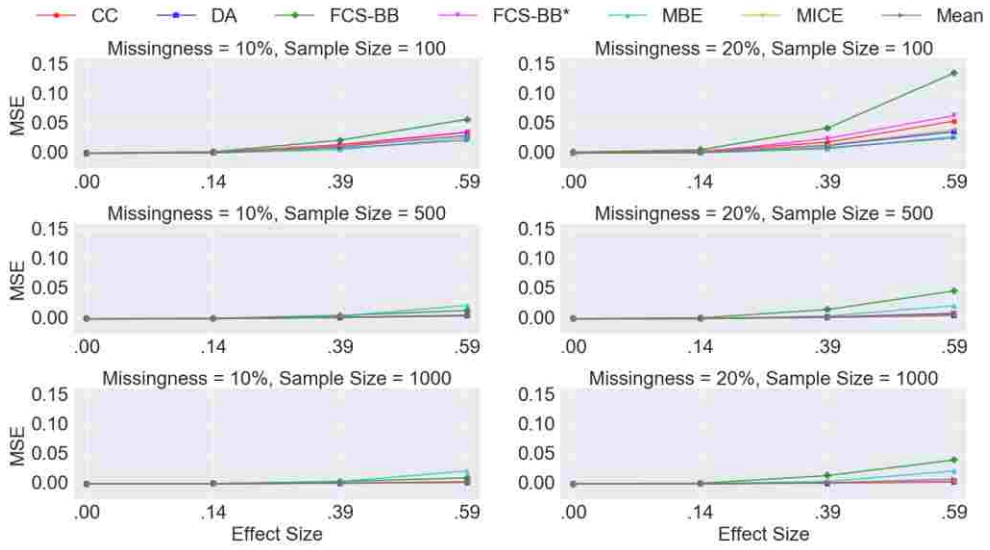
Coverage probability:



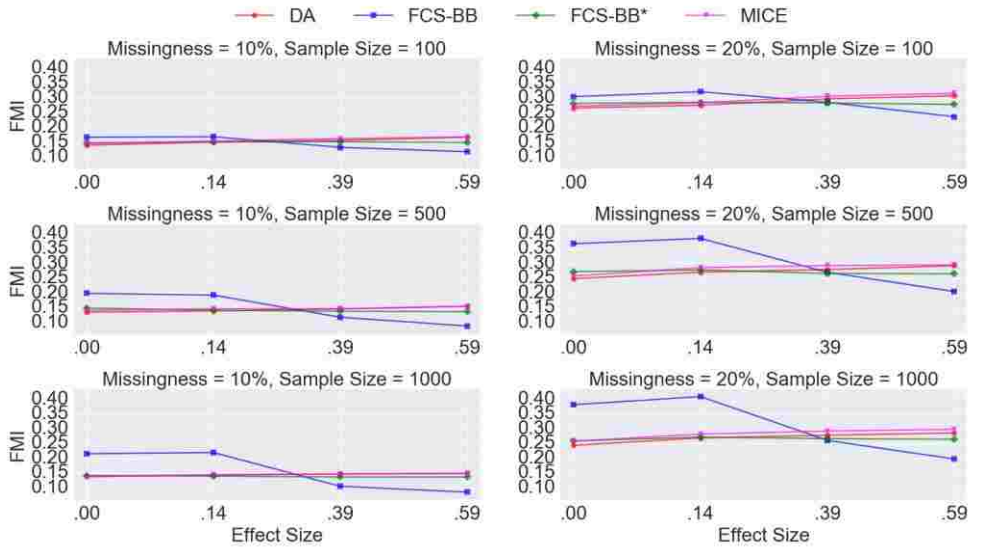
Rejection rate:



Mean squared error:

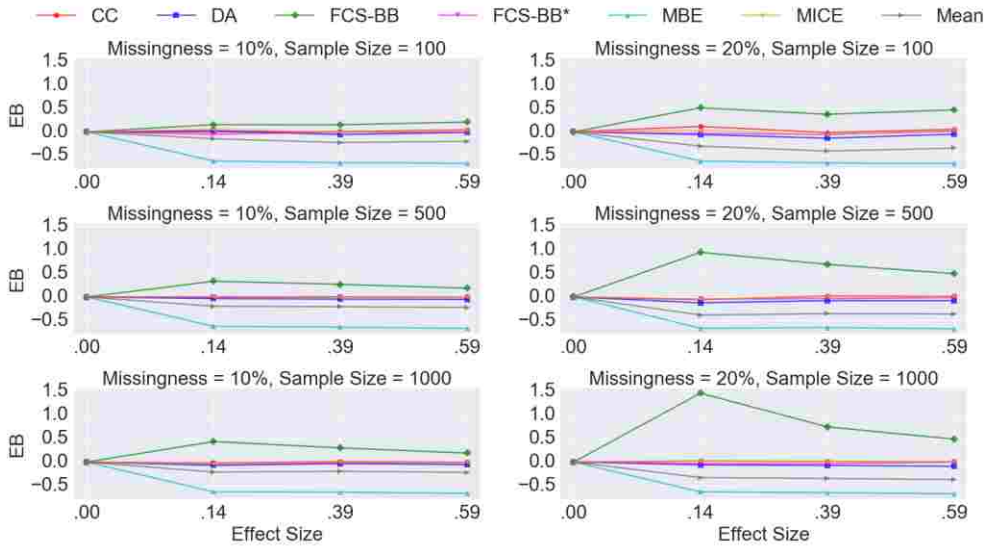


Fraction of missing information:

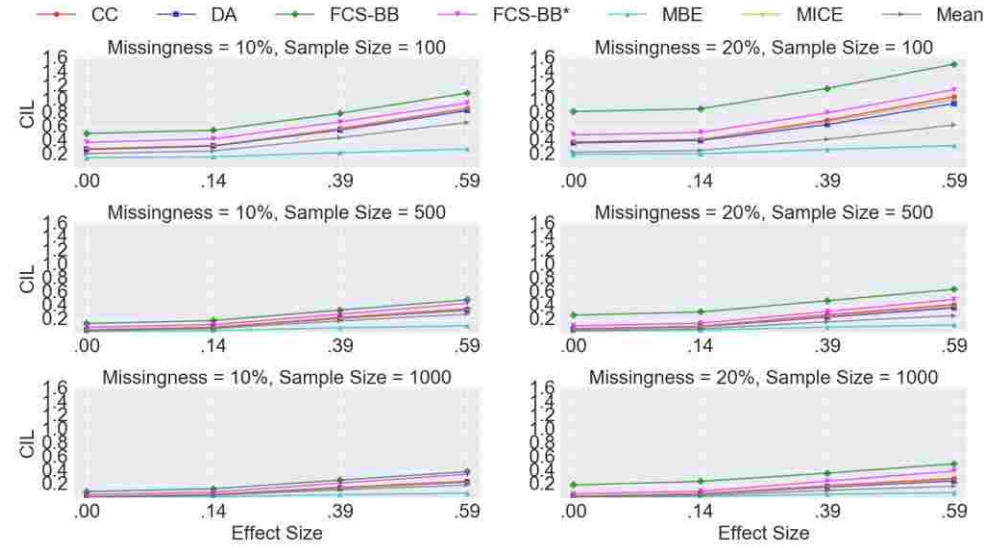


3. Model: Mediation, Mediator: Categorical, Endogenous: Continuous

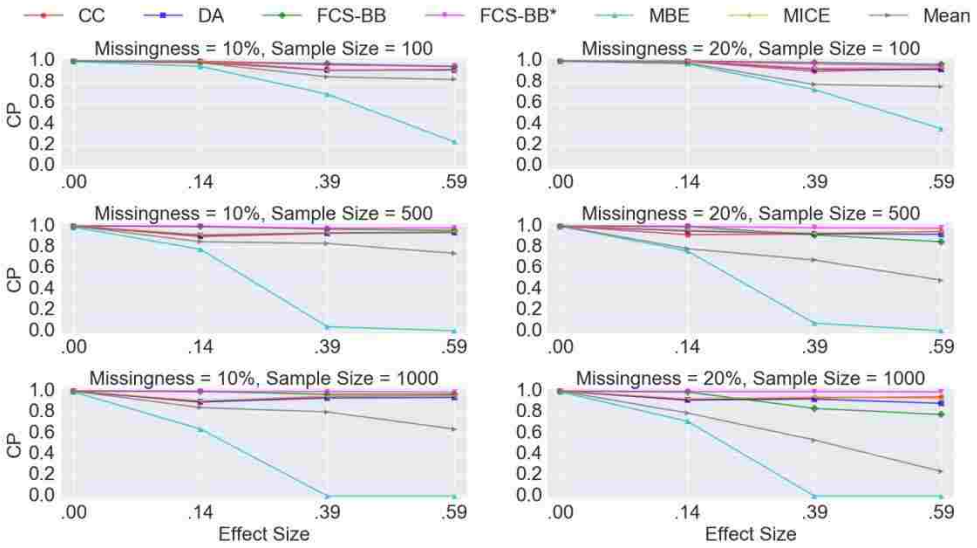
Empirical bias:



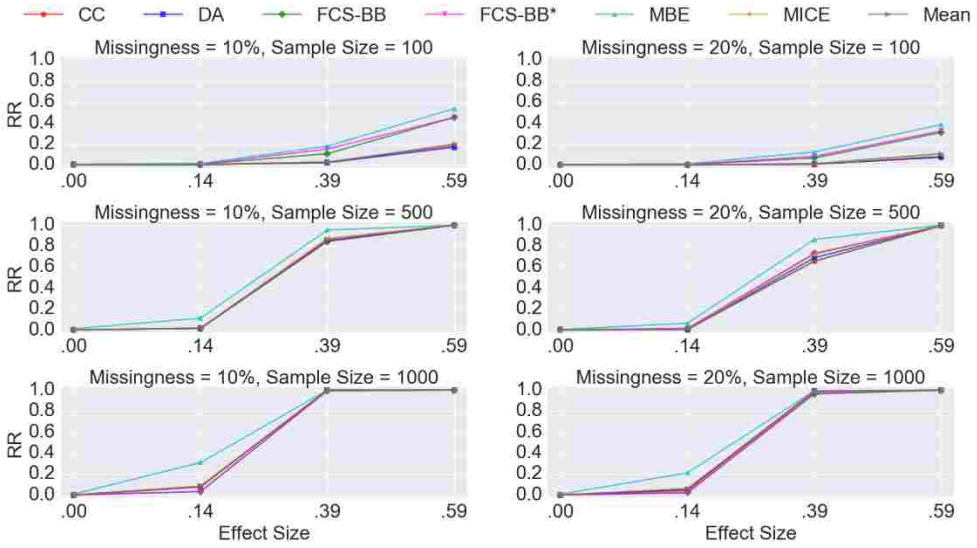
Confidence interval length:



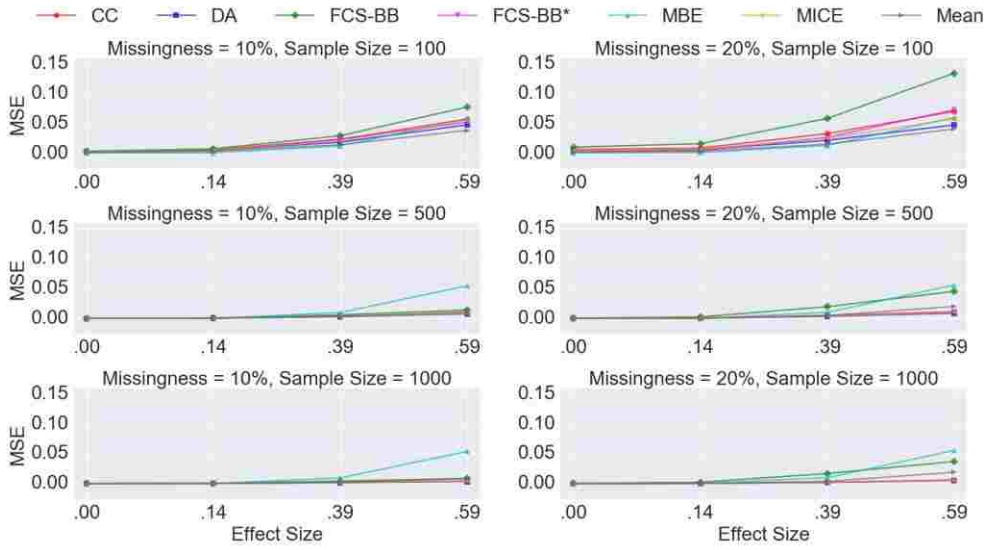
Coverage probability:



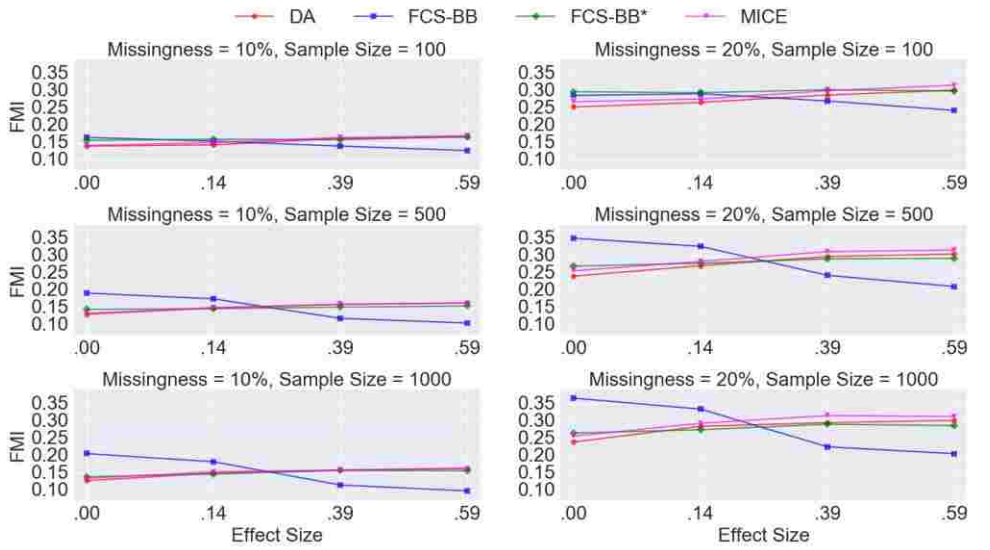
Rejection rate:



Mean squared error:

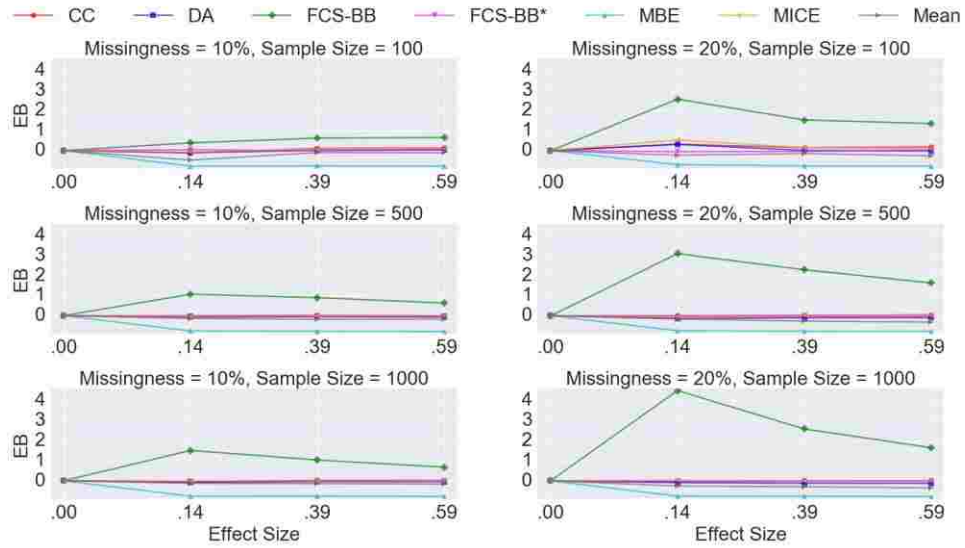


Fraction of missing information:

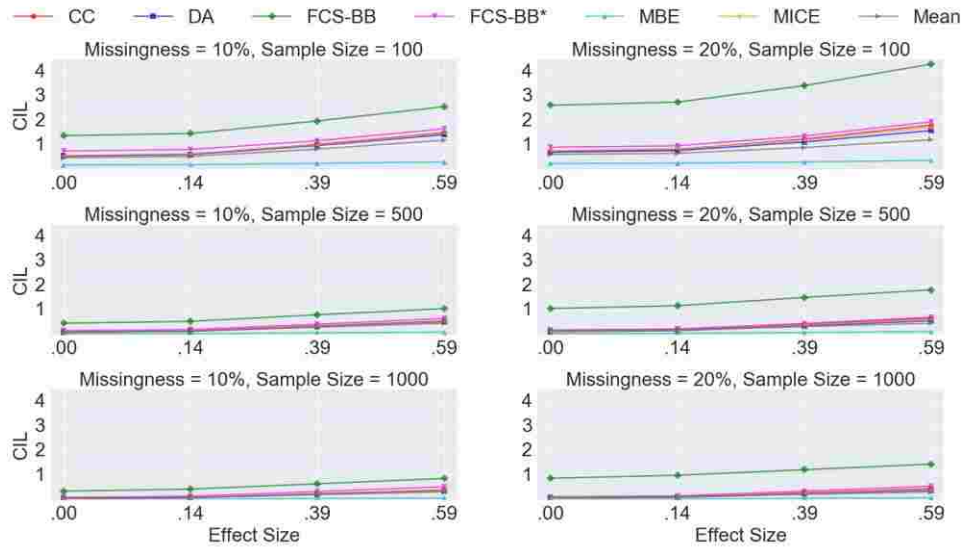


4. Model: Mediation, Mediator: Categorical, Endogenous: Categorical

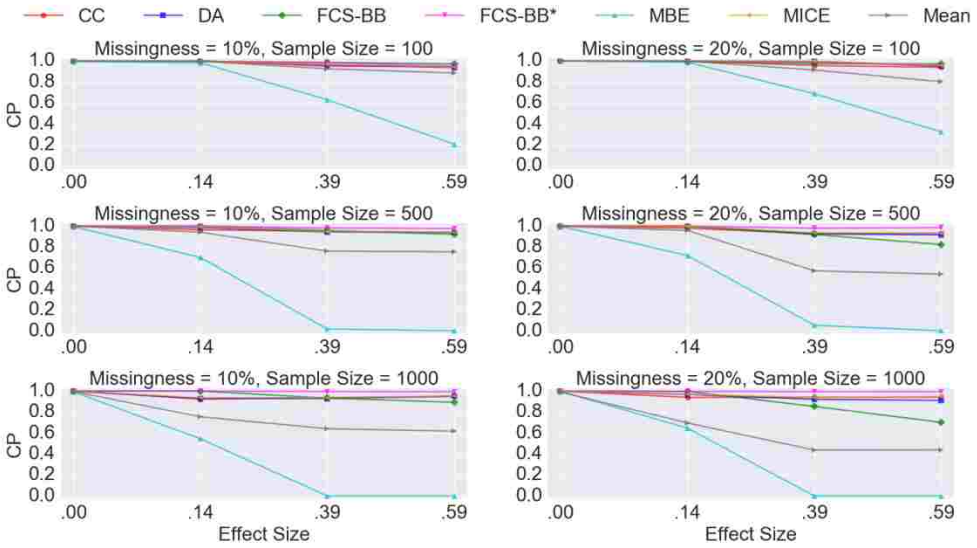
Empirical bias:



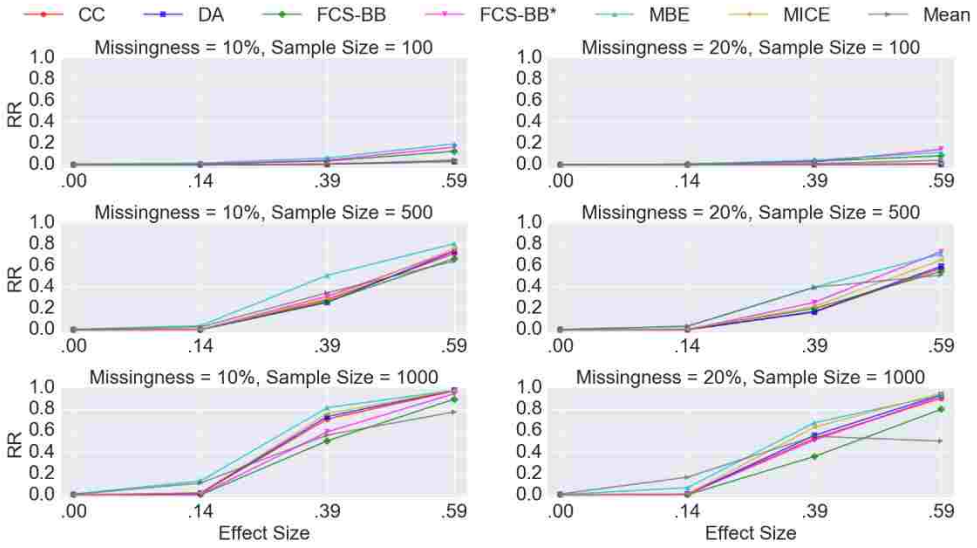
Confidence interval length:



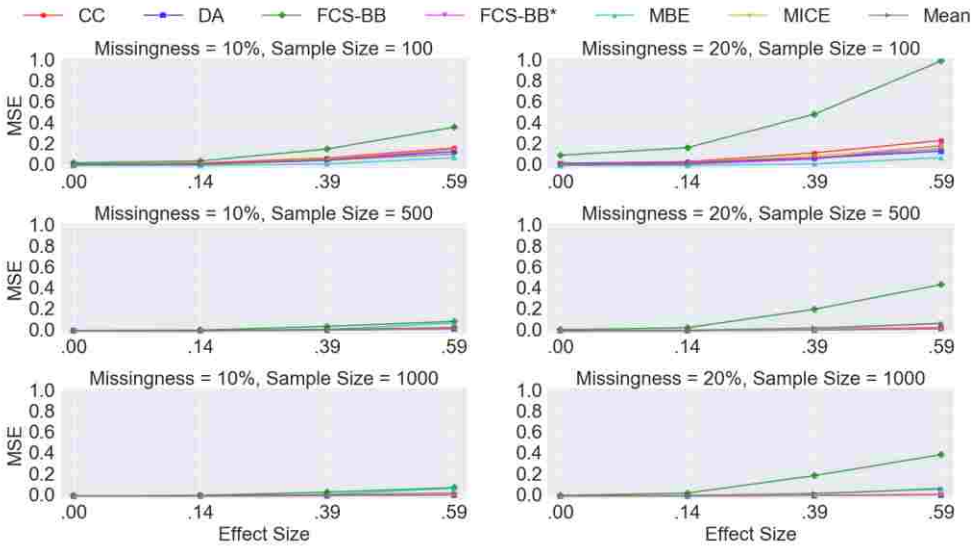
Coverage probability:



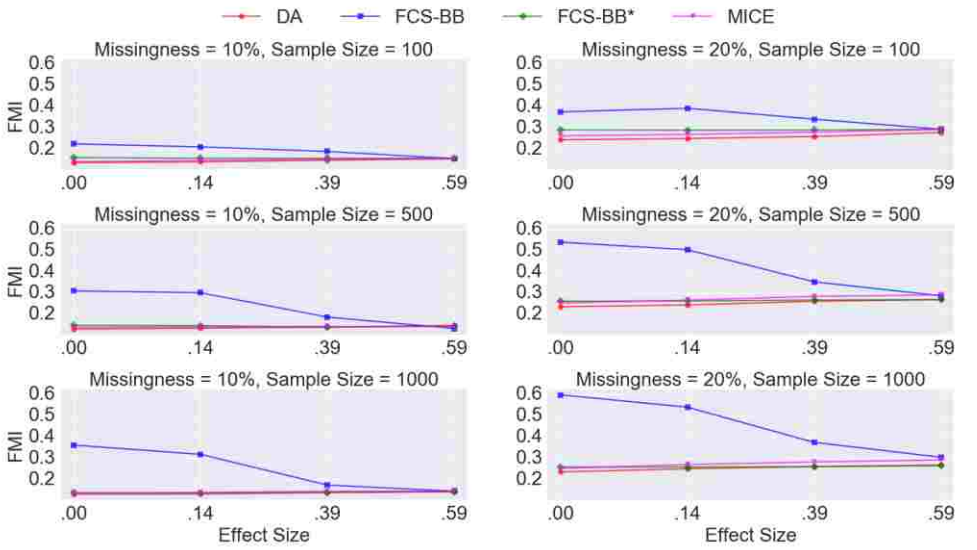
Rejection rate:



Mean squared error:

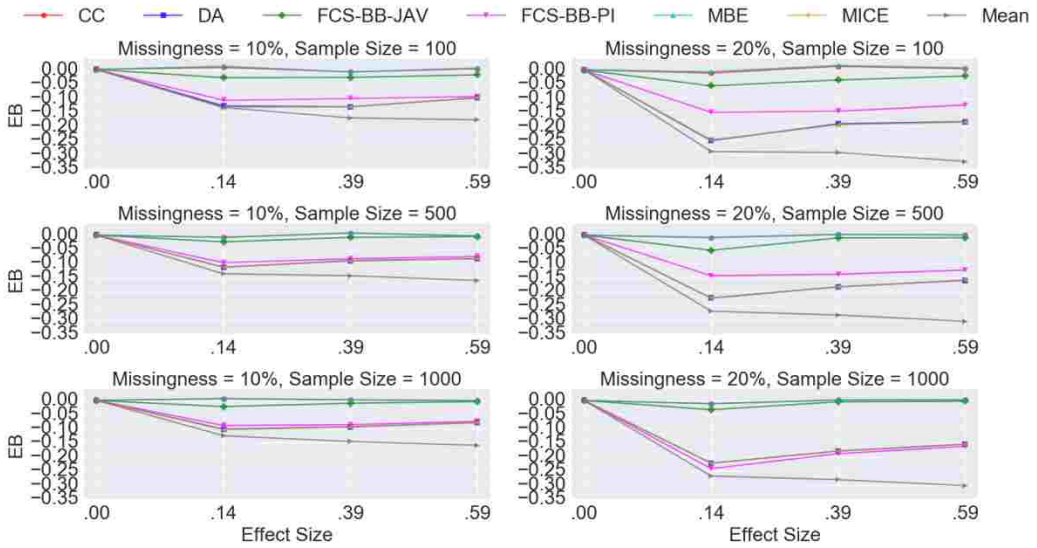


Fraction of missing information:

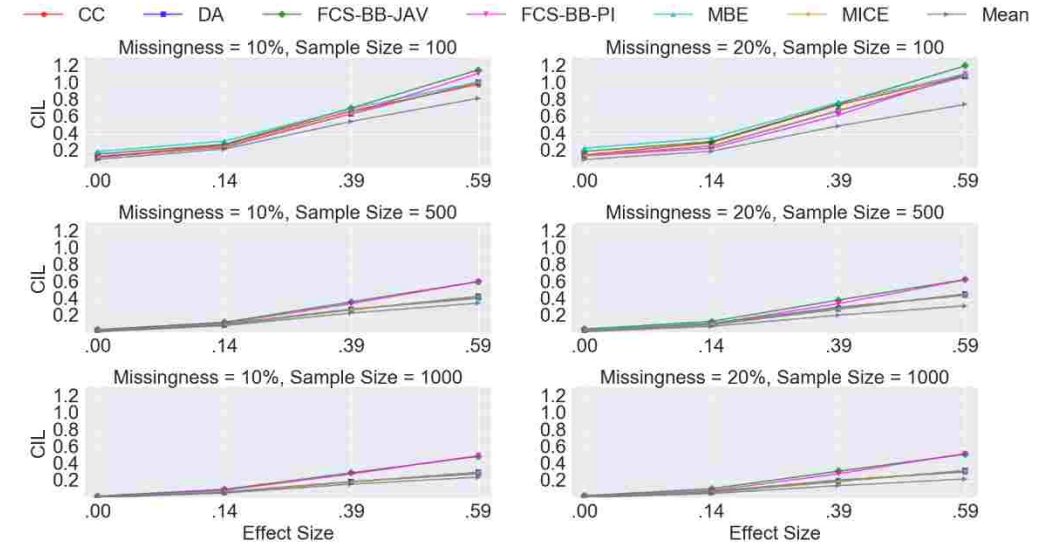


5. Model: Moderated Mediation, Mediator: Continuous, Endogenous: Continuous

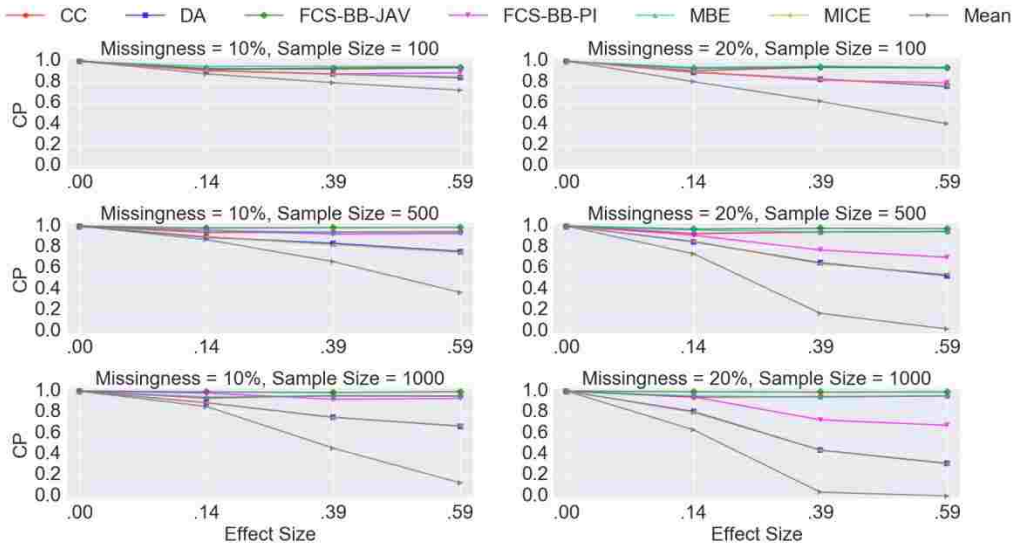
Empirical bias:



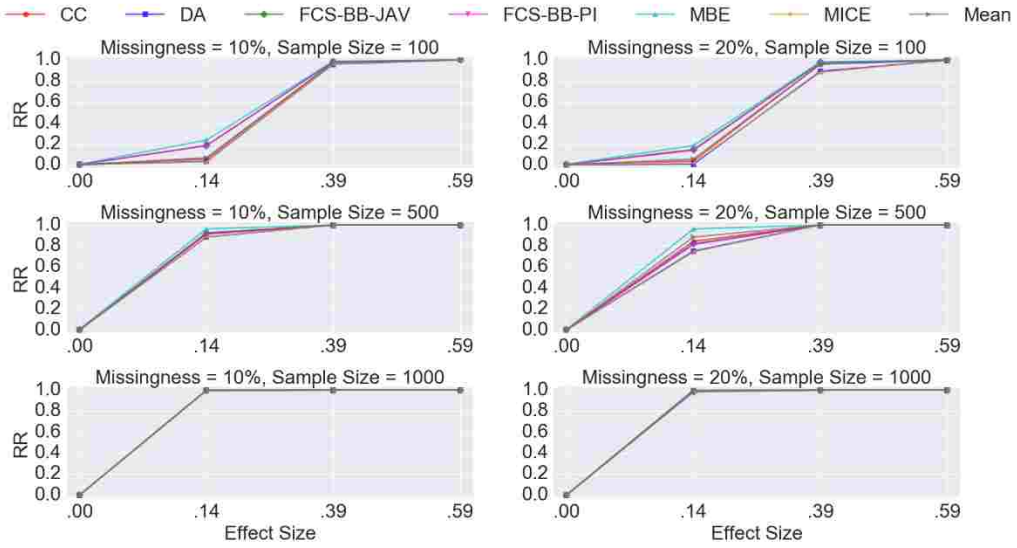
Confidence interval length:



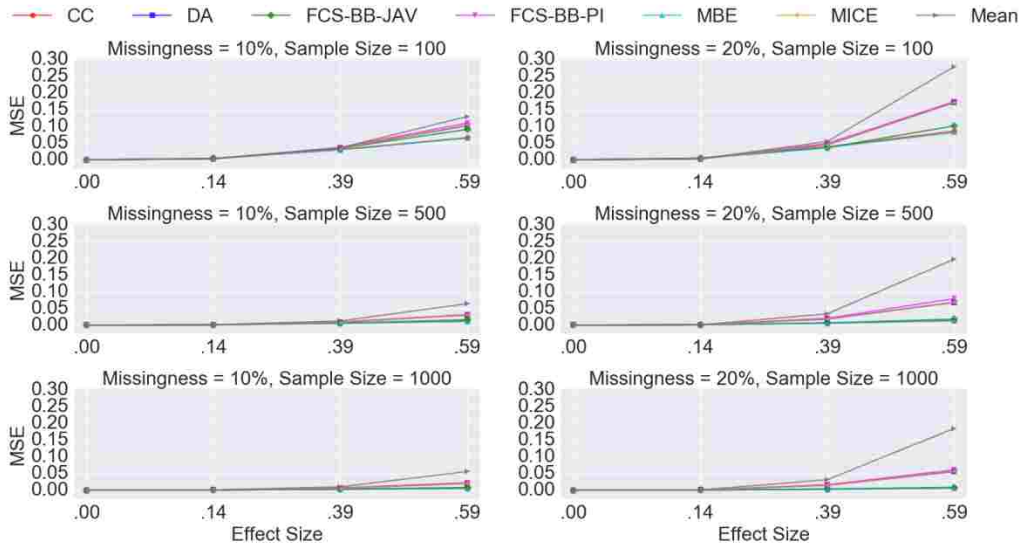
Coverage probability:



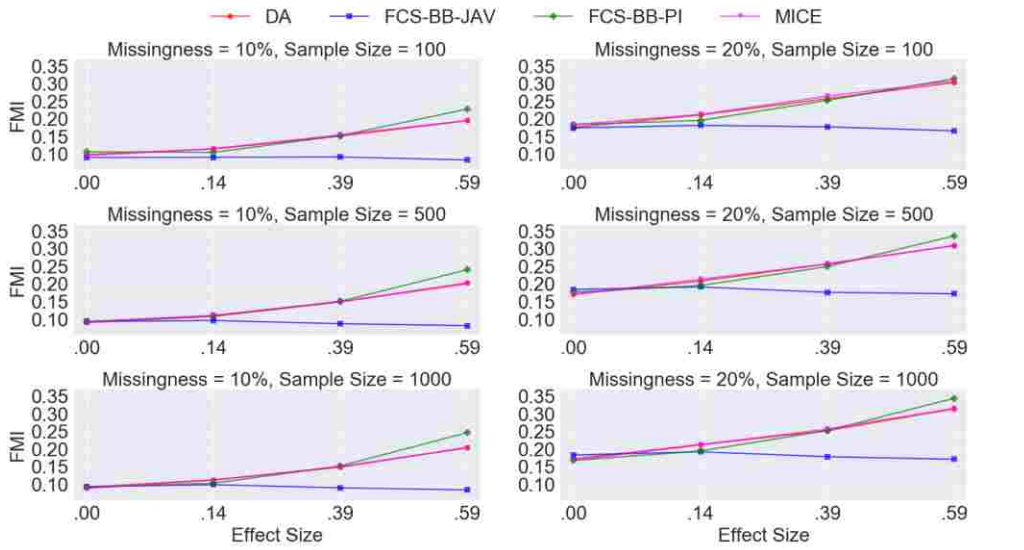
Rejection rate:



Mean squared error:

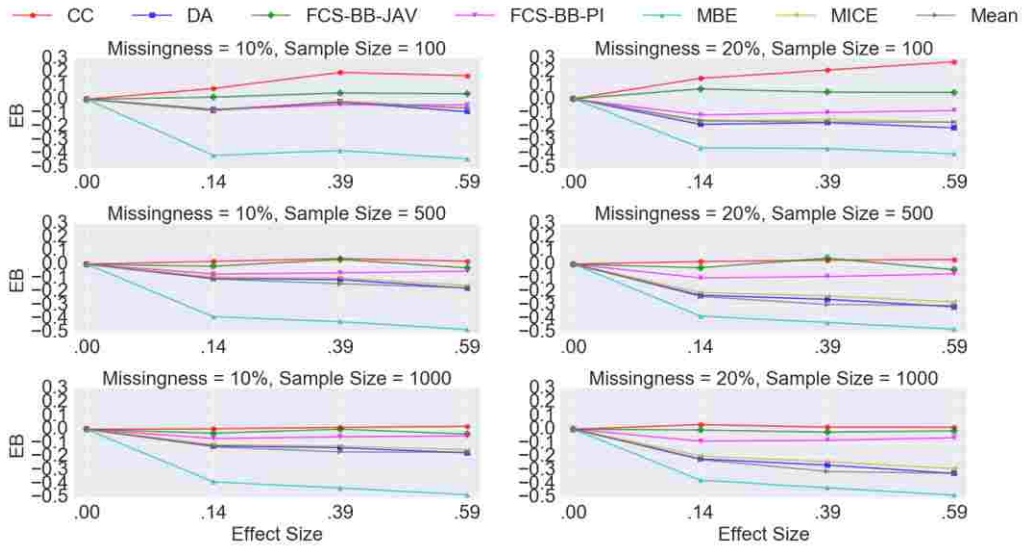


FMI:

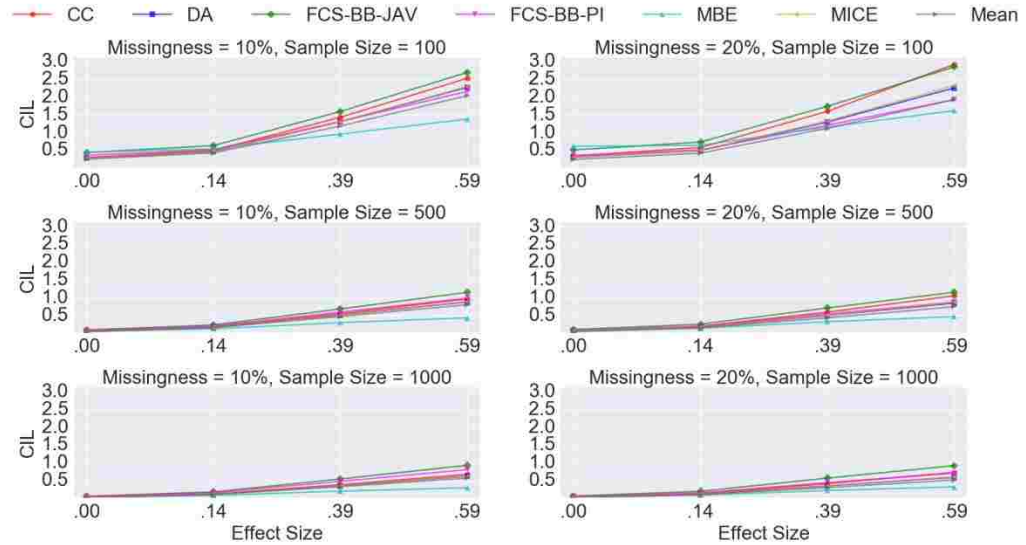


6. Model: Moderated Mediation, Mediator: Continuous, Endogenous: Categorical

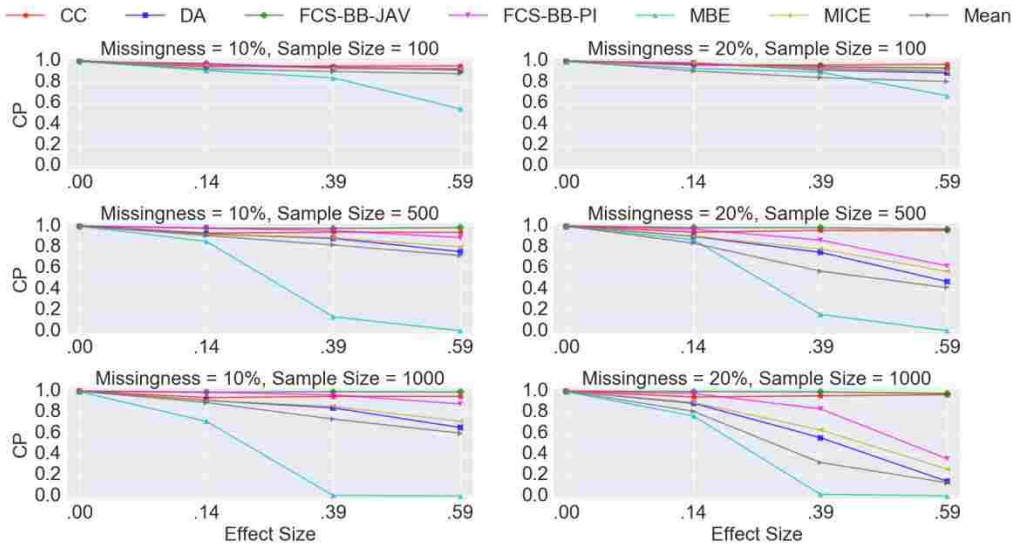
Empirical bias:



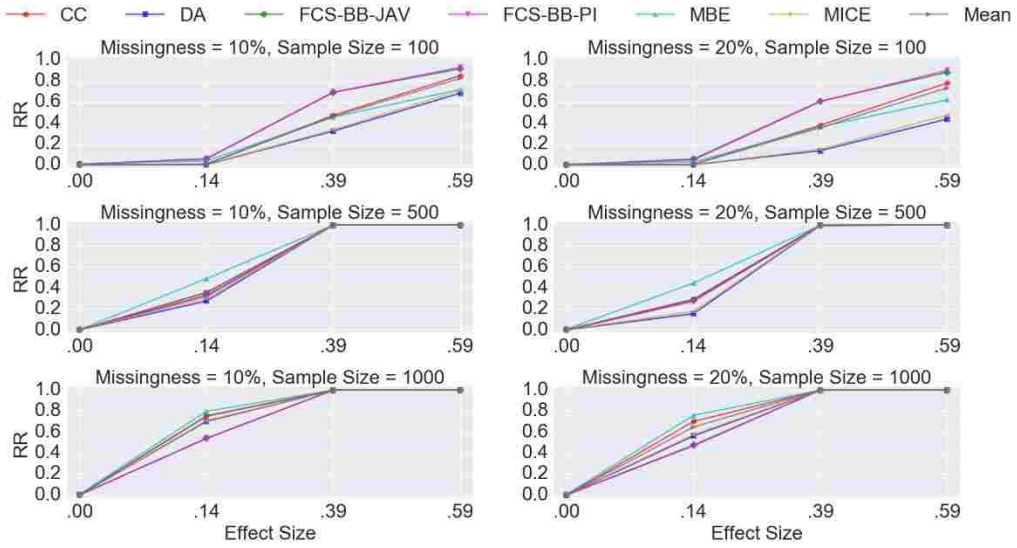
Confidence interval length:



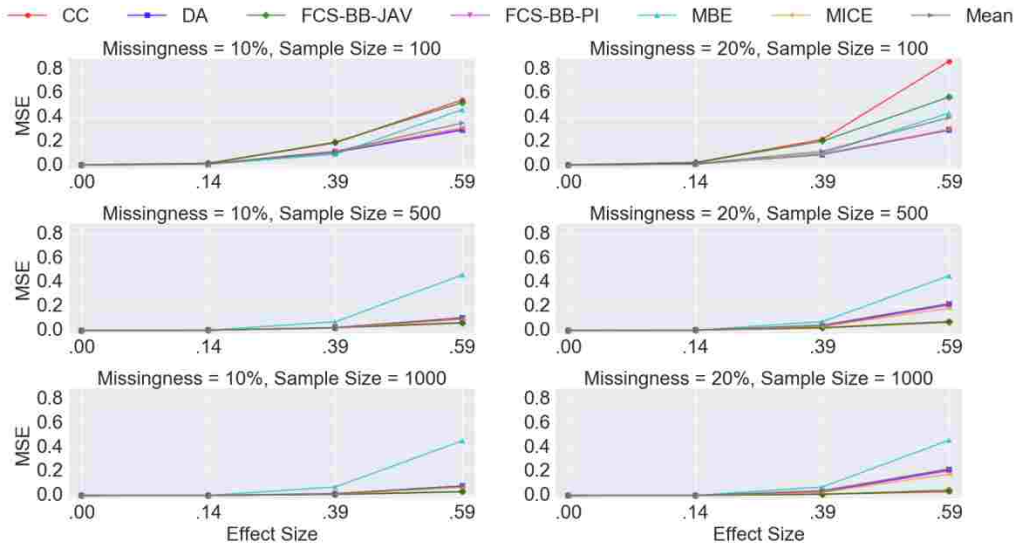
Coverage probability:



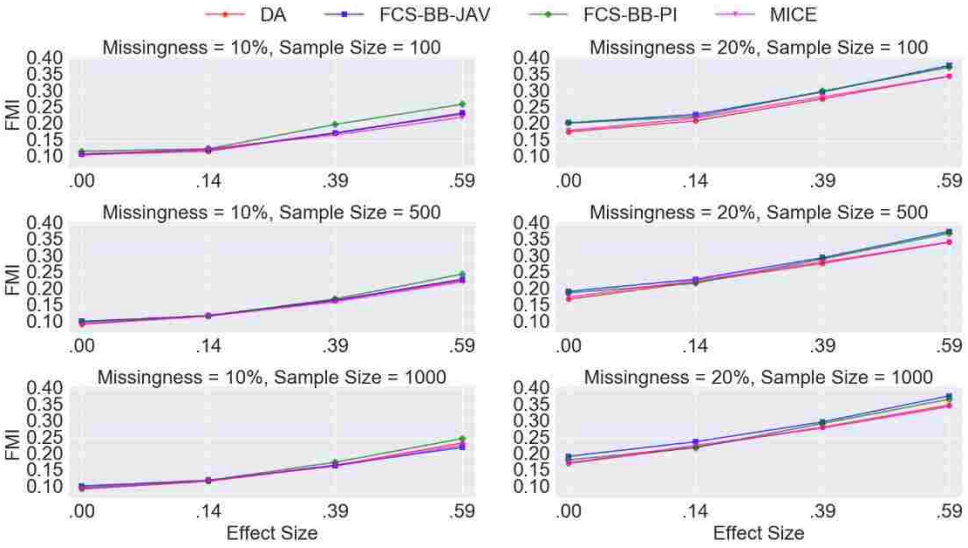
Rejection rate:



Mean squared error:

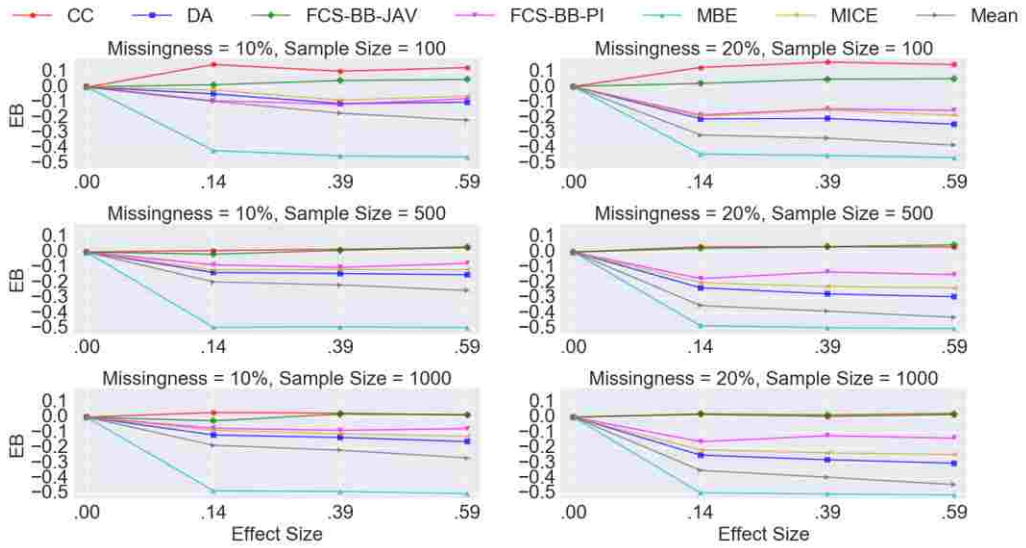


Fraction of missing information:

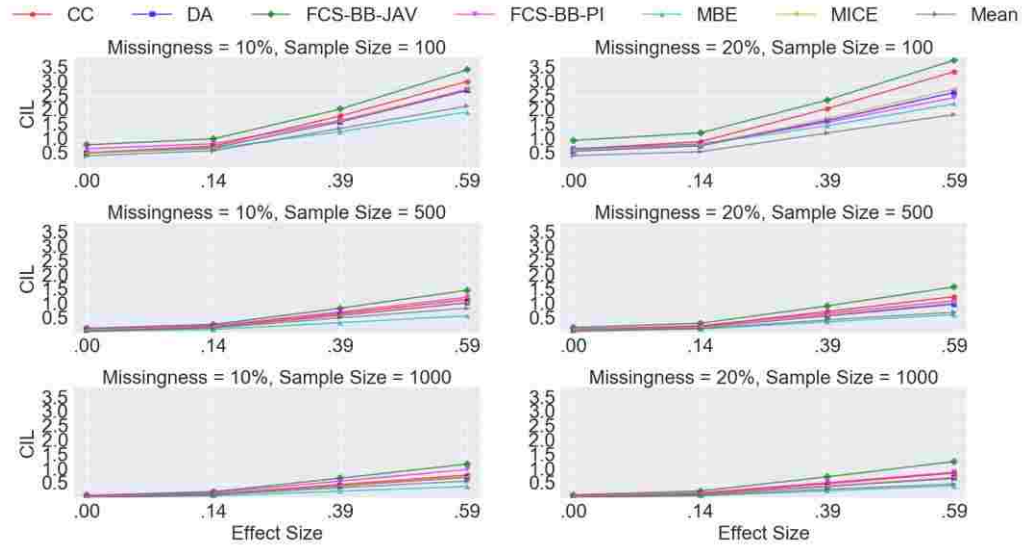


7. Model: Moderated Mediation, Mediator: Categorical, Endogenous: Continuous

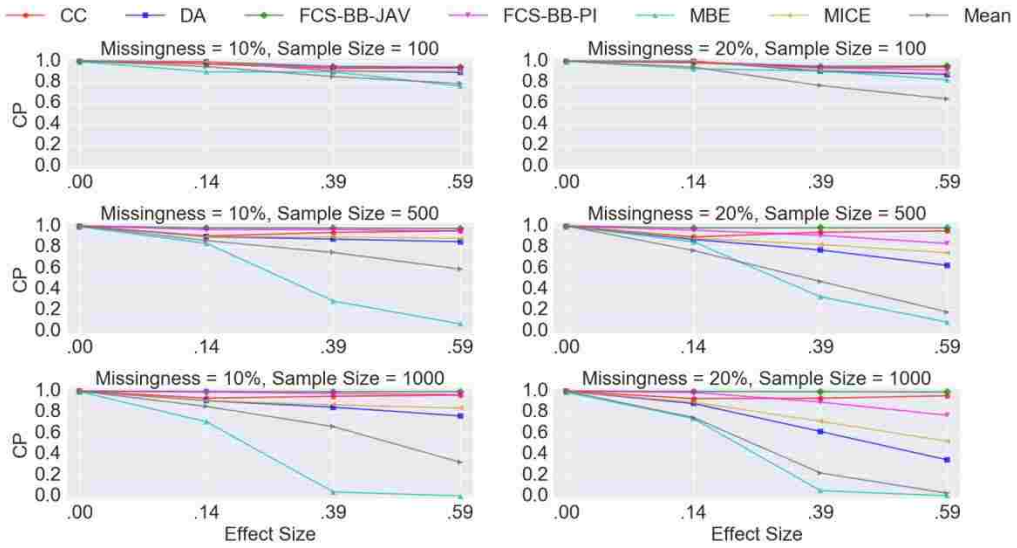
Empirical bias:



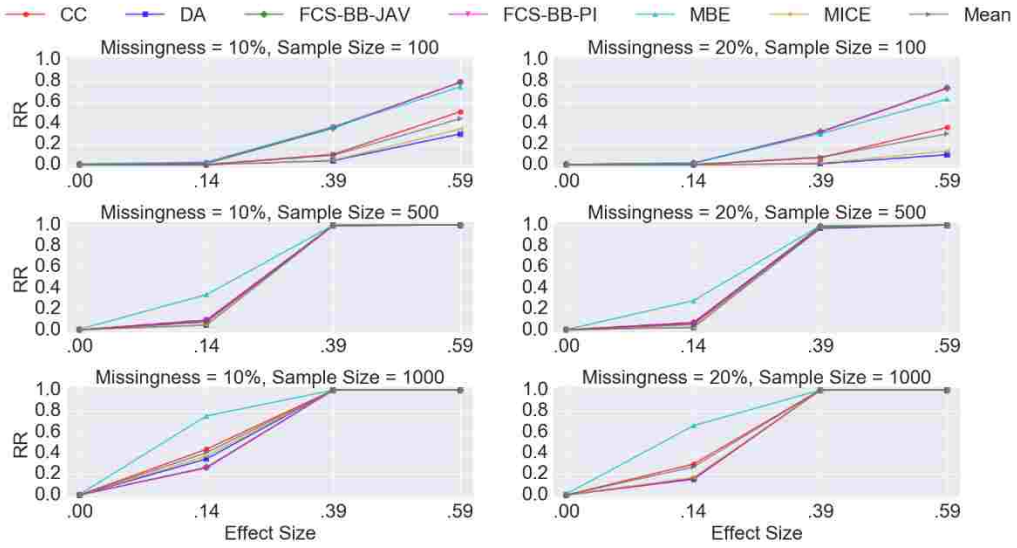
Confidence interval length:



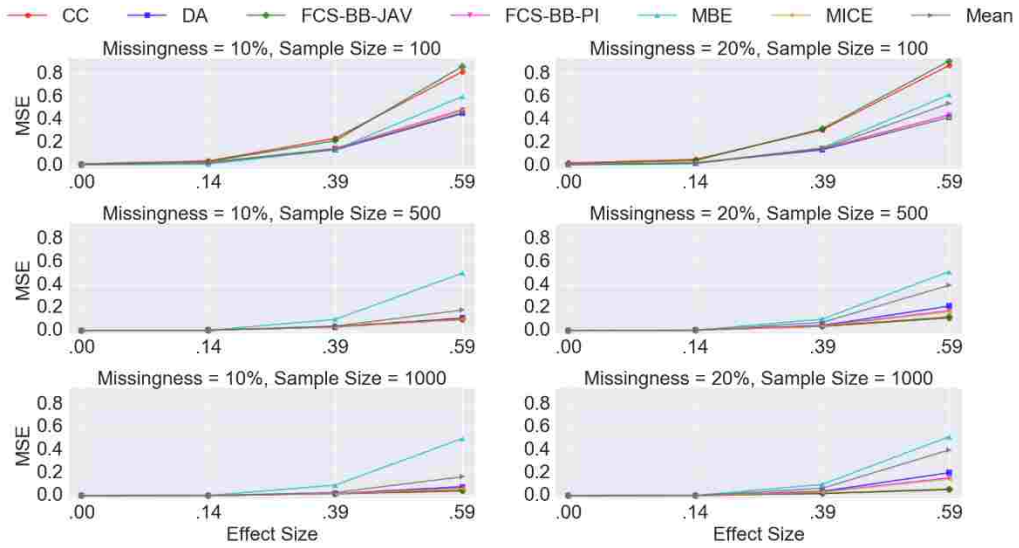
Coverage probability:



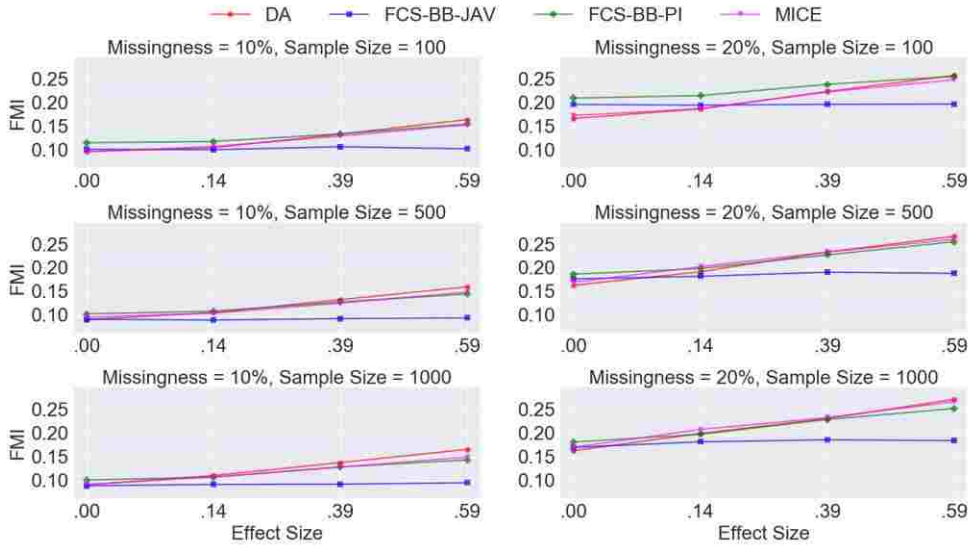
Rejection rate:



Mean squared error:

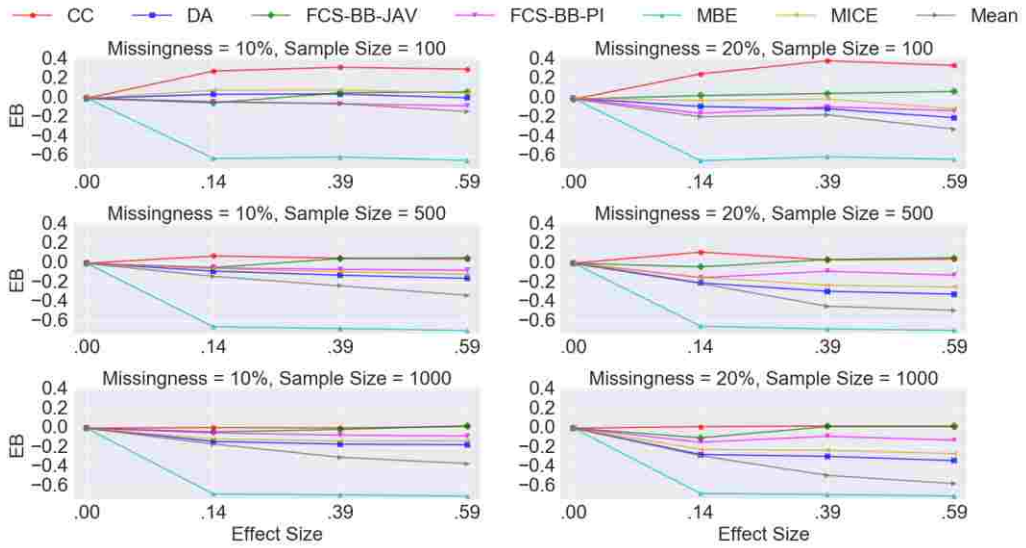


Fraction of missing information:

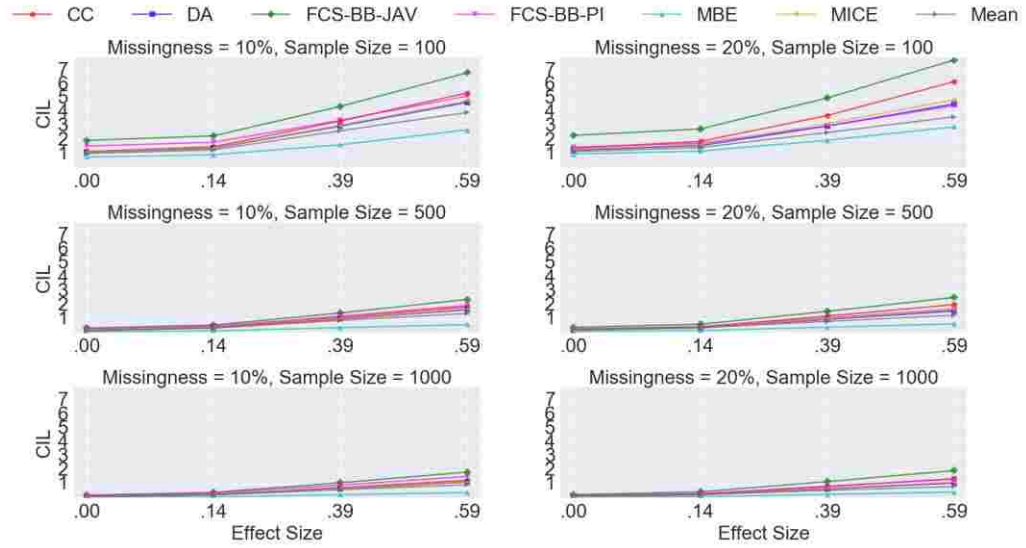


8. Moderated Model: Mediation, Mediator: Categorical, Endogenous: Categorical

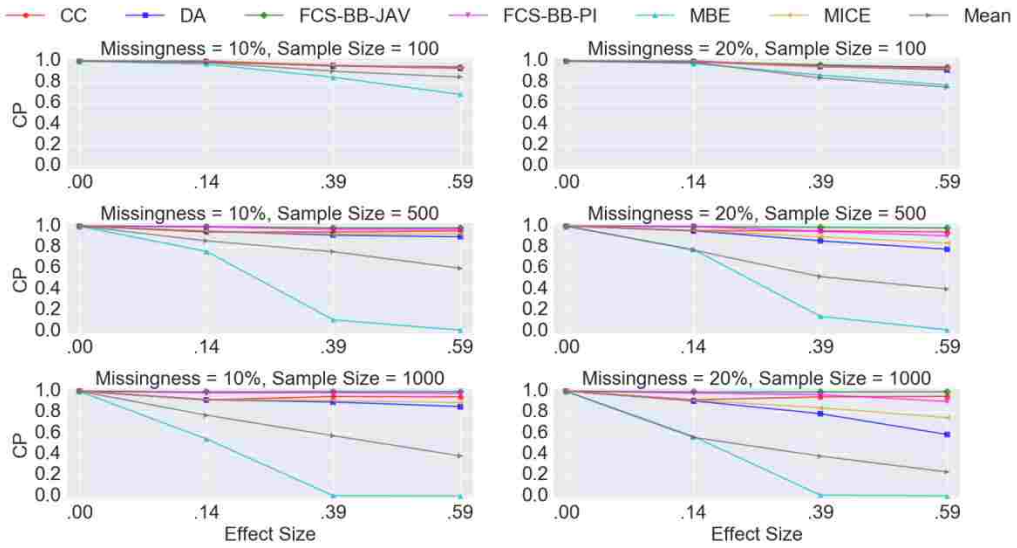
Empirical bias:



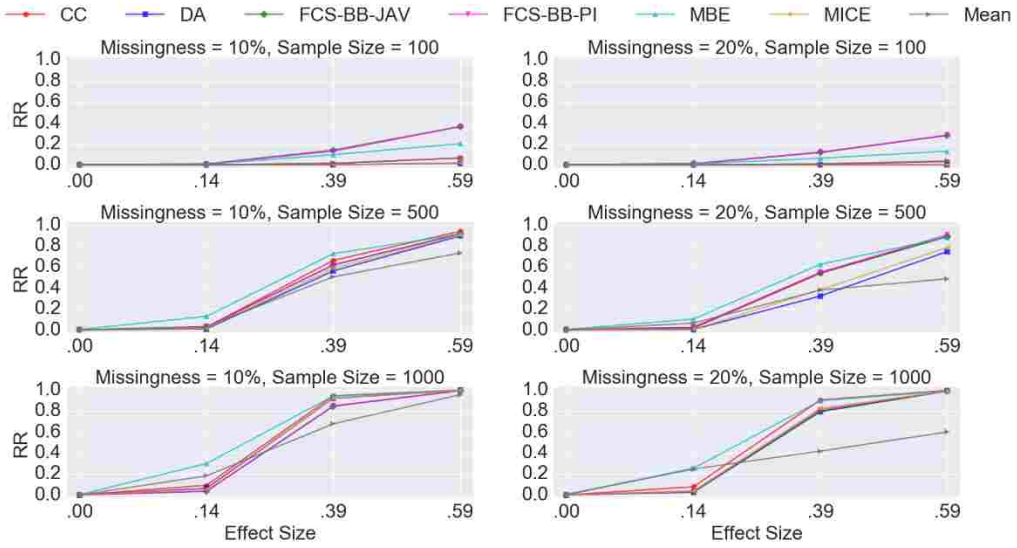
Confidence interval length:



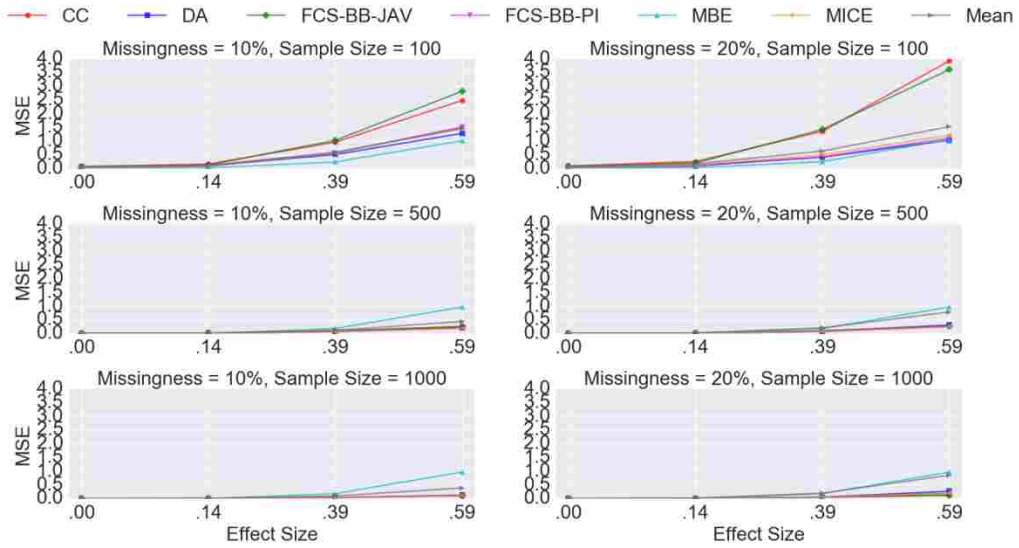
Coverage probability:



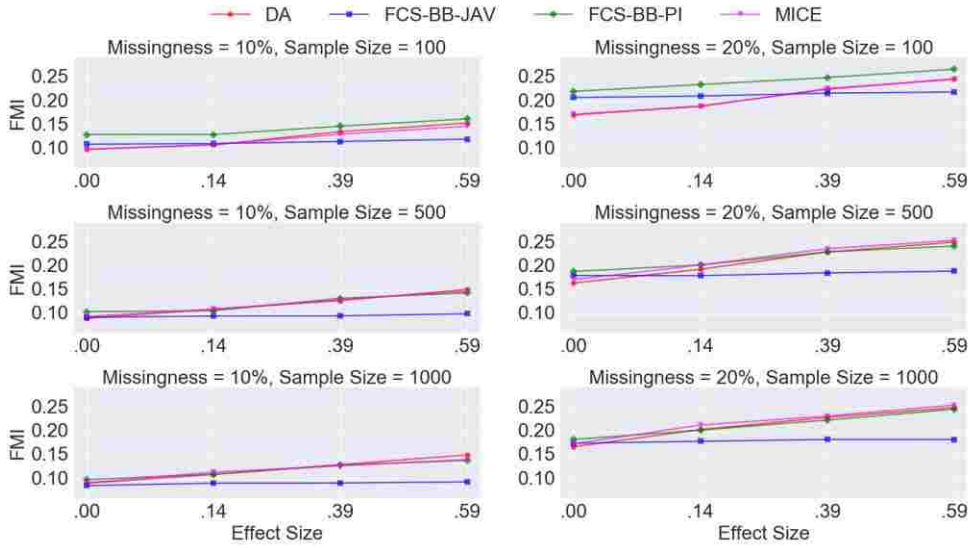
Rejection rate:



Mean squared error:



Fraction of missing information:



VITA

Robert J. Milletich II
Business Address: 250 Mills Godwin Building, Department of Psychology
Old Dominion University, Norfolk, VA 23529-0267
Phone: 757-683-4439
E-mail: rmill040@gmail.com

EDUCATION

Old Dominion University, Norfolk, VA — 2016
Ph.D. Applied Psychological Sciences

Old Dominion University, Norfolk, VA — 2015
M.S. Computational and Applied Mathematics

Old Dominion University, Norfolk, VA — 2012
M.S. Applied Experimental Psychology

Old Dominion University, Norfolk, VA — 2009
B.S. Psychology

PROFESSIONAL EXPERIENCE

Data Scientist at Booz Allen Hamilton (Norfolk, VA) — 2015 – Present
Provide expertise in statistical analysis, machine learning, programming, and data management for clients.

Statistician at Sentara Hospital (Virginia Beach, VA) — 2014 – 2015
Provided expertise in statistical consulting and modeling of patient data.

Instructor at Old Dominion University (Norfolk, VA) — 2010 – 2015
Developed original lectures and materials for courses in statistics at both undergraduate and graduate level.

SELECT PUBLICATIONS

Harrivel, A. R., Stephens, C. L., **Milletich, R. J.**, Heinich, C. M., Last, M. C., Napoli, N. J., et al. (2017). Prediction of cognitive states during flight simulation using multimodal psychophysiological sensing. *Paper to be presented at the American Institute of Aeronautics and Astronautics (AIAA) SciTech Conference January 2017 in Grapevine, Texas.*

Kelley, M. L., **Milletich, R. J.**, Hollis, B. F., Veprinsky, A., Robbins, A. T., & Snell, A. K. (*in press*). Social support and relationship satisfaction as moderators of the stress-mood- alcohol link in U.S. Navy members at predeployment. *Journal of Nervous and Mental Disease*.

Milletich, R. J., Diawara, N., & Jeng, A. (2015). Modeling of deaths due to Ebola virus disease outbreak in western Africa. *International Journal of Statistics in Medical Research,* 306-321.