UCLA
UCLA Electronic Theses and Dissertations

Title
A Hybrid EM Algorithm for Linear Two-Way Interactions with Missing Data

Permalink
https://escholarship.org/uc/item/4zv9h6kn

Author
Kim, Dale

Publication Date
2021

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

A Hybrid EM Algorithm

for Linear Two-Way Interactions

with Missing Data

A dissertation submitted in partial satisfaction

of the requirements for the degree Doctor of Philosophy

in Psychology

by

Dale S. Kim

2021

ABSTRACT OF THE DISSERTATION

A Hybrid EM Algorithm

for Linear Two-Way Interactions

with Missing Data

by

Dale S. Kim

Doctor of Philosophy in Psychology

University of California, Los Angeles, 2021

Professor Steven Paul Reise, Chair

We study an EM algorithm for estimating product-term regression models with missing data. The study of such problems in the likelihood tradition has thus far been restricted to an EM algorithm method using full numerical integration. However, under most missing data patterns, we show that this problem can be solved analytically, and numerical approximations are only needed under specific conditions. Thus we propose a hybrid EM algorithm, which uses analytic solutions when available and approximate solutions only when needed. The theoretical framework of our algorithm is described herein, along with two numerical experiments using both simulated and real data. We show that our algorithm confers higher accuracy to the estimation process, relative to the existing full numerical integration method. We conclude with a discussion of applications, extensions, and topics of further research.

The dissertation of Dale S. Kim is approved.

Craig K. Enders

Hongjing Lu

Han Du

Steven Paul Reise, Committee Chair

University of California, Los Angeles

2021

Contents

List of Figures

List of Tables

ACKNOWLEDGMENTS

I would like to express my gratitude to several of the faculty and staff in the Department of Psychology at UCLA. First, I would like to thank my advisor, Dr. Steven Reise, for his encouragement, support, and guidance, especially during my first few years. Your passion for research and contagious sense of humor was a source of endless positive energy throughout my graduate career. Second, I would like to thank Dr. Peter Bentler, for his constant support and willingness to discuss research ideas. Thank you very much for serving as a co-mentor on several of my academic pursuits, as well as your enthusiastic professional support. Finally, I would like to thank Lisa Lee, for being such an expert resource in navigating the logistical and bureaucratic intricacies of graduate division policy and the individually articulated Ph.D. program. I can't thank you enough for your assistance and support in technical matters, as well as your friendly ear in personal ones.

I would also like to express my gratitude for my colleagues during my time at the University of Washington. First, I would like to thank Dr. Kevin King, who served as my undergraduate research mentor. I would not be where I am today without you taking an interest in my early development as a researcher. Thank you for continuing to be a mentor to me in professional matters, as well as being a research colleague and friend. Second, I would like to thank Dr. Geoffrey Loftus and Dr. Kristine Louie, both of whom also provided me with many opportunities to nurture my nascent academic career. Finally, I would like to thank Dr. Connor McCabe, who is a close friend with whom research is a common interest, among a great many other things. I appreciate you trusting me enough to go and try to pull off some of my crazy research ideas. Here's to a great many successful future endeavors together, both inside and outside academia.

VITA

Education

Ph.D., Statistics   September 2019 - June 2022 (Expected)
University of California, Los Angeles, CA

Ph.D., Quantitative Psychology   September 2015 - December 2021 (Expected)
Minor: Computational Cognition
University of California, Los Angeles, CA

M.S., Statistics   September 2019 - June 2020
University of California, Los Angeles, CA

M.A., Quantitative Psychology   September 2015 - June 2019
University of California, Los Angeles, CA

B.S., Psychology   January 2011 - June 2014
University of Washington, Seattle, WA

Publications

Kim, D. S. & McCabe, C. M. (In press). The Partial Derivative Framework for Substantive Regression Effects. *Psychological Methods*.

McCabe, C. M., Halvorson, M. A., King, K. M., Cao, X., & Kim, D. S., (2021). Interpreting interaction effects in generalized linear models of nonlinear probabilities and counts. *Multivariate Behavioral Research*. Advance online publication. https://doi.org/10.1080/00273171.2020.1868966

Halvorson, M. A., McCabe, C. M., Kim, D. S., Cao, X. & King, K. M. (2021). Making sense of some odd ratios: A tutorial and improvements to present practices in reporting and visualizing quantities of interest for binary and count outcome models. *Psychology of Addictive Behaviors*. Advance online publication. https://doi.org/10.1037/adb0000669

McCabe, C. J., Kim, D. S. & King, K. M. (2018) Tools and Recommendations for the Visual Display of Interactions. *Advances in Methods and Practices in Psychological Science*, 1(2), 147-165. https://doi.org/10.1177/2515245917746792

King, K. M., Kim, D. S. & McCabe, C. J. (2018) Random responses inflate statistical estimates in heavily skewed addictions data. *Drug and Alcohol Dependence*, 183, 102-110. https://doi.org/10.1016/j.drugalcdep.2017.10.033

Kim, D. S., Reise, S. P. & Bentler, P. M. (2018) Identifying Aberrant Data in Structural Equation Models with IRLS-ADF. *Structural Equation Modeling*, 24(3), 343-358. https://doi.org/10.1080/10705511.2017.1379881

Kim, D. S., McCabe, C. J., Yamasaki, B. L., Louie, K. A. & King, K. M. (2018) Detecting Careless Responders with Infrequency Scales Using an Error Balancing Threshold. *Behavior Research Methods*, 50(5), 1960-1970. https://doi.org/10.3758/s13428-017-0964-9

Reise, S. P., Kim, D. S., Mansolf, M. & Widaman, K. F. (2016) Is the Bifactor Model a Better Model or is it Just Better at Modeling Implausible Responses? Application of Iteratively Reweighted Least Squares to the Rosenberg Self-Esteem Scale. *Multivariate Behavioral Research*, 51(6), 818-838. https://doi.org/10.1080/00273171.2016.1243461

Chapter 1

Introduction

We consider the problem of missing data in regression models with product-term predictors. In the psychological sciences, product-term regression models are widely used to test hypotheses pertaining to interactions (Aiken & West, 1991), moderation (Baron & Kenny, 1986), and/or conditional processes (Hayes, 2018). For example, these hypotheses may refer to the difference of an effect between two groups, the dependence of an outcome-predictor relationship on other variables, or the effect of two simultaneous symptoms above and beyond their constituent effects. A considerable amount of methodological research has been dedicated to interpreting these models (Dawson, 2014; McCabe, Kim, & King, 2018; Preacher, Curran, & Bauer, 2006), attesting to their importance and popularity.

However, the estimation of product-term regression models is complicated by the issue of missing data, which is particularly prevalent for data involving human subjects. It can arise from subject dropout, item non-response, logistical errors, or even serve as designed aspect of data collection (Graham, 2009; Raghunathan, 2004). If the mechanism of missing data meets certain conditions, this problem (or design) can be accommodated and consistent estimates can be obtained. In particular, we consider the situation under which the missing data mechanism is called *ignorable* (Schafer, 1997). Colloquially speaking, it means that the probability of an observation being missing does not depend on its own would-be realized value, and that the data value generating mechanism is distinct from the missingness generating mechanism.

Previous literature on this problem can be broadly cast into two categories: correctly

specified and misspecified characterizations of the joint distribution of the data. For product term-regression models, incorrect specification generally occurs for the product terms. The most common type of misspecification is naively assuming the product terms are jointly Gaussian along with their constituent factors (von Hippel, 2009). This has also been called the "just another variable" approach (Seaman, Bartlett, & White, 2012) as it treats the product term simply as another Gaussian random variable. However, this introduces a contradiction of distributional assumptions, as a product of Gaussians random variables cannot itself be Gaussian. While some studies have shown that there may be some conditions under which this method is reasonable (Enders, Baraldi, & Cham, 2014; Seaman et al., 2012), it is not guaranteed to provide unbiased estimates in general (Bartlett, Seaman, White, & Carpenter, 2015; Lüdtke, Robitzsch, & West, 2019; Zhang & Wang, 2017).

To address these issues, correctly specified methods have been developed. This is typically accomplished by a factorizing the joint distribution into a product with conditional distributions. In this way, it can be easier to correctly specify the constituent factored distributions, rather than the original joint distribution itself (e.g., Ibrahim, 1990). Hence, this technique can ensure compatibility between the substantive model of interest, and the overall joint distribution of the data (J. Liu, Gelman, Hill, Su, & Kropko, 2014). It has also been called factored regression modeling (Lüdtke et al., 2019), substantive model compatibility (Bartlett et al., 2015), and model-based handling (Enders, Du, & Keller, 2020).

The application of this technique however, has been largely focused on multiple imputation methods with Markov Chain Monte Carlo (MCMC) under a Bayesian paradigm (Kim, Sugar, & Belin, 2015; Lüdtke, Robitzsch, & West, 2020; Zhang & Wang, 2017). While MCMC methods are widely applicable, convergence assessment is notoriously difficult (Brooks & Roberts, 1998; Cowles & Carlin, 1996; Gelman & Rubin, 1996), and it is known for its slow and inefficient uses of Monte Carlo samples (Gelman & Rubin, 1992; Mossel & Vigoda, 2006). Further, improper applications of MCMC can result in unreliable and/or biased inferences (Cowles, Roberts, & Rosenthal, 1999; Flegal, Haran, &

Illustration of Numerical Integration
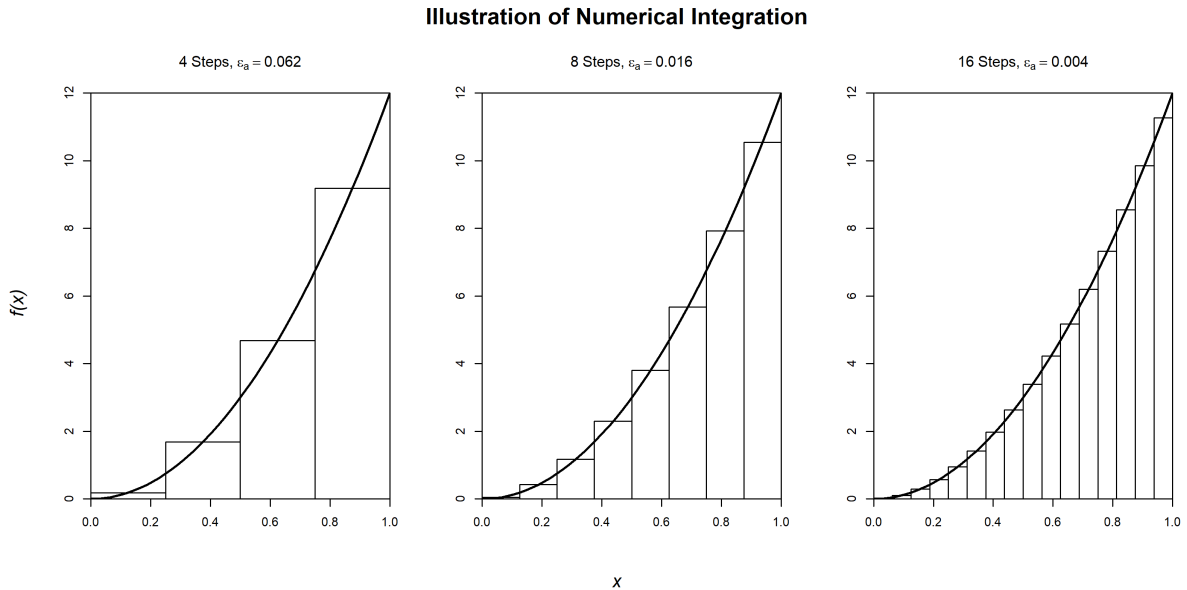
Figure 1.1: Illustration of Riemann midpoint numerical integration. The error of approximation is denoted as ε_a.

Jones, 2008). As such, maintaining the integrity of the procedure requires careful scrutiny and design.

To avoid such issues, we focus on correctly specified likelihood techniques, for which research has been scant. Currently, only an EM algorithm using full numerical integration has been proposed by Lüdtke et al. (2019). While their method is flexible and handles a variety of non-linear models, numerical integration is known to suffer in accuracy and computational complexity as the number of dimensions increase (Hinrichs, Novak, Ullrich, & Woźniakowski, 2014; Simonovits, 2003). Hence, the feasibility of this method is in question even when the number of variables is moderate. Indeed to date, this method has only been tested under very optimistic conditions. Specifically, only the simplest version of the product-term model (two predictors with one product) has been simulated, with only a single variable exhibiting missingness.

To illustrate this problem, consider the numerical integration example displayed in Figure 1.1. We numerically integrate an example function: $f(x) = 12x^2$ on the interval $[0, 1]$ using the Riemann midpoint method. Supposing it is prohibitive to integrate $f(x)$ analytically, we may instead integrate an approximation to $f(x)$ whose integration is much easier. The Riemann midpoint method approximates $f(x)$ with a step function, whose

3
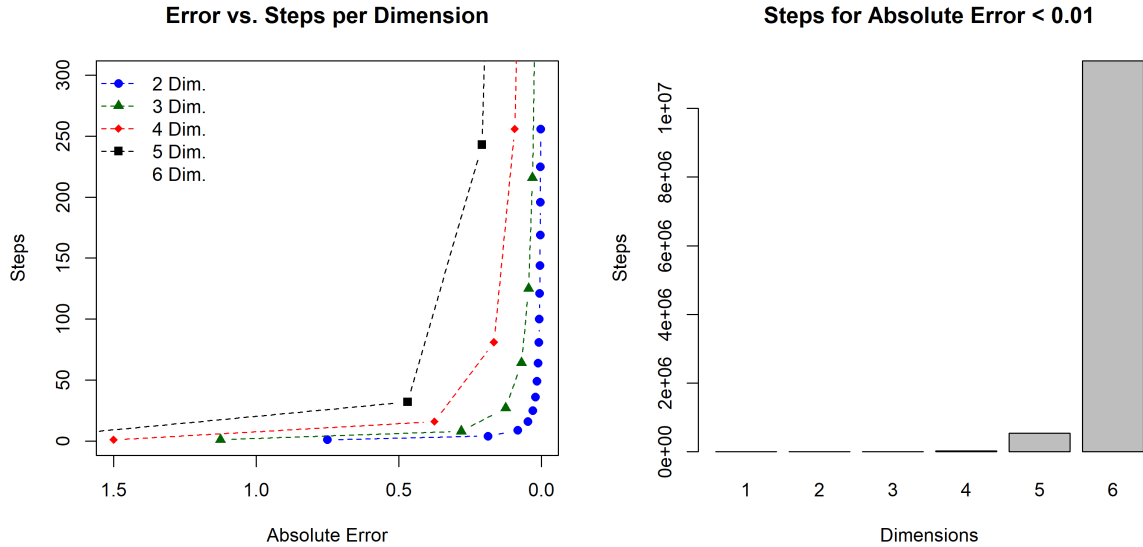
Figure 1.2: Demonstration of numerical integration computational complexity and approximation error.

integral is easily taken by summing the rectangles underneath each step. As illustrated by this one-dimensional example, the error of approximation (ε_a) reduces rather quickly as the number of interpolating steps increase.

A different story arises when we study the multi-dimensional case. Suppose we consider the multi-dimensional generalization of $f(x)$, by independently adding dimensions (i.e., $f(x_1, \ldots, x_p) = \sum_{j=1}^{p} 12x_j^2$). The number of steps required to achieve the same amount of accuracy increases exponentially. In Figure 1.2 (left) we show a contour plot of the error of approximation plotted against the number of steps. Each contour shows that exponentially more steps are required in order to achieve any given level of accuracy. For example, suppose we wanted to achieve a numerical approximation that is within 0.01 of the true value. We can see in Figure 1.2 (right) that as low as the 4-dimensional case, we require about two hundred thousand steps, and the 5-dimensional case requires almost 6.5 million.

Needless to say, numerical integration should not be undertaken unless absolutely necessary, and this is the premise of our proposed research. We propose a hybrid EM algorithm that obviates much of the required calculations done by numerical integration. Specifically, we will show that numerical integration is not necessary for most missing

data patterns, and demonstrate how to use analytic solutions in their place. These exact solutions will yield faster and more accurate estimates relative to their approximate counterparts. Therefore, this research has two main goals: (1) to develop the theoretical motivation of the hybrid EM algorithm and demonstrate its feasibility, and (2) to empirically study the speed and accuracy improvements conferred by the hybrid EM algorithm in practical data scenarios.

1.1 Model and Notation

Let $\boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denote a $p \times 1$ random vector of predictor variables. Then formulate a linear product-term model for a random scalar outcome variable Y as follows:

$$Y = \boldsymbol{d}(\boldsymbol{X})^T \boldsymbol{\beta} + \epsilon, \tag{1.1}$$

where $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$ is a scalar random variable of error terms, $\boldsymbol{\beta}$ is a $d \times 1$ vector of regression coefficients, and $\boldsymbol{d}(\boldsymbol{X})$ is an $d \times 1$ vector-valued design function as follows:

$$\boldsymbol{d}(\boldsymbol{X}) = \begin{bmatrix} 1 & \boldsymbol{X}^T & \overrightarrow{X_j X_k}^T \end{bmatrix}^T, \tag{1.2}$$

where $\overrightarrow{f(\boldsymbol{X})}$ denotes the vector of all unique permutations of $f(\boldsymbol{X})$ over the specified indices. In this case, $\overrightarrow{X_j X_k}$ is a vector whose elements are comprised of all unique permutations of $X_j X_k$ for all $j, k \in \{1, \ldots, p\}$ Hence, $\boldsymbol{d}(\boldsymbol{X})$ is a vector that augments $\boldsymbol{X}$ with a regression intercept and the product terms.

We note that there are two implicit assumptions with this model for the purposes of generality. First, we assume that the vector $\boldsymbol{X}$ contains the substantive model predictors as well as any desired auxiliary variables. Second, it is assumed the substantive model contains all possible two-way products among the variables in $\boldsymbol{X}$. To accommodate the fact that some auxiliary variables or product terms may not be desired in the substantive model, their β coefficient need only be constrained to zero.

5

1.2   Missing Data Assumptions

To recast the data from a predictor-outcome distinction to a missing-observed distinction, we use the following notation. Denote the an augmented data vector as $\boldsymbol{U} = \begin{bmatrix} Y & \boldsymbol{X}^T \end{bmatrix}^T$, which can be reordered as $\begin{bmatrix} \boldsymbol{U}_O^T & \boldsymbol{U}_M^T \end{bmatrix}^T$, where O is the index set of observed variables and M is the index set of missing variables. Further, $\mathbb{P}(\boldsymbol{U})$ is parameterized generally by a vector $\boldsymbol{\theta}$, for which we write $\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{U})$. Then, let $\boldsymbol{R} \in \{0,1\}^{p+1}$ be a binary random vector, which indicates whether the elements of $\boldsymbol{U}$ are observed, and has a probability distribution parameterized by the vector $\boldsymbol{\zeta}$.

We assume that the elements of $\boldsymbol{\theta}$ and $\boldsymbol{\zeta}$ are *distinct*, or that the joint space of $\boldsymbol{\theta}$ and $\boldsymbol{\zeta}$ is simply their Cartesian product $\boldsymbol{\theta} \times \boldsymbol{\zeta}$. Further, we assume the data are *missing at random* (MAR; Rubin, 1976):

$$\mathbb{P}_{\boldsymbol{\zeta}}(\boldsymbol{R}|\boldsymbol{U}) = \mathbb{P}_{\boldsymbol{\zeta}}(\boldsymbol{R}|\boldsymbol{U}_O). \tag{1.3}$$

Taking MAR in tandem with the distinctness of $\boldsymbol{\theta}$ and $\boldsymbol{\zeta}$, we say that the missing data mechanism is *ignorable* (Schafer, 1997).

1.3   The EM Algorithm for Missing Data

The EM algorithm is a two-step iterative procedure for obtaining parameter estimates for models with missing data (Dempster, Laird, & Rubin, 1977). The steps are as follows:

$\boldsymbol{E}$-Step. For any iteration t, define a Q-function given an intial parameter start value $\boldsymbol{\theta}^{(0)}$:

$$\begin{aligned} Q_{\boldsymbol{\theta}^{(t)}}(\boldsymbol{\theta}) &= \mathbb{E}_{\boldsymbol{\theta}^{(t)}} \left[\log \mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{U}) | \boldsymbol{U}_O \right] \\ &= \int_{\boldsymbol{u}_M} \log \mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{U}) \mathbb{P}_{\boldsymbol{\theta}^{(t)}}(\boldsymbol{u}_M | \boldsymbol{u}_O) \, d\boldsymbol{u}_M. \end{aligned} \tag{1.4}$$

$\boldsymbol{M}$-Step.   Maximize the Q-function with respect to $\boldsymbol{\theta}$ and set the result as $\boldsymbol{\theta}^{(t+1)}$:

$$\boldsymbol{\theta}^{(t+1)} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \, Q_{\boldsymbol{\theta}^{(t)}}(\boldsymbol{\theta}), \tag{1.5}$$

where integrating with respect to a vector is shorthand for multiple integration with respect to all elements of the vector (i.e., $\int_{\boldsymbol{z}} f(\boldsymbol{z}) \, d\boldsymbol{z} = \int_{z_1} \cdots \int_{z_p} f(\boldsymbol{z}) \, dz_p \cdots dz_1$, for $\boldsymbol{z} \in \mathbb{R}^p$). Hence, this is an iterative procedure that maximizes the expectation of the complete data log-likelihood, given the observed data. It is known to converge to a local maximum of the likelihood function under very general conditions (Wu, 1983).

1.4   Application to Product-Term Regression Models

For practical uses, the main task of applying the EM algorithm is setting up the Q-function. We do so for product-term regression models by characterizing the joint model of the data as follows:

$$\mathbb{P}(\boldsymbol{U}) = \mathbb{P}(Y|\boldsymbol{X})\mathbb{P}(\boldsymbol{X}), \tag{1.6}$$

where,

$$\mathbb{P}(\boldsymbol{X}) \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$
$$\mathbb{P}(Y|\boldsymbol{X}) \sim \mathcal{N}(\boldsymbol{d}(\boldsymbol{x})^T\boldsymbol{\beta}, \sigma_\epsilon^2). \tag{1.7}$$

Since $\mathbb{P}(\boldsymbol{U})$ factorizes into two Gaussian distributions, it can be written in exponential family form:

$$\mathbb{P}(\boldsymbol{U}) = \exp\left[\boldsymbol{\eta}(\boldsymbol{\theta})^T\boldsymbol{T}(\boldsymbol{U}) - A(\boldsymbol{\theta})\right], \tag{1.8}$$

which yields a Q-function of:

$$
\begin{aligned}
Q_{\boldsymbol{\theta}^{(t)}}(\boldsymbol{\theta}) &= \mathbb{E}_{\boldsymbol{\theta}^{(t)}}\left[\log \mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{U})|\boldsymbol{U}_O\right] \\
&= \mathbb{E}_{\boldsymbol{\theta}^{(t)}}\left[\boldsymbol{\eta}(\boldsymbol{\theta})^T\boldsymbol{T}(\boldsymbol{U}) - A(\boldsymbol{\theta})|\boldsymbol{U}_O\right] \\
&= \boldsymbol{\eta}(\boldsymbol{\theta})^T\mathbb{E}_{\boldsymbol{\theta}^{(t)}}\left[\boldsymbol{T}(\boldsymbol{U})|\boldsymbol{U}_O\right] - A(\boldsymbol{\theta}).
\end{aligned}
\tag{1.9}
$$

where $\boldsymbol{\theta}$ is the vector which contains the unique elements of $\{\boldsymbol{\beta}, \sigma^2, \boldsymbol{\mu}, \boldsymbol{\Sigma}\}$, $\boldsymbol{\eta}(\boldsymbol{\theta})$ is the vector of canonical parameters, and $A(\boldsymbol{\theta})$ is the log-partition function. Hence, constructing the Q-function amounts to deriving $\mathbb{E}\left[\boldsymbol{T}(\boldsymbol{U})|\boldsymbol{U}_O\right]$ per missing data pattern. It can be

shown that $\boldsymbol{T}(\boldsymbol{U})$ is:

$$\boldsymbol{T}(\boldsymbol{U}) = \left[\, Y \;\; Y\overrightarrow{X_j}^T \;\; Y^2 \;\; Y\overrightarrow{X_jX_k}^T \;\; \overrightarrow{X_j}^T \;\; \overrightarrow{X_jX_k}^T \;\; \overrightarrow{X_j^2}^T \;\; \overrightarrow{X_jX_kX_l}^T \;\; \overrightarrow{X_j^2X_k}^T \;\; \overrightarrow{X_jX_kX_lX_m}^T \;\; \overrightarrow{X_j^2X_kX_l}^T \;\; \overrightarrow{X_j^2X_k^2}^T \,\right]^T ,$$

(1.10)

recalling that $\overrightarrow{f(\boldsymbol{X})}$ denotes the vector of all unique permutations of $f(\boldsymbol{X})$ over the specified indices. For example, $\overrightarrow{X_jX_kX_lX_m}$ is the vector comprised of all unique permutations of $X_jX_kX_lX_m$, for all $j, k, l, m \in \{1, \ldots p\}$. The derivation of $\boldsymbol{T}(\boldsymbol{U})$ has been relegated to Appendix A.1. Note that these sufficient statistics only apply to product-term regression models. Different polynomial designs will imply a different $\boldsymbol{T}(\boldsymbol{U})$ vector. We also derive the maximizers of the Q-function in Appendix A.2.

Chapter 2

Missing Data Patterns

The theoretical motivation of this research is the derivation of analytic Q-functions under as many missing data patterns as possible. The form of $\boldsymbol{T}(\boldsymbol{U})$ may appear complex and the possible missing data patterns for $\mathbb{E}\left[\boldsymbol{T}(\boldsymbol{U})|\boldsymbol{U}_O\right]$ are combinatorially large. However, using an appropriate taxonomy, solutions for general classes of missing data patterns can be obtained and applied easily. The only types of missing data patterns (MDP) that need to be considered are as follows:

- **MDP 1:** Y is missing and $\boldsymbol{X}$ has any missingness pattern.

- **MDP 2:** Y is observed and $\boldsymbol{X}$ is patterned such that no product terms are fully missing.

- **MDP 3:** Y is observed and $\boldsymbol{X}$ is patterned such that one or more product terms are fully missing.

We will provide the methods of calculating $\mathbb{E}\left[\boldsymbol{T}(\boldsymbol{U})|\boldsymbol{U}_O\right]$ under each of these patterns. Specifically, we will show that analytic solutions exist for MDP 1 and MDP 2, and computational methods are only necessary for MDP 3.

2.1 Y is Missing, $\boldsymbol{X}$ has Any Missing Data Pattern

MDP 1 is concerned with the case when Y is missing, and $\boldsymbol{X}$ can take on any missingness pattern. We will show that all elements of $\mathbb{E}\left[\boldsymbol{T}(\boldsymbol{U})|\boldsymbol{U}_O\right]$ under this pattern can be

calculated by known functions of $\boldsymbol{\theta}$. Hence, the Q-function for this MDP can always be constructed analytically.

First, we consider the sufficient statistics that are solely a function of $\boldsymbol{X}$ and do not have a Y term. In Equation 1.10 these are the latter 8 (of 12) entries of $\boldsymbol{T}(\boldsymbol{U})$. Note that these entries are all products of the elements of $\boldsymbol{X}$ (e.g., $X_j X_k X_l X_m$ or $X_j^2 X_k^2$). Further, Y is missing in this MDP, so we have $\boldsymbol{U}_O = \boldsymbol{X}_O$. Thus, under MDP 1, we can more generally express the elements of $\mathbb{E}\left[\boldsymbol{T}(\boldsymbol{U})|\boldsymbol{U}_O\right]$ that only depend on $\boldsymbol{X}$ as:

$$\mathbb{E}\left[\prod_{i \in M} X_i^{a_i} \,\middle|\, \boldsymbol{X}_O\right], \tag{2.1}$$

where a_i are non-negative integers.

Our strategy will make use of two key facts. First, Gaussian random vectors are closed under conditioning, hence $\boldsymbol{X}_M|\boldsymbol{X}_O$ is itself a Gaussian random vector whose parameters are functions of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. Second, arbitrary product moments of random vectors can generally be found by appropriately differentiating their moment-generating function (Keener, 2010). Using the Gaussian moment-generating function in this way will remain a key tool for the rest of the theoretical development of this algorithm, so we will state the procedure in the following lemma.

Lemma 1 (Gaussian Product Moments). *Let $\boldsymbol{X}$ be a Gaussian random vector distributed as $\boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then any product moment of the form $\mathbb{E}[\prod_{i=1}^p X_i^{a_i}]$ can be expressed as a function of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$.*

Proof. This follows from a straightforward use of the Gaussian moment-generating function, which is:

$$M_{\boldsymbol{X}}(\boldsymbol{t}) = \exp\left(\boldsymbol{t}^T \boldsymbol{\mu} + \frac{1}{2}\boldsymbol{t}^T \boldsymbol{\Sigma} \boldsymbol{t}\right). \tag{2.2}$$

Then by the moment calculation property, any arbitrary product moment can be calculated with:

$$\frac{\partial^a}{\prod_{i=1}^p \partial t_i^{a_i}} M_{\boldsymbol{X}}(\boldsymbol{t})|_{\boldsymbol{t}=0} = \mathbb{E}\left[\prod_{i=1}^p X_i^{a_i}\right], \tag{2.3}$$

where $a = \sum_{i=1}^p a_i$ and all a_i take non-negative integer values. $\qquad\square$

Some simple didactic applications of the Gaussian moment-generating function have been included in Appendix A.3

From here, the expectation in the form of Equation 2.1 can be obtained by seeing that the parameters of $\boldsymbol{X}_M|\boldsymbol{X}_O$ are:

$$
\begin{aligned}
\boldsymbol{\mu}_c &= \boldsymbol{\mu}_M + \boldsymbol{\Sigma}_{MO}\boldsymbol{\Sigma}_O^{-1}(\boldsymbol{x}_O - \boldsymbol{\mu}_O) \\
\boldsymbol{\Sigma}_c &= \boldsymbol{\Sigma}_M - \boldsymbol{\Sigma}_{MO}\boldsymbol{\Sigma}_O^{-1}\boldsymbol{\Sigma}_{OM},
\end{aligned}
\tag{2.4}
$$

which follows from the well known parameterization of conditioning on Gaussian random vectors. Then by applying Lemma 1 on $\boldsymbol{X}_M|\boldsymbol{X}_O$, we obtain any of its product moments in terms of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. Thus, the latter 8 entries of $\mathbb{E}\left[\boldsymbol{T}(\boldsymbol{U})|\boldsymbol{U}_O\right]$ can be written in terms of $\boldsymbol{\theta}$ analytically.

Among the remaining 4 sufficient statistics, we turn our attention to Y, YX_j and YX_jX_k. Notice that we can consider a general expression that encapsulates the expectation of all three of these statistics by writing them as $\mathbb{E}[YX_j^aX_k^b|\boldsymbol{X}_O]$ for $a, b \in \{0, 1\}$. Then we can re-write this quantity as:

$$
\mathbb{E}[YX_j^aX_k^b|\boldsymbol{X}_O] = \mathbb{E}[\boldsymbol{d}(\boldsymbol{X})^T\boldsymbol{\beta}X_j^aX_k^b|\boldsymbol{X}_O],
\tag{2.5}
$$

which follows from applying the law of total probability and Bayes' rule (see Appendix A.4 for an explicit proof). Noting that $\boldsymbol{d}(\boldsymbol{X})^T\boldsymbol{\beta}$ is a linear combination of products of $\boldsymbol{X}$, we apply Lemma 1 with the linearity properties of the expectation operator to obtain $\mathbb{E}[YX_j^aX_k^b|\boldsymbol{X}_O]$ in terms of $\boldsymbol{\theta}$. Thus the solution for these expectations can be derived analytically as well.

Finally, the remaining expectation is $\mathbb{E}[Y^2|\boldsymbol{X}_O]$. This is derived as follows:

$$
\begin{aligned}
\mathbb{E}\left[Y^2|\boldsymbol{X}_O\right] &= \mathbb{E}\left[\mathbb{E}\left[Y^2|\boldsymbol{X}_M,\boldsymbol{X}_O\right]|\boldsymbol{X}_O\right] \\
&= \mathbb{E}\left[\mathrm{Var}(Y|\boldsymbol{X}) + \mathbb{E}\left[Y|\boldsymbol{X}\right]^2|\boldsymbol{X}_O\right] \\
&= \mathbb{E}\left[\sigma_\epsilon^2 + (\boldsymbol{\beta}^T\boldsymbol{d}(\boldsymbol{X}))^2|\boldsymbol{X}_O\right] \\
&= \sigma_\epsilon^2 + \mathbb{E}\left[\boldsymbol{\beta}^T\boldsymbol{d}(\boldsymbol{X})\boldsymbol{d}(\boldsymbol{X})^T\boldsymbol{\beta}|\boldsymbol{X}_O\right] \\
&= \sigma_\epsilon^2 + \mathbb{E}\left[\sum_{i,j}\beta_i\beta_j(\boldsymbol{d}(\boldsymbol{X})\boldsymbol{d}(\boldsymbol{X})^T)_{ij}\bigg|\boldsymbol{X}_O\right],
\end{aligned}
\tag{2.6}
$$

where $(\boldsymbol{d}(\boldsymbol{X})\boldsymbol{d}(\boldsymbol{X})^T)_{ij}$ refers to the (i,j)th element of $\boldsymbol{d}(\boldsymbol{X})\boldsymbol{d}(\boldsymbol{X})^T$. Once again, since each entry in the matrix $\boldsymbol{d}(\boldsymbol{X})\boldsymbol{d}(\boldsymbol{X})^T$ is a linear combination of products of $\boldsymbol{X}$, we can apply Lemma 1 and the linearity of expectation to write $\mathbb{E}[Y^2|\boldsymbol{X}_O]$ in terms of $\boldsymbol{\theta}$. Thus finally, we have shown that all entries of $\mathbb{E}[\boldsymbol{T}(\boldsymbol{U})|\boldsymbol{U}_O]$ can be written as analytic functions of $\boldsymbol{\theta}$ under MDP 1.

2.2 Y is Observed, $\boldsymbol{X}$ Admits No Fully Missing Products in $\boldsymbol{d}(\boldsymbol{X})$

MDP 2 considers the scenario where Y is observed and $\boldsymbol{X}$ is patterned such that no product terms are fully missing. Equivalently, we can say that $\boldsymbol{X}$ is patterned such that at least one X_j is observed in every product term. In this situation, $\boldsymbol{X}_M|Y,\boldsymbol{X}_O$ takes on a multivariate Gaussian distribution, and thus $\mathbb{E}[\boldsymbol{T}(\boldsymbol{U})|\boldsymbol{U}_O]$ can be completely solved analytically. To see why this is the case, let us re-write the analytical model in Equation 1.1 under the assumptions of MDP 2. First, note that we can separate terms by observed variables and missing variables:

$$
\begin{aligned}
Y &= \boldsymbol{d}(\boldsymbol{X})^T\boldsymbol{\beta} + \epsilon \\
&= \beta_0 + \sum_{j=1}^p \beta_j X_j + \sum_{j\neq k}\beta_{jk}X_jX_k + \epsilon \\
&= \beta_0 + \sum_{j\in O}\beta_j X_j + \sum_{j\in M}\beta_j X_j + \sum_{(j,k)\in O}\beta_{jk}X_jX_k + \sum_{j\in M, k\in O}\beta_{jk}X_jX_k + \epsilon.
\end{aligned}
\tag{2.7}
$$

12

Then, we can regard all $\boldsymbol{X}_O$ as constants and absorb them into the intercept and product-term coefficients as follows:

$$\tilde{\beta}_0 := \beta_0 + \sum_{j \in O} \beta_j X_j + \sum_{(j,k) \in O} \beta_{jk} X_j X_k$$

$$\tilde{\beta}_j := \beta_j + \beta_{jk} X_k, \text{ for } k \in O. \tag{2.8}$$

This allows us to re-write the model only in terms of the missing variables as:

$$Y = \tilde{\beta}_0 + \sum_{j \in M} \tilde{\beta}_j X_j + \epsilon, \tag{2.9}$$

from which we can write for any fixed $m \in M$:

$$X_m = \frac{Y - \tilde{\beta}_0 - \sum_{j \in M \setminus m} \tilde{\beta}_j X_j - \epsilon}{\tilde{\beta}_m}. \tag{2.10}$$

Thus, any X_m is a linear combination of other Gaussian random variables, therefore must be Gaussian itself. Hence $\boldsymbol{X}_M | Y, \boldsymbol{X}_O$ follows a multivariate Gaussian distribution. The derivation of the exact probability distribution $\mathbb{P}(\boldsymbol{X}_M | Y, \boldsymbol{X}_O)$ can be found in Appendix A.5.

Since Y is observed in this missing data pattern, $\mathbb{E}[\boldsymbol{T}(\boldsymbol{U}) | \boldsymbol{U}_O]$ only concerns product functions of $\boldsymbol{X}$. Hence, we only need to apply Lemma 1 to obtain these expectations, as $\boldsymbol{X}_M | Y, \boldsymbol{X}_O$ is a multivariate Gaussian. Thus, under MDP 2, $\mathbb{E}[\boldsymbol{T}(\boldsymbol{U}) | \boldsymbol{U}_O]$ can be written as a function of $\boldsymbol{\theta}$ and solved analytically.

2.3 Y is Observed, $\boldsymbol{X}$ Admits Fully Missing Products in $d(\boldsymbol{X})$

MDP 3 concerns the case where Y is observed and $\boldsymbol{X}$ is patterned such that product terms are fully missing. In this situation, the entries of $\mathbb{E}[\boldsymbol{T}(\boldsymbol{U}) | \boldsymbol{U}_O]$ may be difficult to derive analytically, or admit no closed-form. This can be seen from the following characterization

of X_m for $m \in M$ under MDP 3:

$$
\begin{aligned}
Y &= \boldsymbol{d}(\boldsymbol{X})^T \boldsymbol{\beta} + \epsilon \\
&= \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \sum_{j \neq k} \beta_{jk} X_j X_k + \epsilon \\
&= \beta_0 + \beta_m X_m + \sum_{j \neq m} \beta_j X_j + \sum_{j \neq m} \beta_{jm} X_j X_m + \sum_{j \neq k \neq m} \beta_{jk} X_j X_k + \epsilon \\
&= \beta_0 + \left(\beta_m + \sum_{j \neq m} \beta_{jm} X_j \right) X_m + \sum_{j \neq m} \beta_j X_j + \sum_{j \neq k \neq m} \beta_{jk} X_j X_k + \epsilon \\
\Rightarrow X_m &= \frac{Y - \beta_0 - \sum_{j \neq m} \beta_j X_j - \sum_{j \neq k \neq m} \beta_{jk} X_j X_k - \epsilon}{\beta_m + \sum_{j \neq m} \beta_{jm} X_j}.
\end{aligned}
\tag{2.11}
$$

From here we can see that X_m is a sum consisting of Gaussian ratio and product Gaussian ratio random variables. The moments or moment generation function of such random variables are difficult to derive and not readily available. Thus, this is the only missing data pattern for which numerical integration is used to obtain $\mathbb{E}[\boldsymbol{T}(\boldsymbol{U})|\boldsymbol{U}_O]$. That is, we approximate $\mathbb{E}[\boldsymbol{T}(\boldsymbol{U})|\boldsymbol{U}_O]$ with

$$
\begin{aligned}
\mathbb{E}[\boldsymbol{T}(\boldsymbol{U})|\boldsymbol{U}_O] &= \frac{\int_{\boldsymbol{u}_M} \boldsymbol{T}(\boldsymbol{u}_M, \boldsymbol{u}_O) \mathbb{P}(\boldsymbol{u}_M | \boldsymbol{u}_O) \, d\boldsymbol{u}_M}{\int_{\boldsymbol{u}_M} \mathbb{P}(\boldsymbol{u}_M | \boldsymbol{u}_O) \, d\boldsymbol{u}_M} \\
&= \frac{\mathbb{P}(\boldsymbol{u}_O)}{\mathbb{P}(\boldsymbol{u}_O)} \frac{\int_{\boldsymbol{u}_M} \boldsymbol{T}(\boldsymbol{u}_M, \boldsymbol{u}_O) \mathbb{P}(\boldsymbol{u}_M, \boldsymbol{u}_O) \, d\boldsymbol{u}_M}{\int_{\boldsymbol{u}_M} \mathbb{P}(\boldsymbol{u}_{Mg}, \boldsymbol{u}_O) \, d\boldsymbol{u}_M} \\
&\approx \frac{\sum_{g=1}^{G} \boldsymbol{T}(\boldsymbol{u}_{Mg}, \boldsymbol{u}_O) \mathbb{P}(\boldsymbol{u}_{Mg}, \boldsymbol{u}_O)}{\sum_{g=1}^{G} \mathbb{P}(\boldsymbol{u}_{Mg}, \boldsymbol{u}_O)},
\end{aligned}
\tag{2.12}
$$

where $\boldsymbol{u}_{Mg}$ is the gth grid point over the domain of $\boldsymbol{u}_M$ for numerical integration. Note that the purpose of dividing by $1 = \int_{\boldsymbol{u}_M} \mathbb{P}(\boldsymbol{u}_M | \boldsymbol{u}_O) \, d\boldsymbol{u}_M$ is to cancel out $\mathbb{P}(\boldsymbol{u}_O)$ from the numerator. This allows us to perform calculations in terms of $\mathbb{P}(\boldsymbol{u}_M, \boldsymbol{u}_O)$, rather than $\mathbb{P}(\boldsymbol{u}_M | \boldsymbol{u}_O)$, thus the latter need not be derived.

2.4 Summary of Results

We propose to construct a hybrid EM method that uses the analytic results derived in this Chapter for MDPs 1 and 2, and numerical integration for MDP 3. A summary of these

results are displayed in Table 2.1. This table essentially serves as a map that provides us with the method of calculation for any sufficient statistic's conditional expectation given any MDP.

$\mathbb{E}[\boldsymbol{T}(\boldsymbol{U})|\boldsymbol{U}_O]$	Y^2	Y	$Y\overrightarrow{X_j}^T$	$Y\overrightarrow{X_jX_k}^T$	$\mathbb{E}[\prod_{i=1}^p X_i^{a_i}]$
MDP 1	Eq. 2.6		Eq. 2.5		Lemma 1
MDP 2			Observed		
MDP 3					Numerical Integration

Table 2.1: Mosaic table of techniques to obtain $\mathbb{E}[\boldsymbol{T}(\boldsymbol{U})|\boldsymbol{U}_O]$ by MDP.

Chapter 3

Empirical Studies

Given that the hybrid EM algorithm minimizes the use of numerical approximations, we now investigate the impact this has on data analysis. This was done via two empirical studies: a traditional simulation study over a variety of conditions, and a smaller scale simulation study using real data with psychological measures. We compared empirical parameter bias, empirical mean square error, confidence interval coverage rates, and computation time between the hybrid EM method and full numerical integration.

3.1   Simulation Conditions

For the traditional simulation study, we sought to study estimator performance over a variety of realistic settings. These settings were

- Estimation method: hybrid EM (HYB) and full numerical integration (NI).

- Sample size (n): 50, 100, 250, 500, and 1000.

- Number of predictors (p): 3, 5, 10, and 15.

- Proportion of missingness (φ_{MIS}): 0.10, 0.20, and 0.30.

- Proportion among missing data from MDP 3 (φ_{MDP3}): 0, 0.05, 0.10, 0.15, 0.20, 0.25, 0.50, 0.75, and 1.

A fully factorial design was used for a total of $2 \times 5 \times 4 \times 3 \times 9 = 1080$ conditions. The number of simulations per condition was set to 100.

16

3.2 Data Generation

The parameters for $\boldsymbol{X}$ were generated in the following way:

$$\boldsymbol{\mu} \sim \mathcal{U}_p(-3, 3)$$
$$\boldsymbol{\Sigma} = \boldsymbol{DCD}, \tag{3.1}$$

where $\boldsymbol{D}$ is a diagonal matrix of variances, with the diagonal distributed as $\mathcal{U}_p(1, \sqrt{3})$ and $\boldsymbol{C}$ is a constant correlation matrix with a unity diagonal and off-diagonal entries of 0.3. Thus, each $\boldsymbol{\Sigma}$ is generated from the same underlying correlation matrix, but scaled accordingly with random variance entries. Once these parameters were drawn, $\boldsymbol{X}$ was sampled from $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

The parameters of the model were generated as follows. First, the number of product-terms in the model was set to half the number of predictors rounded down, i.e., $\left\lfloor \frac{p}{2} \right\rfloor$. Therefore the number of design variables in the model was

$$d = 1 + p + \left\lfloor \frac{p}{2} \right\rfloor, \tag{3.2}$$

including the intercept. $\boldsymbol{X}$ variables were chosen at random (uniformly) to form product-terms in $\boldsymbol{d}(\boldsymbol{X})$. Then, an adjusted R^2 parameter and $\boldsymbol{\beta}$ vector were simulated using:

$$R_a^2 \sim \mathcal{U}_1(0.1, 0.5)$$
$$\boldsymbol{\beta} \sim \mathcal{U}_d(-3, 3). \tag{3.3}$$

Given a sample of $\boldsymbol{x}$ vectors and a sampled $\boldsymbol{\beta}$, we can algebraically solve for a σ_ϵ^2 such that the drawn R_a^2 is achieved (see Appendix A.6 for details). This allows us to draw errors terms with $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$ and finally calculate the outcome with $Y = \boldsymbol{d}(\boldsymbol{X})^T \boldsymbol{\beta} + \epsilon$.

Once $\boldsymbol{X}$ and Y were generated, the observed data indicator $\boldsymbol{R}$ was generated under a MAR mechanism. This was done by randomly selecting a non-product variable in $\boldsymbol{X}$ (with uniform probability) to serve as an always-observed "missingness anchor" variable designated X_a, to determine missingness in all other variables. This ensured the MAR

assumption was always met. Using an intermediate latent propensity variable based on X_a, we assigned MDPs 1, 2, and 3 such that both φ_{MIS} and φ_{MDP3} were met according to the simulation condition and MDPs 1 and 2 were divided equally among the remaining missing cases. The exact mathematical details of this procedure can be found in Appendix A.7.

3.3 Estimation Methods

The EM algorithm outlined in Sections 1.3 and 1.4 was used to estimate the simulated datasets via the HYB and NI methods. The true model specification that was generated along with the data was assumed to be known during estimation. The HYB method used this information to categorize cases according to MDPs 1, 2, and 3 as defined in Chapter 2. For cases under MDPs 1 and 2, the E-step was calculated analytically according to the methods described in Sections 2.1 and 2.2, respectively. Numerical integration was used by the HYB method for cases under MDP 3, and by the NI method for all MDPs. The minimum number of integration points was set to be as close to 1,000, as possible. Following the existing `mdmb` packge (Robitzsch & Lüdtke, 2021) in the `R` language (R Core Team, 2021), we used Riemann midpoint numerical integration as exemplified in Figure 1.1. Confidence intervals were computed via non-parametric bootstrap (Efron & Tibshirani, 1994) using 1,000 bootstrap samples.

To minimize programming and computational variations between the two methods, both the HYB and NI methods were programmed using as much overlapping code as possible. This was done in the `R` language (R Core Team, 2021), primarily using the `Rcpp` (Eddelbuettel & François, 2011) and `RcppArmadillo` (Eddelbuettel & Sanderson, 2014) packages, and the `Armadillo` package (Sanderson & Curtin, 2016) for the `C++` language. We numerically verified the accuracy of the NI method through small-scale simulation comparisons with the existing `mdmb` package. These results can be found in Appendix A.8.

3.4 Performance Metrics

Estimator performance metrics were empirical bias, empirical mean square error (MSE), 95% confidence interval coverage rates, and computation time elapsed. For bias, MSE, and 95% confidence interval coverage rates, their component-wise counterparts were collected per coefficient per dataset. That is, we collected

$$
\begin{aligned}
\text{Deviation} &:= \hat{\beta}_j - \beta_j \\
\text{Square Error} &:= (\hat{\beta}_j - \beta_j)^2 \\
\text{Coverage} &:= I(\widehat{LL}_j \leq \beta_j \leq \widehat{UL}_j),
\end{aligned}
\tag{3.4}
$$

where $\widehat{LL}_j$ and $\widehat{UL}_j$ are the 2.5% and 97.5% empirical percentiles over the bootstrapped samples for $\hat{\beta}_j$ and $I(\cdot)$ is the indicator function. The average of these quantities provide the empirical bias, empirical MSE, and 95% confidence interval coverage rates, respectively. We analyzed the performance metrics by plotting the marginal means and conducting descriptive linear regressions of our outcomes as a function of simulation settings: method, n, p, φ_{MIS}, and φ_{MDP3}.

3.5 Results

Marginal plots of empirical bias, empirical MSE, coverage rates and time outcomes are displayed in Figures 3.1, 3.2, 3.3, and 3.4, respectively. The results of the linear regressions are displayed in Tables 3.1, 3.2, 3.3, and 3.4, also respectively. For the regressions, method was indicator coded ($\text{HYB} = 0, \text{NI} = 1$) and interacted with all predictors. The order variable was included as a control (via indicator coding), and is defined as the number of $\boldsymbol{X}$ variables that serve as the the coefficient's regressor. This is zero for intercepts, one for first-order terms (Order_1), and two for product-order terms (Order_2).

3.5.1 Empirical Bias

The marginal plot on empirical bias shows that both HYB and NI are generally unbiased across all conditions. The regression coefficients shows that bias does not vary linearly with any of the predictors for HYB, and only very small effects may exist for NI, if at all. Since the maximum likelihood estimates of $\boldsymbol{\beta}$ are unbiased, the missing data mechanism is MAR, and all models are guaranteed to be correctly specified, we would expect any bias to come from the method of estimation. Our results show very little empirical bias incurred by both methods.

3.5.2 Empirical Mean Square Error

The marginal plots show empirical MSE decreasing with n for both HYB and NI. This is congruent with maximum likelihood theory, as $\hat{\boldsymbol{\beta}}$ is known to be asymptotically consistent. However, the effect is more pronounced for HYB than NI, with the difference between the two being smaller at lower n. This is also reflected by the linear regression cofficients.

For p, the MSE increases in a nearly linear fashion, with NI being higher than HYB. This can be attributed to the fact that both HYB and NI implement numerical integration in their algorithms, which loses accuracy as the number of variables increase. Further, the difference between the two methods slightly increases as p increases, which is also reflected by the descriptive linear regressions. This can be accounted for by the fact that HYB uses less numerical integration than NI.

Both methods see an increase in MSE as φ_{MIS} increases in a nearly linear fashion. The variance of the MLE is well-known to increase with the proportion of missing data, also known as the missing information principle (Orchard & Woodbury, 1972). Since MSE is the squared bias plus the variance of the estimate, it follows that MSE would increase with φ_{MIS}. Once again, both the marginal plot and descriptive regression shows that the NI method has a higher rate of increase than HYB. This may be attributable to the fact that NI requires more numerical integration than HYB as φ_{MIS} increases.

The marginal plot of φ_{MDP3} shows that the MSE of NI is greater than (or equal to) that of HYB. The MSE of NI was fairly unaffected by changes in φ_{MDP3}, while the MSE

of HYB increased with φ_{MDP3}. The means are identical at $\varphi_{\text{MDP3}} = 1$, reflecting the fact that both methods use full numerical integration in this condition.

3.5.3 Coverage Rate

The marginal plots show increases in coverage rates toward 0.95 as n increases, with HYB being close to 0.95 than NI over all n. This is attributable to the fact that bootstrap approximations become more accurate as the number samples in the original data increase (Glivenko-Cantelli theorem; Tucker, 1959). However, with p, φ_{MIS}, and φ_{MDP3}, we observe little to no trends, with coverage rates at about 0.93 to 0.94, overall exhibiting slight under-coverage. Across all three variables, HYB is generally closer to the nominal rate than NI. The one exception is when $\varphi_{\text{MDP3}} = 1$ as both methods are identical.

3.5.4 Time

Both the marginal plots and regression coefficients show that the time to complete estimation increases with all four variables n, p, φ_{MIS}, and φ_{MDP3}. For the variables n, p, and φ_{MIS}, there is very little difference between HYB and NI. For φ_{MDP3} however, we see that HYB is substantially faster than NI at lower φ_{MDP3}, with the difference between the two decreasing to zero at $\varphi_{\text{MDP3}} = 1$.

3.5.5 Summary

A high-level summary of the marginal trends is as follows.

- Both methods exhibited very little bias.

- The MSE of HYB was less than or equal to NI across all conditions. The difference increased with n, p, and φ_{MIS}, and decreased with φ_{MDP3}.

- Under most conditions, HYB had coverage rates closer to the nominal rate than NI. Generally, the difference was within 1%, and both methods slightly under-covered in all conditions.

- The time difference between the methods show that HYB is generally faster than NI, though the difference is small. The exception is when $\varphi_{\mathrm{MDP3}} = 0$, where the difference is substantial, but shrinks to zero as φ_{MDP3} approaches 1.
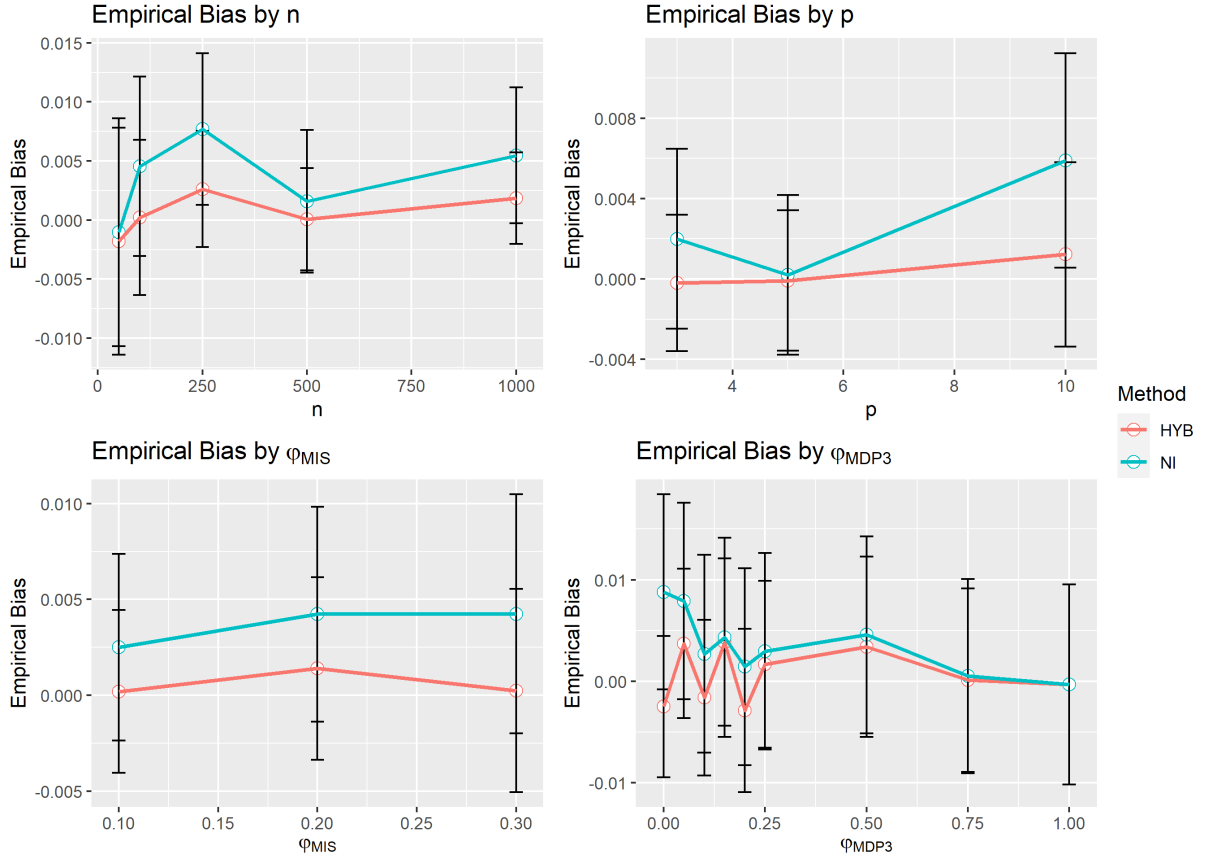
Figure 3.1: Marginal view of bias by n, p, φ_{MIS}, and φ_{MDP3}, stratified by method. Error bars indicate 95% confidence intervals.

Coefficient	Estimate	S. E.	t-value	p-value
Intercept	6.85×10^{-3}	7.50×10^{-3}	9.13×10^{-1}	≤ 0.361
n	1.27×10^{-6}	4.34×10^{-6}	2.93×10^{-1}	≤ 0.770
p	-1.30×10^{-4}	3.50×10^{-4}	-3.72×10^{-1}	≤ 0.710
ρ_{MIS}	-1.24×10^{-2}	1.85×10^{-2}	-6.72×10^{-1}	≤ 0.502
ρ_{MDP3}	1.29×10^{-3}	4.63×10^{-3}	2.79×10^{-1}	≤ 0.780
NI	-1.03×10^{-1}	1.06×10^{-2}	-9.70	≤ 0.001
NI $\times n$	3.29×10^{-5}	6.11×10^{-6}	5.37	≤ 0.001
NI $\times p$	9.85×10^{-3}	4.94×10^{-4}	2.00×10^{1}	≤ 0.001
NI $\times \rho_{\text{MIS}}$	1.90×10^{-1}	2.60×10^{-2}	7.30	≤ 0.001
NI $\times \rho_{\text{MDP3}}$	-6.58×10^{-2}	6.53×10^{-3}	-1.01×10^{1}	≤ 0.001
Order$_1$	-5.13×10^{-3}	5.82×10^{-3}	-8.83×10^{-1}	≤ 0.377
Order$_2$	-3.66×10^{-3}	6.19×10^{-3}	-5.92×10^{-1}	≤ 0.554
NI $\times$ Order$_1$	1.24×10^{-2}	8.20×10^{-3}	1.51	≤ 0.131
NI $\times$ Order$_2$	2.73×10^{-2}	8.73×10^{-3}	3.13	≤ 0.002

Table 3.1: Descriptive linear model of bias as a function of method, n, p, φ_{MIS}, φ_{MDP3}, and Order (control variable), along with their interactions with NI.

Figure 3.2: Marginal view of mean square error by n, p, φ_{MIS}, and φ_{MDP3}, stratified by method. Error bars indicate 95% confidence intervals.

Coefficient	Estimate	S. E.	t-value	p-value
Intercept	2.13	3.17×10^{-2}	6.71×10^{1}	≤ 0.001
n	-1.67×10^{-3}	1.83×10^{-5}	-9.12×10^{1}	≤ 0.001
p	1.96×10^{-1}	1.48×10^{-3}	1.33×10^{2}	≤ 0.001
ρ_{MIS}	2.70	7.80×10^{-2}	3.47×10^{1}	≤ 0.001
ρ_{MDP3}	6.54×10^{-1}	1.96×10^{-2}	3.35×10^{1}	≤ 0.001
NI	5.95×10^{-1}	4.47×10^{-2}	1.33×10^{1}	≤ 0.001
NI $\times n$	5.13×10^{-4}	2.58×10^{-5}	1.99×10^{1}	≤ 0.001
NI $\times p$	6.41×10^{-2}	2.09×10^{-3}	3.07×10^{1}	≤ 0.001
NI $\times \rho_{\mathrm{MIS}}$	1.69	1.10×10^{-1}	1.54×10^{1}	≤ 0.001
NI $\times \rho_{\mathrm{MDP3}}$	-7.39×10^{-1}	2.76×10^{-2}	-2.68×10^{1}	≤ 0.001
Order$_1$	-3.30	2.46×10^{-2}	-1.34×10^{2}	≤ 0.001
Order$_2$	-3.21	2.62×10^{-2}	-1.23×10^{2}	≤ 0.001
NI $\times$ Order$_1$	-1.23	3.47×10^{-2}	-3.56×10^{1}	≤ 0.001
NI $\times$ Order$_2$	-1.14	3.69×10^{-2}	-3.08×10^{1}	≤ 0.001

Table 3.2: Descriptive linear model of mean square error as a function of method, n, p, φ_{MIS}, φ_{MDP3}, and Order (control variable), along with their interactions with NI.
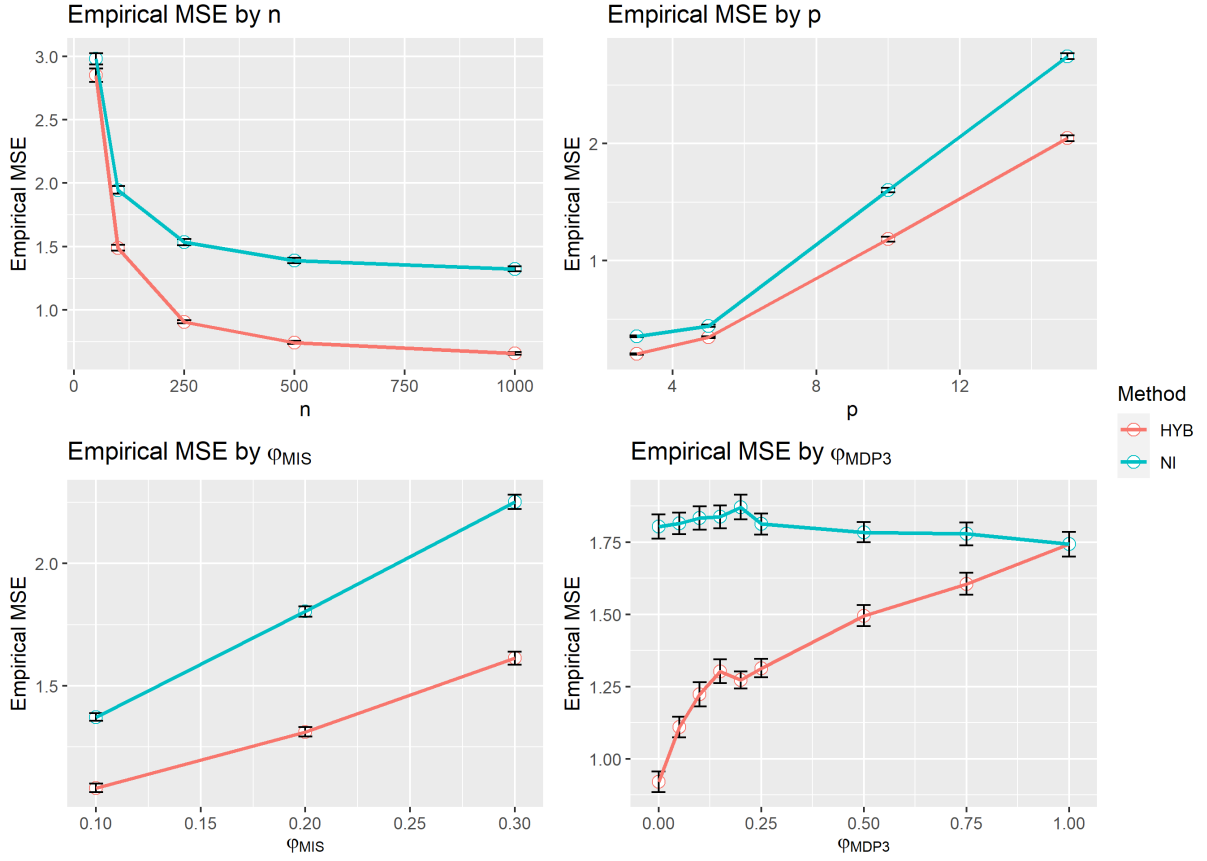
Figure 3.3: Marginal view of coverage rate by n, p, φ_{MIS}, and φ_{MDP3}, stratified by method. Error bars indicate 95% confidence intervals.

Coefficient	Estimate	S. E.	t-value	p-value
Intercept	9.23×10^{-1}	1.44×10^{-3}	6.39×10^{2}	≤ 0.001
n	1.73×10^{-5}	8.32×10^{-7}	2.08×10^{1}	≤ 0.001
p	6.78×10^{-4}	6.72×10^{-5}	1.01×10^{1}	≤ 0.001
ρ_{MIS}	2.44×10^{-2}	3.54×10^{-3}	6.88	≤ 0.001
ρ_{MDP3}	-2.22×10^{-4}	8.90×10^{-4}	-2.50×10^{-1}	≤ 0.803
NI	3.02×10^{-3}	2.04×10^{-3}	1.48	≤ 0.139
NI $\times n$	-9.16×10^{-6}	1.18×10^{-6}	-7.78	≤ 0.001
NI $\times p$	-6.99×10^{-4}	9.51×10^{-5}	-7.35	≤ 0.001
NI $\times \rho_{\mathrm{MIS}}$	-3.85×10^{-3}	5.01×10^{-3}	-7.68×10^{-1}	≤ 0.442
NI $\times \rho_{\mathrm{MDP3}}$	7.89×10^{-3}	1.26×10^{-3}	6.27	≤ 0.001
Order$_1$	-1.03×10^{-3}	1.12×10^{-3}	-9.20×10^{-1}	≤ 0.357
Order$_2$	-1.98×10^{-3}	1.19×10^{-3}	-1.66	≤ 0.097
NI $\times$ Order$_1$	-8.37×10^{-5}	1.58×10^{-3}	-5.30×10^{-2}	≤ 0.958
NI $\times$ Order$_2$	7.39×10^{-4}	1.68×10^{-3}	4.39×10^{-1}	≤ 0.661

Table 3.3: Descriptive linear model of coverage rate as a function of method, n, p, φ_{MIS}, φ_{MDP3}, and Order (control variable), along with their interactions with NI.
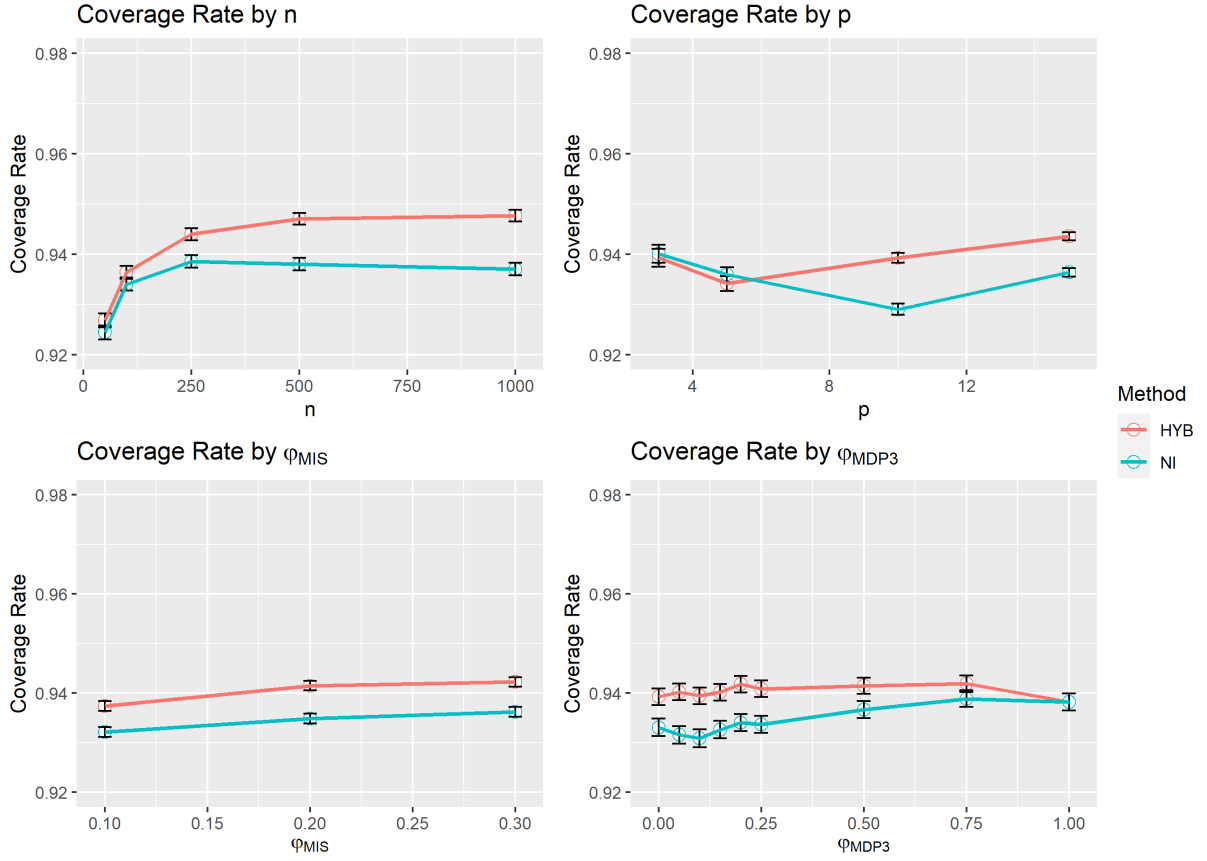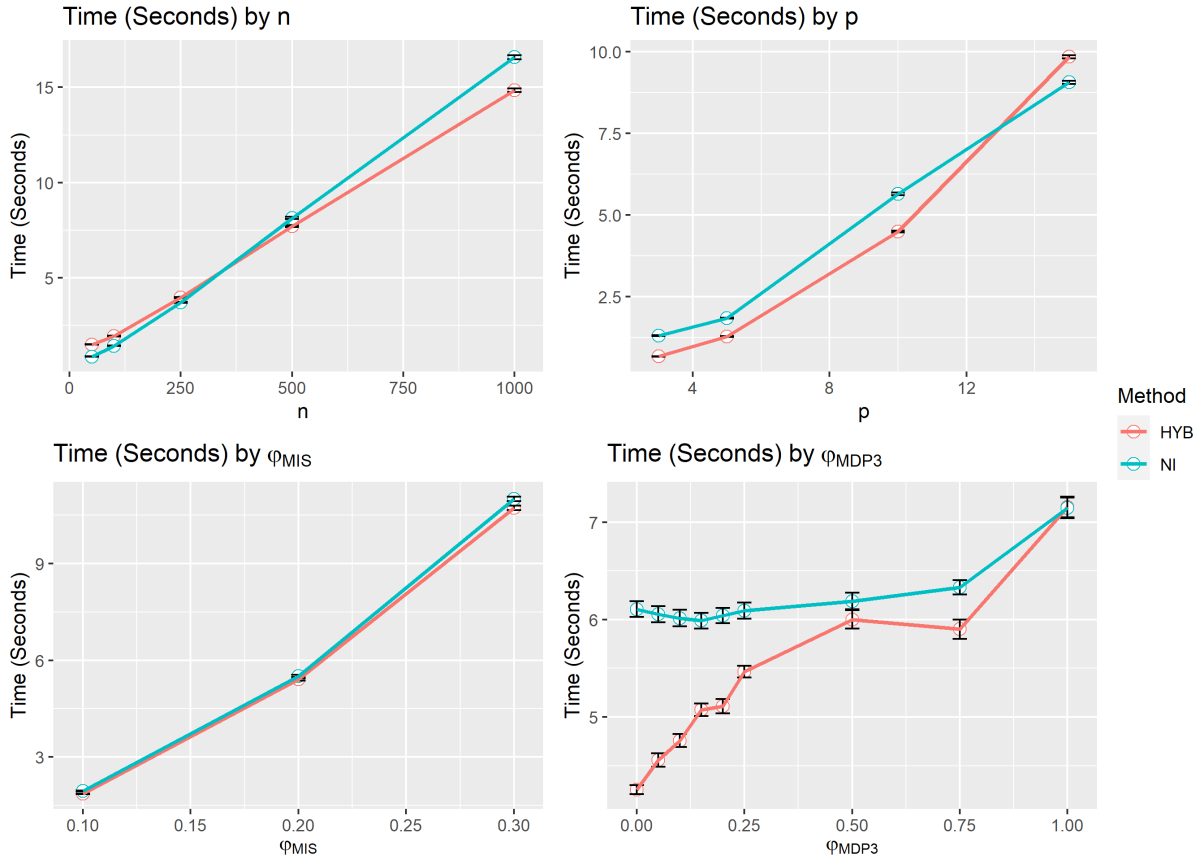
Figure 3.4: Marginal view of average time by n, p, φ_{MIS}, and φ_{MDP3}, stratified by method. Error bars indicate 95% confidence intervals.

Coefficient	Estimate	S. E.	t-value	p-value
Intercept	-1.75×10^1	5.33×10^{-2}	-3.29×10^2	≤ 0.001
n	1.42×10^{-2}	3.08×10^{-5}	4.62×10^2	≤ 0.001
p	8.29×10^{-1}	2.48×10^{-3}	3.33×10^2	≤ 0.001
ρ_{MIS}	4.44×10^1	1.31×10^{-1}	3.39×10^2	≤ 0.001
ρ_{MDP3}	1.48	3.29×10^{-2}	4.50×10^1	≤ 0.001
NI	7.01×10^{-1}	7.54×10^{-2}	9.29	≤ 0.001
NI $\times n$	2.51×10^{-3}	4.35×10^{-5}	5.78×10^1	≤ 0.001
NI $\times p$	-1.44×10^{-1}	3.51×10^{-3}	-4.09×10^1	≤ 0.001
NI $\times \rho_{\mathrm{MIS}}$	9.15×10^{-1}	1.85×10^{-1}	4.95	≤ 0.001
NI $\times \rho_{\mathrm{MDP3}}$	-7.87×10^{-1}	4.65×10^{-2}	-1.69×10^1	≤ 0.001
Order$_1$	-1.58×10^{-1}	4.13×10^{-2}	-3.83	≤ 0.001
Order$_2$	-2.08×10^{-1}	4.40×10^{-2}	-4.74	≤ 0.001
NI $\times$Order$_1$	1.23×10^{-1}	5.84×10^{-2}	2.11	≤ 0.035
NI $\times$Order$_2$	1.69×10^{-1}	6.22×10^{-2}	2.72	≤ 0.007

Table 3.4: Descriptive linear model of time as a function of method, n, p, φ_{MIS}, φ_{MDP3}, and Order (control variable), along with their interactions with NI.

3.6 Real Data Study

For our second simulation study, datasets were generated by bootstrapping samples using a real dataset as a basis, where simulated missingness was artificially inserted. We analyzed measures of psychopathology from the Adolescent Brain Cognitive Development (ABCD) Study (https://abcdstudy.org). The ABCD is a large, multi-site study, whose data are publicly available (Volkow et al., 2018), which was approved by the institutional review boards of the participating sites (Clark et al., 2018). To avoid potential clustering effects by site and to reduce the sample size to a more realistic scale, one site was selected randomly with uniform probabity to provide the basis of our data ($n = 604$).

We considered a linear model of conduct disorder (CD) as a function of attention deficit hyperactivity disorder (ADHD), depression (DEP), their product-term (ADHD $\times$ DEP), controlled for by anxiety (ANX) and oppositional defiant disorder (ODD). Previous work has shown comorbidity among these variables (Angold, Costello, & Erkanli, 1999; Jensen, Martin, & Cantwell, 1997). Measures were taken using summary scores of the child behavior checklist (CBCL; Achenbach & Rescorla, 2001). The selected site incidentally had no missing data on these variables.

For each bootstrapped data set, missingness was inserted using the same MAR generating procedure as the traditional simulation study. We fixed $\varphi_{\text{MIS}} = 0.3$ and $\varphi_{\text{MDP3}} = 1/3$, which set equal proportions among all MDPs. For performance metrics, we collected deviation and square error as in Equation 3.4 from the HYB and NI methods. The least squares estimates of the original non-missing data set was considered as the true parameters.

Boxplots of the results are displayed in Figure 3.5, and averages are shown in Table 3.5. On average, the HYB method had biases closer to zero than NI across all parameter estimates. Likewise, the HYB method also had MSEs closer to zero than NI across all parameter estimates except for the product-term coefficient, where NI had a marginally smaller MSE by 3.9×10^{-4}.
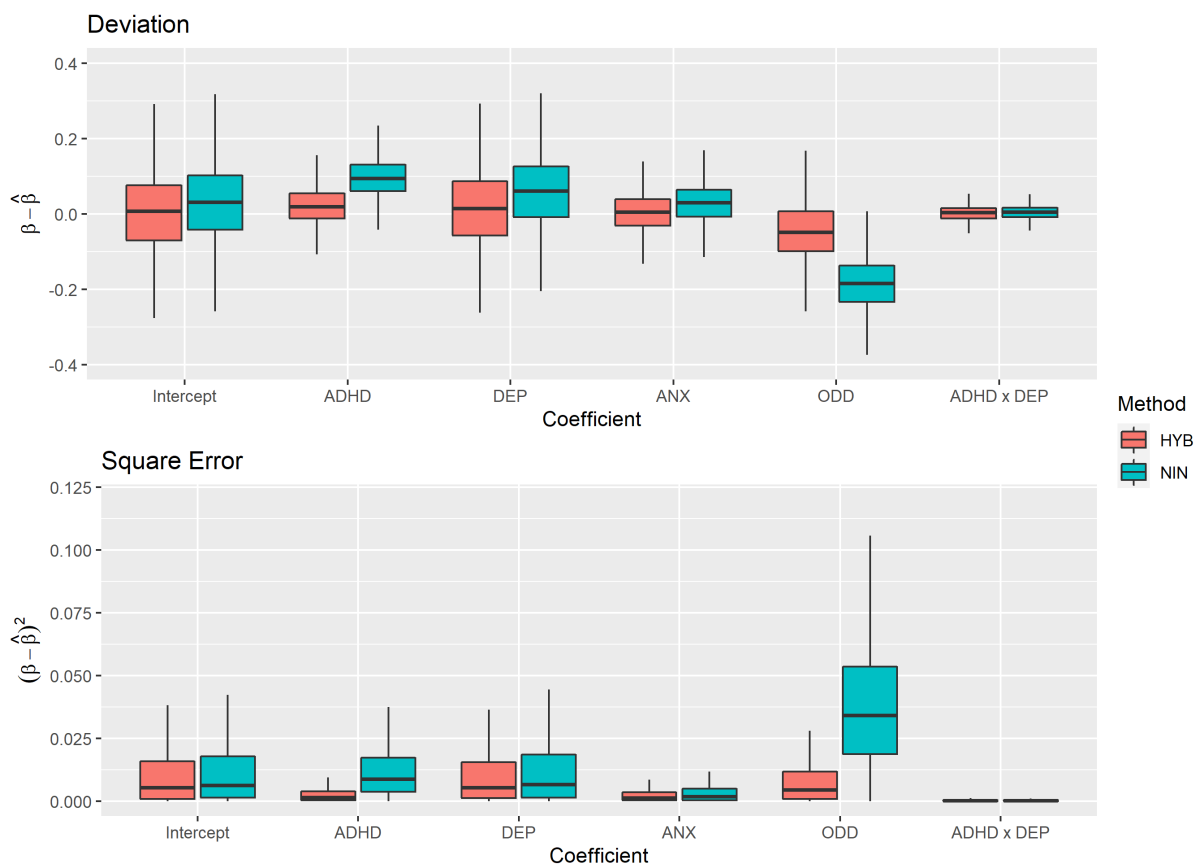
Figure 3.5: Boxplots of deviations and square errors of the parameter estimates in the real data simulation.

	Bias		MSE	
Coefficient	HYB	NI	HYB	NI
Intercept	0.0074	0.0299	0.0123	0.0131
ADHD	0.0217	0.0974	0.00322	0.0122
DEP	0.0141	0.0603	0.0117	0.0138
ANX	0.0048	0.0293	0.00286	0.00382
OPP	-0.0463	-0.184	0.00861	0.0393
ADHD $\times$ DEP	0.00161	0.00317	0.000383	0.000344
Means	0.000543	0.00603	0.00651	0.0138

Table 3.5: Bias and MSE averaged within method and coefficient.

Chapter 4

Discussion

In this research, we sought to improve the EM algorithm for linear models with two-way product terms by deriving analytic E-steps for large classes of MDPs. These derivations were used to develop a hybrid approach to the EM algorithm, where analytic E-steps were used whenever possible and numerical integration was used otherwise. Through two simulation studies, we showed that the hybrid approach outperformed, or performed as well as, the existing method of full numerical integration on several performance metrics. The difference between the two methods diminished as the proportion of missing data that require numerical integration for the hybrid approach increased (φ_{MDP3}). Otherwise, the degree to which the hybrid method out performed full numerical integration grew with n, p, and φ_{MIS}.

4.1 Limitations and Future Extensions

The current research focused on regression models with two-way product-terms. While this is a very common model in the psychological sciences, occasionally higher-power polynomials and multi-way product-terms are called for. The approach of deriving analytic E-steps for two-way product term models generalizes in a straightforward manner to all polynomial forms. Higher powers of predictors are readily accommodated by Lemma 1, since the Gaussian moment-generating function can be used to obtain any non-negative integer moment. Arbitrarily large product-terms can also be accommodated, although more limited due to the nature of MDPs 2 and 3. MDP 2 relies on the fact that any

missing X_j is linear with respect to all other missing X_{-j} conditional on the observed data. This holds as long as both factors of any product-term are not missing simultaneously. For product-terms with more than two factors, the conditional linearity strategy could only accommodate one factor of the product-term to be missing. Otherwise, the missing predictor would be conditionally non-linear and numerical integration would be required. Therefore, as the number of factors in a product-term increases, it becomes less likely that the conditional linearity strategy can be used, diminishing the advantages of the hybrid approach.

Other extensions can be made to discrete predictors. Ordinal data can be accommodated using existing methods for estimating polychoric correlations assuming the data are a result of discretizing Gaussian variables (Olsson, 1979). For nominal data, techniques from Gaussian mixture models may also be incorporated into estimation. Such extensions would greatly advance the applicability of the hybrid method as both ordinal and nominal data are commonly used.

For numerical integration, we focused exclusively on the Riemann midpoint method. Both the speed and accuracy of numerical integration may be improved via other approximating functions (e.g., Q. Liu & Pierce, 1994), and/or by adaptive methods (e.g., Rabe-Hesketh, Skrondal, & Pickles, 2002). Monte Carlo methods of integration also exist, including Gibbs and Metropolis-Hasting variants (Levine & Casella, 2001; Wei & Tanner, 1990), which can handle high-dimensions more efficiently. However, these methods require more samples as the number of EM iterations progress, so that Monte Carlo error does not interfere with EM convergence assessment (Booth & Hobert, 1999). The investigation of these issues could lead to improvements in both the hybrid and numerical integration approaches.

A final area of further research pertain to implementation of standard errors. While we employed bootstrap standard errors in the current research, other methods exist that utilize numerical differentiation of either the EM operation (Meng & Rubin, 1991) or the observed Fisher score function (Jamshidian & Bentler, 1999). While these methods are able to directly approximate the observed-data Hessian matrix, their computational

complexity is on the order of the number of parameters squared (Jamshidian & Jennrich, 2000). This makes computation time a practical issue for larger models. Direct analytic derivation of the observed-data Hessian is also possible, but often deemed too tedious to undertake (Hartley, 1958). However, this may be feasible with a proper MDP taxonomy, such as MDPs 1, 2, and 3 described in Chapter 2, or with symbolic computation methods.

4.2 Implications and Conclusions

An interesting application of the hybrid approach pertains to planned missingness designs. In planned missingness designs, there is a large degree of control over missing data patterns. The benefits of the hybrid approach can be fully realized if the missing data patterns are manipulated to completely avoid MDP 3. This can be done by designing a planned missingness study to contain a block of items to be delivered to all participants (a non-missing block, e.g., item set X in a 3-form design; Graham, Taylor, Olchowski, & Cumsille, 2006). Then MDP 3 can be completly avoided by assigning one of the factors of each product-term to the non-missing data block.

Ultimately, the accuracy of scientifically inferential quantities, such as parameter estimates and coverage intervals are fundamental aspects for the applied use of any methodology. Improvements in these areas directly impact the ability to detect effects accurately. The wide use of product-term regression models and the prevalence of missing data issues makes our hybrid EM method highly applicable to problems in the psychological sciences. A final step of this research would be to develop and disseminate a user-friendly software implementation, allowing accessible use of the hybrid EM method to the end-user researcher.

Appendix

A.1 Sufficient Statistics of the Joint Model

In this section we derive the vector of sufficient statistics $\boldsymbol{T}(\boldsymbol{U})$. Noting that $\mathbb{P}(\boldsymbol{U}) = \mathbb{P}(Y|\boldsymbol{X})\mathbb{P}(\boldsymbol{X})$, and $\mathbb{P}(Y|\boldsymbol{X})$ and $\mathbb{P}(\boldsymbol{X})$ are both exponential family distributions, we can derive the sufficient statistics in a piecewise manner. It is well known that $\boldsymbol{T}(\boldsymbol{X}) = \begin{bmatrix} \boldsymbol{X}^T & \overrightarrow{X_j X_k}^T & \overrightarrow{X_j^2}^T \end{bmatrix}^T$ since $\mathbb{P}(\boldsymbol{X})$ is Gaussian. Thus, what remains is determining the entries of $\boldsymbol{T}(\boldsymbol{U})$ that are contributed by $\mathbb{P}(Y|\boldsymbol{X})$. Since $\mathbb{P}(Y|\boldsymbol{X})$ is also Gaussian, we can determine its sufficient statistics by examining the exponential as follows:

$$
\begin{aligned}
\mathbb{P}(Y|\boldsymbol{X}) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}\left(Y - \boldsymbol{d}(\boldsymbol{X})^T\boldsymbol{\beta}\right)^2\right) \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}\left(Y^2 - 2Y\boldsymbol{d}(\boldsymbol{X})^T\boldsymbol{\beta} + \boldsymbol{\beta}^T\boldsymbol{d}(\boldsymbol{X})\boldsymbol{d}(\boldsymbol{X})^T\boldsymbol{\beta}\right)\right).
\end{aligned}
\tag{1}
$$

For which we have Y^2, $Y\boldsymbol{d}(\boldsymbol{X})^T$, and $\boldsymbol{d}(\boldsymbol{X})\boldsymbol{d}(\boldsymbol{X})^T$. Y^2 is self-explanatory and $Y\boldsymbol{d}(\boldsymbol{X})^T = \begin{bmatrix} Y & Y\boldsymbol{X}^T & Y\overrightarrow{X_j X_k}^T \end{bmatrix}^T$. Then what remains for us to analyze is the $\boldsymbol{d}(\boldsymbol{X})\boldsymbol{d}(\boldsymbol{X})^T$ term. By simply checking the elements of this outer product we obtain:

$$
\boldsymbol{d}(\boldsymbol{X})\boldsymbol{d}(\boldsymbol{X})^T = \begin{bmatrix} 1 & X^T & \overrightarrow{X_j X_k}^T \\ X & XX^T & X\overrightarrow{X_j X_k}^T \\ \overrightarrow{X_j X_k} & \overrightarrow{X_j X_k}X^T & \overrightarrow{X_j X_k}\overrightarrow{X_j X_k}^T \end{bmatrix}.
\tag{2}
$$

Then we can see that:

$$\text{vec}(\overrightarrow{X_j X_k} X^T) = \boldsymbol{P}_1 \left[\overrightarrow{X_j X_k X_l}^T \quad \overrightarrow{X_j^2 X_k}^T \right]^T$$

$$\text{vech}(\overrightarrow{X_j X_k} \overrightarrow{X_j X_k}^T) = \boldsymbol{P}_2 \left[\overrightarrow{X_j X_k X_l X_m}^T \quad \overrightarrow{X_j^2 X_k X_l}^T \quad \overrightarrow{X_j^2 X_k^2}^T \right]^T,$$

(3)

where $\text{vec}(\cdot)$ and $\text{vech}(\cdot)$ are the vectorization and half-vectorization functions, respectively. $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$ are permutation matrices that yields the appropriate ordering. Finally, we combine the sufficient statistics of $\mathbb{P}(Y|\boldsymbol{X})$ and $\mathbb{P}(\boldsymbol{X})$, we obtain:

$$\boldsymbol{T}(\boldsymbol{U}) = \left[Y \; Y\overrightarrow{X_j}^T \; Y^2 \; Y\overrightarrow{X_j X_k}^T \; \overrightarrow{X_j}^T \; \overrightarrow{X_j X_k}^T \; \overrightarrow{X_j^2}^T \; \overrightarrow{X_j X_k X_l}^T \; \overrightarrow{X_j^2 X_k}^T \; \overrightarrow{X_j X_k X_l X_m}^T \; \overrightarrow{X_j^2 X_k X_l}^T \; \overrightarrow{X_j^2 X_k^2}^T \right]^T.$$

(4)

A.2   Maximization of the Q-function

The Q-function for a sample can be written as follows:

$$\begin{aligned} Q_{\boldsymbol{\theta}^{(t)}}(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{\theta}^{(t)}} \left[\log \mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{u}_i) | \boldsymbol{u}_{Oi} \right] \\ &= \sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{\theta}^{(t)}} \left[\log \mathbb{P}_{\boldsymbol{\theta}}(y_i|\boldsymbol{x}_i) + \log \mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{x}_i) | \boldsymbol{u}_{Oi} \right]. \end{aligned}$$

(5)

Maximizing with respect to $\boldsymbol{\beta}$, the relevant terms of the Q-function are:

$$\begin{aligned} Q_{\boldsymbol{\theta}^{(t)}}(\boldsymbol{\beta}) &= -\frac{1}{2\sigma_\epsilon^2} \sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{\theta}^{(t)}} \left[(y_i - \boldsymbol{d}(\boldsymbol{x}_i)^T \boldsymbol{\beta})^2 \big| \boldsymbol{u}_{Oi} \right] + c \\ &= -\frac{1}{2\sigma_\epsilon^2} \sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{\theta}^{(t)}} \left[\boldsymbol{\beta}^T \boldsymbol{d}(\boldsymbol{x}_i) \boldsymbol{d}(\boldsymbol{x}_i)^T \boldsymbol{\beta} - 2 y_i \boldsymbol{d}(\boldsymbol{x}_i)^T \boldsymbol{\beta} \big| \boldsymbol{u}_{Oi} \right] + c'. \end{aligned}$$

(6)

The first partial derivative is then

$$\frac{\partial Q_{\boldsymbol{\theta}^{(t)}}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -\frac{1}{2\sigma_\epsilon^2} \sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{\theta}^{(t)}} \left[2 \boldsymbol{d}(\boldsymbol{x}_i) \boldsymbol{d}(\boldsymbol{x}_i)^T \boldsymbol{\beta} - 2 y_i \boldsymbol{d}(\boldsymbol{x}_i) \big| \boldsymbol{u}_{Oi} \right],$$

(7)

which we set to zero and obtain the first-order condition of

$$\sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{\theta}^{(t)}} \left[2\boldsymbol{d}(\boldsymbol{x}_i)\boldsymbol{d}(\boldsymbol{x}_i)^T \middle| \boldsymbol{u}_{Oi} \right] \boldsymbol{\beta} = \sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{\theta}^{(t)}} \left[2y_i \boldsymbol{d}(\boldsymbol{x}_i) | \boldsymbol{u}_{Oi} \right]. \tag{8}$$

Thus our estimate of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{\theta}^{(t)}} \left[\boldsymbol{d}(\boldsymbol{x}_i)\boldsymbol{d}(\boldsymbol{x}_i)^T \middle| \boldsymbol{u}_{Oi} \right] \right)^{-1} \sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{\theta}^{(t)}} \left[y_i \boldsymbol{d}(\boldsymbol{x}_i) | \boldsymbol{u}_{Oi} \right]. \tag{9}$$

Maximizing with respect to σ_ϵ^2, we return to Equation 5 and once again collect the relevant terms as

$$Q_{\boldsymbol{\theta}^{(t)}}(\sigma_\epsilon^2) = -\frac{n}{2} \log(\sigma_\epsilon^2) - \frac{1}{2\sigma_\epsilon^2} \sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{\theta}^{(t)}} \left[(y_i - \boldsymbol{d}(\boldsymbol{x}_i)^T \boldsymbol{\beta})^2 \middle| \boldsymbol{u}_{Oi} \right] + c. \tag{10}$$

The first partial derivative is then

$$\frac{\partial Q_{\boldsymbol{\theta}^{(t)}}(\sigma_\epsilon^2)}{\partial \sigma_\epsilon^2} = -\frac{n}{2\sigma_\epsilon^2} + \frac{1}{2\sigma_\epsilon^4} \sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{\theta}^{(t)}} \left[(y_i - \boldsymbol{d}(\boldsymbol{x}_i)^T \boldsymbol{\beta})^2 \middle| \boldsymbol{u}_{Oi} \right], \tag{11}$$

which yields the first-order condition and estimate of σ_ϵ^2 as

$$\frac{n}{2\sigma_\epsilon^2} = \frac{1}{2\sigma_\epsilon^4} \sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{\theta}^{(t)}} \left[(y_i - \boldsymbol{d}(\boldsymbol{x}_i)^T \boldsymbol{\beta})^2 \middle| \boldsymbol{u}_{Oi} \right]$$

$$\Rightarrow \hat{\sigma}_\epsilon^2 = \frac{\sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{\theta}^{(t)}} \left[(y_i - \boldsymbol{d}(\boldsymbol{x}_i)^T \hat{\boldsymbol{\beta}})^2 \middle| \boldsymbol{u}_{Oi} \right]}{n}. \tag{12}$$

To express this quantity strictly in terms of the expected sufficient statistics, we expand the quadratic term to obtain

$$\hat{\sigma}_\epsilon^2 = \frac{\sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{\theta}^{(t)}} \left[(y_i - \boldsymbol{d}(\boldsymbol{x}_i)\hat{\boldsymbol{\beta}})^2 \middle| \boldsymbol{u}_{Oi} \right]}{n}$$

$$= \frac{\sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{\theta}^{(t)}} \left[y_i^2 | \boldsymbol{u}_{Oi} \right]}{n} - 2\frac{\sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{\theta}^{(t)}} \left[y_i \boldsymbol{d}(\boldsymbol{x}_i)^T | \boldsymbol{u}_{Oi} \right]}{n} \hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\beta}}^T \frac{\sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{\theta}^{(t)}} \left[\boldsymbol{d}(\boldsymbol{x}_i)\boldsymbol{d}(\boldsymbol{x}_i)^T | \boldsymbol{u}_{Oi} \right]}{n} \hat{\boldsymbol{\beta}}. \tag{13}$$

For simplicity, let us express $\hat{\boldsymbol{\beta}}$ with the following shorthand notation:

$$
\boldsymbol{A} := \sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{\theta}^{(t)}} \left[\boldsymbol{d}(\boldsymbol{x}_i) \boldsymbol{d}(\boldsymbol{x}_i)^T \big| \boldsymbol{u}_{Oi} \right]
$$
$$
\boldsymbol{b} := \sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{\theta}^{(t)}} \left[y_i \boldsymbol{d}(\boldsymbol{x}_i) | \boldsymbol{u}_{Oi} \right]
$$
$$
\Rightarrow \hat{\boldsymbol{\beta}} = \boldsymbol{A}^{-1} \boldsymbol{b}.
$$

(14)

Re-expressing Equation 13 with these quantities, we have:

$$
\begin{aligned}
\hat{\sigma}_\epsilon^2 &= \frac{\sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{\theta}^{(t)}} \left[y_i^2 | \boldsymbol{u}_{Oi} \right]}{n} - 2\frac{\boldsymbol{b}^T}{n}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\beta}}^T \frac{\boldsymbol{A}}{n}\hat{\boldsymbol{\beta}} \\
&= \frac{\sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{\theta}^{(t)}} \left[y_i^2 | \boldsymbol{u}_{Oi} \right]}{n} - 2\frac{\boldsymbol{b}^T \boldsymbol{A}^{-1} \boldsymbol{b}}{n} + \frac{\boldsymbol{b}^T \boldsymbol{A}^{-1} \boldsymbol{A} \boldsymbol{A}^{-1} \boldsymbol{b}}{n} \\
&= \frac{\sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{\theta}^{(t)}} \left[y_i^2 | \boldsymbol{u}_{Oi} \right]}{n} - \frac{\boldsymbol{b} \boldsymbol{A}^{-1} \boldsymbol{b}}{n}
\end{aligned}
$$

(15)

which finally yields

$$
\hat{\sigma}_\epsilon^2 = \frac{\sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{\theta}^{(t)}} \left[y_i^2 | \boldsymbol{u}_{Oi} \right]}{n} - \frac{\sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{\theta}^{(t)}} \left[y_i \boldsymbol{d}(\boldsymbol{x}_i) | \boldsymbol{u}_{Oi} \right] \left(\sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{\theta}^{(t)}} \left[\boldsymbol{d}(\boldsymbol{x}_i) \boldsymbol{d}(\boldsymbol{x}_i)^T | \boldsymbol{u}_{Oi} \right] \right)^{-1}}{n}
$$
$$
\times \sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{\theta}^{(t)}} \left[y_i \boldsymbol{d}(\boldsymbol{x}_i) | \boldsymbol{u}_{Oi} \right]
$$

(16)

Returning to Equation 5 and collecting the relevant terms for $\boldsymbol{\mu}$ we have

$$
\begin{aligned}
Q_{\boldsymbol{\theta}^{(t)}}(\boldsymbol{\mu}) &= -\frac{1}{2} \sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{\theta}^{(t)}} \left[(\boldsymbol{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}) \big| \boldsymbol{u}_{Oi} \right] + c \\
&= -\frac{n}{2} \boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} + \sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{\theta}^{(t)}} \left[\boldsymbol{x}_i | \boldsymbol{u}_{Oi} \right]^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} + c'.
\end{aligned}
$$

(17)

This yields a first partial derivative of

$$
\frac{\partial Q_{\boldsymbol{\theta}^{(t)}}(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}} = -n \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} + \boldsymbol{\Sigma}^{-1} \sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{\theta}^{(t)}} \left[\boldsymbol{x}_i | \boldsymbol{u}_{Oi} \right],
$$

(18)

and thus our first-order condition and estimator for $\boldsymbol{\mu}$ is

$$n\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} = \boldsymbol{\Sigma}^{-1}\sum_{i=1}^{n}\mathbb{E}_{\boldsymbol{\theta}^{(t)}}\left[\boldsymbol{x}_i|\boldsymbol{u}_{Oi}\right]$$
$$\Rightarrow \hat{\boldsymbol{\mu}} = \frac{\sum_{i=1}^{n}\mathbb{E}_{\boldsymbol{\theta}^{(t)}}\left[\boldsymbol{x}_i|\boldsymbol{u}_{Oi}\right]}{n} \tag{19}$$

Finally, for $\boldsymbol{\Sigma}$, we once again collect the relevant terms from the Q-function as follows:

$$Q_{\boldsymbol{\theta}^{(t)}}(\boldsymbol{\Sigma}) = -\frac{n}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{i=1}^{n}\mathbb{E}_{\boldsymbol{\theta}^{(t)}}\left[(\boldsymbol{x}_i - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})\big|\boldsymbol{u}_{Oi}\right] + c, \tag{20}$$

which yields a first partial derivative of

$$\frac{\partial Q_{\boldsymbol{\theta}^{(t)}}(\boldsymbol{\Sigma})}{\partial\boldsymbol{\Sigma}} = -\frac{n}{2}\boldsymbol{\Sigma}^{-1} + \frac{1}{2}\boldsymbol{\Sigma}^{-1}\left(\sum_{i=1}^{n}\mathbb{E}_{\boldsymbol{\theta}^{(t)}}\left[(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^T\big|\boldsymbol{u}_{Oi}\right]\right)\boldsymbol{\Sigma}^{-1}, \tag{21}$$

where we used the identities of $\dfrac{\partial\log|\boldsymbol{A}|}{\partial\boldsymbol{A}} = \boldsymbol{A}^{-1}$ and $\dfrac{\partial\boldsymbol{b}^T\boldsymbol{A}^{-1}\boldsymbol{c}}{\partial\boldsymbol{A}} = -\left(\boldsymbol{A}^{-1}\boldsymbol{c}\boldsymbol{b}^T\boldsymbol{A}^{-1}\right)^T$ for an invertible matrix $\boldsymbol{A}$, and vectors $\boldsymbol{b}$ and $\boldsymbol{c}$. This gives us the first-order condition and estimator of $\boldsymbol{\Sigma}$ as

$$\frac{n}{2}\boldsymbol{\Sigma}^{-1} = \frac{1}{2}\boldsymbol{\Sigma}^{-1}\left(\sum_{i=1}^{n}\mathbb{E}_{\boldsymbol{\theta}^{(t)}}\left[(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^T\big|\boldsymbol{u}_{Oi}\right]\right)\boldsymbol{\Sigma}^{-1}$$
$$\Rightarrow n\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma} = \boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}\left(\sum_{i=1}^{n}\mathbb{E}_{\boldsymbol{\theta}^{(t)}}\left[(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^T\big|\boldsymbol{u}_{Oi}\right]\right)\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma} \tag{22}$$
$$\Rightarrow \hat{\boldsymbol{\Sigma}} = \frac{\sum_{i=1}^{n}\mathbb{E}_{\boldsymbol{\theta}^{(t)}}\left[(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}})(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}})^T\big|\boldsymbol{u}_{Oi}\right]}{n}.$$

As with $\hat{\sigma}_\epsilon^2$ this can be written in terms of the sufficient statistics with

$$\hat{\boldsymbol{\Sigma}} = \frac{\sum_{i=1}^{n}\mathbb{E}_{\boldsymbol{\theta}^{(t)}}\left[\boldsymbol{x}_i\boldsymbol{x}_i^T\big|\boldsymbol{u}_{Oi}\right]}{n} - \frac{\hat{\boldsymbol{\mu}}\sum_{i=1}^{n}\mathbb{E}_{\boldsymbol{\theta}^{(t)}}\left[\boldsymbol{x}_i|\boldsymbol{u}_{Oi}\right]^T}{n} - \frac{\sum_{i=1}^{n}\mathbb{E}_{\boldsymbol{\theta}^{(t)}}\left[\boldsymbol{x}_i|\boldsymbol{u}_{Oi}\right]\hat{\boldsymbol{\mu}}^T}{n} + \hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^T$$
$$= \frac{\sum_{i=1}^{n}\mathbb{E}_{\boldsymbol{\theta}^{(t)}}\left[\boldsymbol{x}_i\boldsymbol{x}_i^T\big|\boldsymbol{u}_{Oi}\right]}{n} - \hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^T. \tag{23}$$

A.3 Gaussian MGF Examples

For didactic purposes, the first, second and second-cross moments of the Gaussian distribution are derived here using the moment generating function. Let $\boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then for convenience, define the exponentiated quadratic form as $\eta := \boldsymbol{t}^T \boldsymbol{\mu} + \frac{1}{2} \boldsymbol{t}^T \boldsymbol{\Sigma} \boldsymbol{t}$, and write the MGF as $M_{\boldsymbol{X}}(\boldsymbol{t}) = \exp(\eta)$. As a preliminary observeration, using basic vector calculus, notice that the gradient of η with respect to $\boldsymbol{t}$ is

$$\nabla_{\boldsymbol{t}} \eta = \boldsymbol{\mu} + \boldsymbol{\Sigma} \boldsymbol{t}. \tag{24}$$

Therefore the partial derivative of η with respect to the jth element of $\boldsymbol{t}$ is

$$\frac{\partial \eta}{\partial t_j} = \mu_j + \boldsymbol{\Sigma}_{\cdot j}^T \boldsymbol{t}, \tag{25}$$

where $\boldsymbol{\Sigma}_{\cdot j}$ is the jth column of $\boldsymbol{\Sigma}$. With this in mind, the first, second, and second-cross partial derivatives of $M_{\boldsymbol{X}}(\boldsymbol{t})$ are as follows.

$$\begin{aligned} \frac{\partial M_{\boldsymbol{X}}(\boldsymbol{t})}{\partial t_j} &= \left[\frac{\partial \exp(\eta)}{\partial \eta} \right] \left[\frac{\partial \eta}{\partial t_j} \right] \\ &= \exp(\eta)(\mu_j + \boldsymbol{\Sigma}_{\cdot j}^T \boldsymbol{t}), \end{aligned} \tag{26}$$

which follows by the chain rule. We can leverage this result into the second partial derivative since

$$\begin{aligned} \frac{\partial^2 M_{\boldsymbol{X}}(\boldsymbol{t})}{\partial t_j^2} &= \left[\frac{\partial}{\partial t_j} \exp(\eta) \right] (\mu_j + \boldsymbol{\Sigma}_{\cdot j}^T \boldsymbol{t}) + \exp(\eta) \left[\frac{\partial}{\partial t_j} (\mu_j + \boldsymbol{\Sigma}_{\cdot j}^T \boldsymbol{t}) \right] \\ &= \exp(\eta)(\mu_j + \boldsymbol{\Sigma}_{\cdot j}^T \boldsymbol{t})^2 + \exp(\eta) \boldsymbol{\Sigma}_{jj}, \end{aligned} \tag{27}$$

which follows from the product rule and $\boldsymbol{\Sigma}_{jj}$ is the element of $\boldsymbol{\Sigma}$ in the jth row and jth column, or σ_j^2. Finally, the second-cross partial derivative is

$$\begin{aligned} \frac{\partial^2 M_{\boldsymbol{X}}(\boldsymbol{t})}{\partial t_j \partial t_k} &= \left[\frac{\partial}{\partial t_k} \exp(\eta) \right] (\mu_j + \boldsymbol{\Sigma}_{\cdot j}^T \boldsymbol{t}) + \exp(\eta) \left[\frac{\partial}{\partial t_k} (\mu_j + \boldsymbol{\Sigma}_{\cdot j}^T \boldsymbol{t}) \right] \\ &= \exp(\eta)(\mu_k + \boldsymbol{\Sigma}_{\cdot k}^T \boldsymbol{t})(\mu_j + \boldsymbol{\Sigma}_{\cdot j}^T \boldsymbol{t}) + \exp(\eta) \boldsymbol{\Sigma}_{jk}. \end{aligned} \tag{28}$$

From here, the Gaussian moments can be obtained by evaluating all these partial derivatives at $\boldsymbol{t} = \boldsymbol{0}$.

$$
\begin{aligned}
\mathbb{E}[X_j] &= \left. \frac{\partial M_{\boldsymbol{X}}(\boldsymbol{t})}{\partial t_j} \right|_{\boldsymbol{t}=\boldsymbol{0}} \\
&= \exp(0)(\mu_j + \boldsymbol{\Sigma}_{\cdot j}^T \boldsymbol{0}) \\
&= \mu_j \\
\mathbb{E}[X_j^2] &= \left. \frac{\partial^2 M_{\boldsymbol{X}}(\boldsymbol{t})}{\partial t_j^2} \right|_{\boldsymbol{t}=\boldsymbol{0}} \\
&= \exp(0)(\mu_j + \boldsymbol{\Sigma}_{\cdot j}^T \boldsymbol{0})^2 + \exp(0)\boldsymbol{\Sigma}_{jj} \\
&= \mu_j^2 + \sigma_j^2 \\
\mathbb{E}[X_j X_k] &= \left. \frac{\partial^2 M_{\boldsymbol{X}}(\boldsymbol{t})}{\partial t_j \partial t_k} \right|_{\boldsymbol{t}=\boldsymbol{0}} \\
&= \exp(0)(\mu_k + \boldsymbol{\Sigma}_{\cdot k}^T \boldsymbol{0})(\mu_j + \boldsymbol{\Sigma}_{\cdot j}^T \boldsymbol{0}) + \exp(0)\boldsymbol{\Sigma}_{jk} \\
&= \mu_j \mu_k + \sigma_{jk}.
\end{aligned}
\tag{29}
$$

A.4 Re-expression of $\mathbb{E}[Y X_j^a X_k^b | X_O]$ Under MDP 1

Denote the index set of missing data except j and k as $A = M \setminus \{j, k\}$. Then we can re-express $\mathbb{E}[Y X_j^a X_k^b | \boldsymbol{X}_O]$ as follows:

$$
\begin{aligned}
\mathbb{E}[Y X_j^a X_k^b | \boldsymbol{X}_O] &= \int_y \int_{x_j} \int_{x_k} y x_j^a x_k^b \mathbb{P}(y, x_j, x_k | \boldsymbol{x}_O) \, dx_j \, dx_k \, dy \\
&= \int_y \int_{x_j} \int_{x_k} y x_j^a x_k^b \int_{\boldsymbol{x}_A} \mathbb{P}(y, x_j, x_k, \boldsymbol{x}_A | \boldsymbol{x}_O) \, d\boldsymbol{x}_A \, dx_j \, dx_k \, dy \\
&= \int_y \int_{x_j} \int_{x_k} \int_{\boldsymbol{x}_A} y x_j^a x_k^b \mathbb{P}(y | x_j, x_k, \boldsymbol{x}_A, \boldsymbol{x}_O) \mathbb{P}(x_j, x_k, \boldsymbol{x}_A | \boldsymbol{x}_O) \, d\boldsymbol{x}_A \, dx_j \, dx_k \, dy \\
&= \int_{x_j} \int_{x_k} \int_{\boldsymbol{x}_A} \mathbb{E}[Y | \boldsymbol{X}] x_j^a x_k^b \mathbb{P}(x_j, x_k, \boldsymbol{x}_A | \boldsymbol{x}_O) \, d\boldsymbol{x}_A \, dx_j \, dx_k \\
&= \int_{\boldsymbol{x}_M} \boldsymbol{d}(\boldsymbol{X})^T \boldsymbol{\beta} x_j^a x_k^b \mathbb{P}(\boldsymbol{x}_M | \boldsymbol{x}_O) \, d\boldsymbol{x}_M \\
&= \mathbb{E}[\boldsymbol{d}(\boldsymbol{X})^T \boldsymbol{\beta} X_j^a X_k^b | \boldsymbol{X}_O],
\end{aligned}
\tag{30}
$$

where $\mathbb{P}(x_j, x_k, \boldsymbol{x}_A | \boldsymbol{x}_O) = \mathbb{P}(\boldsymbol{x}_M | \boldsymbol{x}_O)$ since $M = A \cup \{j, k\}$.

A.5   Derivation of $\mathbb{P}(X_M|Y, X_O)$

Consider a quadratic form of Equation 1.1:

$$Y = \beta_0 + \boldsymbol{\beta}_f^T \boldsymbol{X} + \frac{1}{2} \boldsymbol{X}^T \boldsymbol{\beta}_s \boldsymbol{X} + \epsilon, \tag{31}$$

where, without the loss of generality, we assume $\boldsymbol{X}$ is ordered as $X = \begin{bmatrix} \boldsymbol{X}_O^T & \boldsymbol{X}_M^T \end{bmatrix}^T$ and $\boldsymbol{\beta}_f$ and $\boldsymbol{\beta}_s$ are defined as

$$\boldsymbol{\beta}_f := \begin{bmatrix} \boldsymbol{\beta}_O^T & \boldsymbol{\beta}_M^T \end{bmatrix}^T$$

$$\boldsymbol{\beta}_s := \begin{bmatrix} \boldsymbol{\beta}_{OO} & \boldsymbol{\beta}_{OM} \\ \boldsymbol{\beta}_{OM} & \boldsymbol{\beta}_{MM} \end{bmatrix}, \tag{32}$$

which contain the regression coefficients that correspond to the first-order and product terms of $\boldsymbol{d}(\boldsymbol{X})$, respectively. Note that the diagonal entries of $\boldsymbol{\beta}_s$ are zeroes and under MDP 2 $\boldsymbol{\beta}_{MM}$ is a zero matrix. We can then re-write the error term as a function of $\boldsymbol{X}_M$ as follows

$$
\begin{aligned}
Y &= \beta_0 + \boldsymbol{\beta}_f^T \boldsymbol{X} + \frac{1}{2} \boldsymbol{X}^T \boldsymbol{\beta}_s \boldsymbol{X} + \epsilon \\
&= \beta_0 + \boldsymbol{\beta}_O^T \boldsymbol{X}_O + \boldsymbol{\beta}_M^T \boldsymbol{X}_M + \frac{1}{2} \boldsymbol{X}_O^T \boldsymbol{\beta}_{OO} \boldsymbol{X}_O + \boldsymbol{X}_O^T \boldsymbol{\beta}_{OM} \boldsymbol{X}_M + \epsilon \\
&= \beta_0 + \boldsymbol{\beta}_O^T \boldsymbol{X}_O + \frac{1}{2} \boldsymbol{X}_O^T \boldsymbol{\beta}_{OO} \boldsymbol{X}_O + (\boldsymbol{\beta}_M^T + \boldsymbol{X}_O^T \boldsymbol{\beta}_{OM}) \boldsymbol{X}_M + \epsilon \\
\Rightarrow \epsilon &= Y - \beta_0 - \boldsymbol{\beta}_O^T \boldsymbol{X}_O - \frac{1}{2} \boldsymbol{X}_O^T \boldsymbol{\beta}_{OO} \boldsymbol{X}_O - (\boldsymbol{\beta}_M^T + \boldsymbol{X}_O^T \boldsymbol{\beta}_{OM}) \boldsymbol{X}_M \\
\epsilon &= \alpha_0 - \boldsymbol{\alpha}_1^T \boldsymbol{X}_M,
\end{aligned}
\tag{33}
$$

which follows by defining

$$
\begin{aligned}
\alpha_0 &:= Y - \beta_0 - \boldsymbol{\beta}_O^T \boldsymbol{X}_O - \frac{1}{2} \boldsymbol{X}_O^T \boldsymbol{\beta}_{OO} \boldsymbol{X}_O \\
\boldsymbol{\alpha}_1 &:= \boldsymbol{\beta}_M + \boldsymbol{\beta}_{OM}^T \boldsymbol{X}_O.
\end{aligned}
\tag{34}
$$

From here, we can proceed with a direct derivation of $\mathbb{P}(\boldsymbol{X}_M | Y, \boldsymbol{X}_O)$ as follows:

$$
\begin{aligned}
\mathbb{P}(\boldsymbol{X}_M | Y, \boldsymbol{X}_O) & \propto \mathbb{P}(Y, \boldsymbol{X}_M, \boldsymbol{X}_O) \\
& \propto \mathbb{P}(Y | \boldsymbol{X}_M, \boldsymbol{X}_O) \mathbb{P}(\boldsymbol{X}_M | \boldsymbol{X}_O) \\
& \propto \exp\left[-\frac{1}{2} \sigma_\epsilon^{-2} (\alpha_0 - \boldsymbol{\alpha}_1^T \boldsymbol{X}_M)^2 \right] \exp\left[-\frac{1}{2} (\boldsymbol{X}_M - \boldsymbol{\mu}_c)^T \boldsymbol{\Sigma}_c^{-1} (\boldsymbol{X}_M - \boldsymbol{\mu}_c) \right] \\
& \propto \exp\left[-\frac{1}{2} \sigma_\epsilon^{-2} (\boldsymbol{X}_M^T \boldsymbol{\alpha}_1 \boldsymbol{\alpha}_1^T \boldsymbol{X}_M - 2\alpha_0 \boldsymbol{\alpha}_1^T \boldsymbol{X}_M) \right] \\
& \qquad\qquad\qquad\qquad \times \exp\left[-\frac{1}{2} (\boldsymbol{X}_M^T \boldsymbol{\Sigma}_c^{-1} \boldsymbol{X}_M - 2\boldsymbol{\mu}_c^T \boldsymbol{\Sigma}_c^{-1} \boldsymbol{X}_M) \right] \\
& = \exp\left[-\frac{1}{2} (\boldsymbol{X}_M^T (\boldsymbol{\Sigma}_c^{-1} + \sigma_\epsilon^{-2} \boldsymbol{\alpha}_1 \boldsymbol{\alpha}_1^T) \boldsymbol{X}_M + 2(\boldsymbol{\mu}_c^T \boldsymbol{\Sigma}_c^{-1} + \sigma_\epsilon^{-2} \alpha_0 \boldsymbol{\alpha}_1^T) \boldsymbol{X}_M) \right] \\
& = \exp\left[-\frac{1}{2} (\boldsymbol{X}_M - \boldsymbol{b})^T \boldsymbol{A} (\boldsymbol{X}_M - \boldsymbol{b}) - \frac{1}{2} \boldsymbol{b}^T \boldsymbol{A} \boldsymbol{b} \right],
\end{aligned}
\tag{35}
$$

where we completed the square by using the shorthand notation:

$$
\begin{aligned}
\boldsymbol{A} & := \boldsymbol{\Sigma}_c^{-1} + \sigma_\epsilon^{-2} \boldsymbol{\alpha}_1 \boldsymbol{\alpha}_1^T \\
\boldsymbol{b} & := \boldsymbol{A}^{-1} (\boldsymbol{\Sigma}_c^{-1} \boldsymbol{\mu}_c + \sigma_\epsilon^{-2} \alpha_0 \boldsymbol{\alpha}_1).
\end{aligned}
\tag{36}
$$

This ultimately implies

$$
\begin{aligned}
& \mathbb{P}(\boldsymbol{X}_M | Y, \boldsymbol{X}_O) \propto \exp\left[-\frac{1}{2} (\boldsymbol{X}_M - \boldsymbol{b})^T \boldsymbol{A} (\boldsymbol{X}_M - \boldsymbol{b}) \right] \\
& \Rightarrow \mathbb{P}(\boldsymbol{X}_M | Y, \boldsymbol{X}_O) \sim \mathcal{N}(\boldsymbol{b}, \boldsymbol{A}^{-1}),
\end{aligned}
\tag{37}
$$

where $\boldsymbol{A}$ and $\boldsymbol{b}$ are functions of $\boldsymbol{\theta}$ as defined in Equation 36. We note that this is the multivariate generalization to the one-dimensional result found by Kim et al. (2015).

A.6 Generating σ_ϵ^2

Given a sampled $\boldsymbol{x}$, $\boldsymbol{\beta}$, and R_a^2, we algebraically solve for σ_ϵ^2 in the following manner. First, note that σ_ϵ^2 can be written as a function of R^2 by the identity:

$$R^2 = \frac{\sigma_{\hat{Y}}^2}{\sigma_{\hat{Y}}^2 + \sigma_\epsilon^2}$$

$$\sigma_{\hat{Y}}^2 R^2 + \sigma_\epsilon^2 R^2 = \sigma_{\hat{Y}}^2$$

$$\sigma_\epsilon^2 R^2 = \sigma_{\hat{Y}}^2 - \sigma_{\hat{Y}}^2 R^2 \tag{38}$$

$$\sigma_\epsilon^2 R^2 = \sigma_{\hat{Y}}^2 (1 - R^2)$$

$$\sigma_\epsilon^2 = \frac{\sigma_{\hat{Y}}^2 (1 - R^2)}{R^2}.$$

In turn, $\hat{\sigma}_{\hat{Y}}^2$ can be estimated from $\boldsymbol{x}$ and $\boldsymbol{\beta}$ by the following calculations:

$$\hat{Y}_i = \boldsymbol{\beta}^T \boldsymbol{d}(\boldsymbol{x}_i)$$

$$\boldsymbol{\mu}_{\hat{Y}} = \sum_{i=1}^n \frac{\hat{Y}_i}{n} \tag{39}$$

$$\hat{\sigma}_{\hat{Y}}^2 = \sum_{i=1}^n \frac{(\hat{Y}_i - \boldsymbol{\mu}_{\hat{Y}})^2}{n - 1},$$

Finally, the conversion between R^2 and R_a^2 was obtained by simply algebraically manipulating the definition R_a^2

$$R_a^2 = 1 - (1 - R^2)\frac{n - 1}{n - d}$$

$$\Rightarrow R^2 = 1 - (1 - R_a^2)\frac{n - d}{n - 1}. \tag{40}$$

Thus, using $\boldsymbol{x}$, $\boldsymbol{\beta}$, and R_a^2, we can generate a σ_ϵ^2 parameter and simulate the errors as $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$ to finally obtain outcome with $Y = \boldsymbol{d}(\boldsymbol{X})^T \boldsymbol{\beta} + \epsilon$.

A.7 Generating $\boldsymbol{R}$

The first step in generating $\boldsymbol{R}$ was assigning missing data patterns. Recall that the simulation parameter φ_{MIS} was the proportion of missing data overall and φ_{MDP3} was the proportion *among the missing data* assigned to MDP 3. Let us define φ_j as the proportion

of the overall data assigned to MDP j. We set these quantities as follows:

$$\varphi_1 = \varphi_2 := \frac{1}{2}\varphi_{\text{MIS}}(1 - \varphi_{\text{MDP3}})$$
$$\varphi_3 := \varphi_{\text{MIS}}\varphi_{\text{MDP3}}$$

(41)

That is, φ_3 is simply determined by φ_{MIS} and φ_{MDP3}, and φ_1 and φ_2 share the remaining proportion of missing data is equally. Subsequently, let us define n_j as the number of samples assigned to MDP j. This was calculated as $n_j = \lceil n\varphi_j \rceil$, or $n\varphi_j$ rounded up. Therefore, even if n is small, each MDP is assigned at least one case so long as its corresponding $\varphi_j > 0$.

Once all n_j were determined, we randomly (with uniform probability) selected a non-product variable in $\boldsymbol{X}$ to serve as a "missingness anchor" variable (denoted X_a, with mean μ_a and variance σ_a^2) to determine the missingness in all other variables. This was done through an intermediate latent propensity variable, denoted R^*, defined as

$$R^* := \zeta \left(\frac{X_a - \mu_a}{\sigma_a} \right) + \left(\sqrt{1 - \zeta^2} \right) \epsilon_0,$$

(42)

where $\epsilon_0 \sim \mathcal{N}(0,1)$. This parameterization sets $\text{Cor}(R^*, X_a) = \zeta$, which allows for simple control over the MAR mechanism. We set $\zeta = 0.7$ for all conditions. Thus given n samples of X_a, we used Equation 42 to draw n samples of R^*. Then to determine missingness, the greatest $n_{MIS} := n_1 + n_2 + n_3$ values of the sample of R^* were assigned randomly to MDPs 1, 2, or 3 with uniform probability, such that the total of each MDP was equal to its corresponding n_j.

Once MDPs were assigned, missingness was drawn as follows. First, to guarantee that the assigned missingness pattern was followed, one X_j (or a product-term pair of X_j and X_k) had their R_{X_j} set to zero, selected with uniform probability among all MDP appropriate $\boldsymbol{X}$ variables. Then we set

- MDP 1: $R_Y = 0$ and $R_{X_j} \sim \text{Bernoulli}(0.5)$, for all $j \neq a$.

- MDP 2: $R_Y = 1$ and $R_{X_j} \sim \text{Bernoulli}(0.5)$, for all $j \neq a$ such that X_j is not a factor of any product-terms in $\boldsymbol{d}(\boldsymbol{X})$. Then for pairs all (X_j, X_k) that form a product-terms

42

in $\boldsymbol{d}(\boldsymbol{X})$, we draw R_{X_j} sequentially:

- If both R_{X_j} and R_{X_k} have not been drawn yet, then select one with uniform probability and draw as Bernoulli(0.5).

- If only one R_{X_j} has been set, then $R_{X_k} = 1 - R_{X_j}$.

- MDP 3: $R_Y = 1$ and $R_j = R_k \sim$ Bernoulli(0.5), for all pairs $(j \neq a, k \neq a)$ such that (X_j, X_k) form a product-term in $\boldsymbol{d}(\boldsymbol{X})$.

A.8   Numerical Software Comparison

To verify the numerical compatibility of our estimation methods with existing software, we compared the NI parameter estimates with that of the `mdmb` R package (Robitzsch & Lüdtke, 2021). We ran an abbreviated version of the simulation study described in Section 3.1. The conditions were:

- Sample size (n): 500 and 1000.

- Number of predictors (p): 3 and 5.

- Proportion of missingness (φ_{MIS}): 0.30.

- Proportion among missing data from MDP 3 $(\varphi_{\mathrm{MDP3}})$: 1/3.

The number of simulations was set to 100 per condition, using the data generating procedure described in Section 3.2. We calculated the absolute differences and logged absolute differences between individual parameter estimates and summarized them in Table A.8.1. A histogram of the logged absolute differences are displayed in Figure A.8.1. The mean absolute difference between the two methods is 0.017 ± 0.001 indicating the two methods are similar numerically. Further, the NI method has marginally better parameter recovery metrics, with an empirical bias of 0.0005, and empirical MSE of 0.0244, compared to an empirical bias and MSE of -0.0007 and 0.0303, respectively, from MDMB.
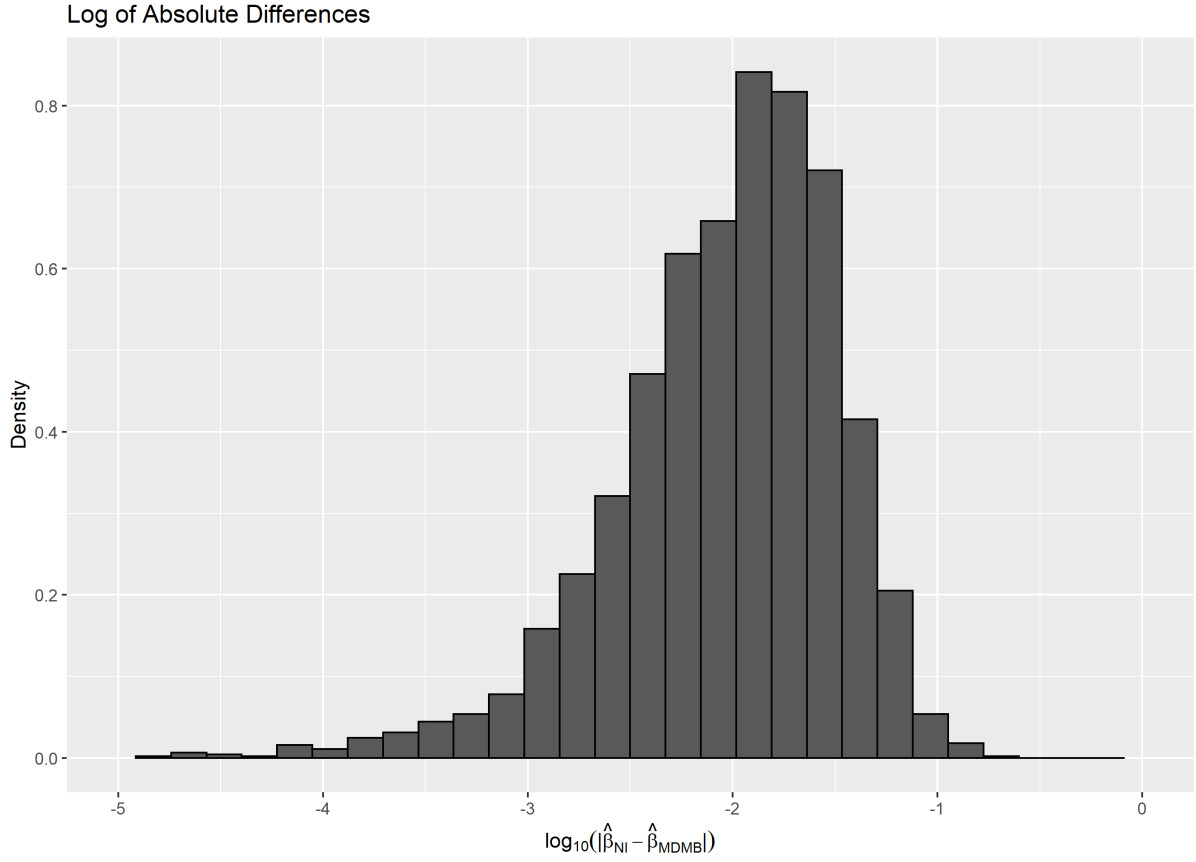
Figure A.8.1: Histogram of $\log_{10}$ absolute differences.

Summary	$\lvert\hat{\beta}_{NI} - \hat{\beta}_{MDMB}\rvert$	$\log_{10}\lvert\hat{\beta}_{NI} - \hat{\beta}_{MDMB}\rvert$
Max.	2.10×10^{-1}	-0.68
Q_3	2.25×10^{-2}	-1.65
Mean	1.66×10^{-2}	-2.03
Median	1.15×10^{-2}	-1.94
Q_1	4.69×10^{-3}	-2.33
Min.	5.19×10^{-6}	-5.28

Table A.8.1: Summary mean and quartiles for the absolute and logged absolute differences between NI and MDMB parameter estimates.

References

Achenbach, T. M., & Rescorla, L. A. (2001). *Manual for the ASEBA school-age forms & profiles: An integrated system of multi-informant assessment.* Burlington, VT: University of Vermont.

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions.* Thousand Oaks, CA: SAGE Publications.

Angold, A., Costello, E. J., & Erkanli, A. (1999). Comorbidity. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, *40*(1), 57–87.

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychology research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*(6), 1173–1182.

Bartlett, J. W., Seaman, S. R., White, I. R., & Carpenter, J. R. (2015). Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. *Statistical Methods in Medical Research*, *24*(4), 462–487.

Booth, J. G., & Hobert, J. P. (1999). Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society: Series B*, *61*(1), 265–285.

Brooks, S. P., & Roberts, G. O. (1998). Convergence assessment techniques for Markov chain Monte Carlo. *Statistics and Computing*, *8*(4), 319–335.

Clark, D. B., Fisher, C. B., Bookheimer, S., Brown, S. A., Evans, J. H., Hopfer, C., . . . Yurgelun-Todd, D. (2018). Biomedical ethics and clinical oversight in multisite observational neuroimaging studies with children and adolescents: The ABCD experience. *Developmental Cognitive Neuroscience*, *32*, 143–154.

Cowles, M. K., & Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics:

A comparative review. *Journal of the American Statistical Association*, *91*(434), 883–904.

Cowles, M. K., Roberts, G. O., & Rosenthal, J. S. (1999). Possible biases induced by MCMC convergence diagnostics. *Journal of Statistical Computation and Simulation*, *64*(1), 87–104.

Dawson, J. F. (2014). Moderation in management research: What, why, When, and how. *Journal of Business and Psychology*, *29*(1), 1–19.

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

Eddelbuettel, D., & François, R. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, *40*(8), 1–18. Retrieved from `https://www.jstatsoft.org/v40/i08/` doi: 10.18637/jss.v040.i08

Eddelbuettel, D., & Sanderson, C. (2014, March). Rcpparmadillo: Accelerating r with high-performance c++ linear algebra. *Computational Statistics and Data Analysis*, *71*, 1054–1063. Retrieved from `http://dx.doi.org/10.1016/j.csda.2013.02.005`

Efron, B., & Tibshirani, R. J. (1994). *An Introduction to the Bootstrap*. Boca Raton, FL: Chapman & Hall.

Enders, C. K., Baraldi, A. N., & Cham, H. (2014). Estimating interaction effects with incomplete predictor variables. *Psychological Methods*, *19*(1), 39–55.

Enders, C. K., Du, H., & Keller, B. T. (2020). A model-based imputation procedure for multilevel regression models with random coefficients, interaction effects, and nonlinear terms. *Psychological Methods*, *25*(1), 88–112.

Flegal, J. M., Haran, M., & Jones, G. L. (2008). Markov chain Monte Carlo: Can we trust the third significant figure? *Statistical Science*, *23*(2), 250–260.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472.

Gelman, A., & Rubin, D. B. (1996). Markov chain Monte Carlo methods in biostatistics. *Statistical Methods in Medical Research*, *5*(4), 339–355.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, *60*(1), 549–576.

Graham, J. W., Taylor, B. J., Olchowski, A. E., & Cumsille, P. E. (2006). Planned missing data designs in psychological research. *Psychological Methods*, *11*(4), 323–343.

Hartley, H. O. (1958). Maximum likelihood estimation from incomplete data. *Biometrics*, *14*(2), 174–194.

Hayes, A. F. (2018). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach* (2nd ed.). New York, NY: The Guilford Press.

Hinrichs, A., Novak, E., Ullrich, M., & Woźniakowski, H. (2014). The curse of dimensionality for numerical integration of smooth functions. *Mathematics of Computation*, *83*, 2853–2863.

Ibrahim, J. G. (1990). Incomplete data in generalized linear models. *Journal of the American Statistical Association*, *85*(411), 765–769.

Jamshidian, M., & Bentler, P. M. (1999). ML estimation of mean and covariance structures with missing data using complete data routines. *Journal of Educational and Behavioral Statistics*, *24*(1), 21–41.

Jamshidian, M., & Jennrich, R. I. (2000). Standard errors for EM estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *62*(2), 257–270.

Jensen, P. S., Martin, D., & Cantwell, D. P. (1997). Comorbidity in ADHD: Implications for research, practice, and DSM-V. *Journal of the American Academy of Child & Adolescent Psychiatry*, *36*(8), 1065–1079.

Keener, R. W. (2010). *Theoretical statistics: Topics for a core course*. New York, NY: Springer.

Kim, S., Sugar, C. A., & Belin, T. R. (2015). Evaluating model-based imputation methods for missing covariates in regression models with interactions. *Statistics in Medicine*, *34*(11), 1876–1888.

Levine, R. A., & Casella, G. (2001). Implementations of the Monte Carlo EM algorithm. *Journal of Computational and Graphical Statistics*, *10*(3), 422–439.

Liu, J., Gelman, A., Hill, J., Su, Y.-S., & Kropko, J. (2014). On the stationary distribution

of iterative imputations. *Biometrika*, *101*(1), 155–173.

Liu, Q., & Pierce, D. A. (1994). A note on Gauss-Hermite quadrature. *Biometrika*, *81*(3), 624–629.

Lüdtke, O., Robitzsch, A., & West, S. G. (2019). Analysis of interactions and nonlinear effects with missing data: A factored regression modeling approach using maximum likelihood estimation. *Multivariate Behavioral Research*.

Lüdtke, O., Robitzsch, A., & West, S. G. (2020). Regression models invovling nonlinear effects with missing data: A sequential modeling approach using bayesian estimation. *Psychological Methods*, *25*(2), 157–181.

McCabe, C. J., Kim, D. S., & King, K. M. (2018). Improving present practices in the visual display of interactions. *Advances in Methods and Practices in Psychological Science*, *1*(2), 147–165.

Meng, X.-L., & Rubin, D. B. (1991). Using EM to obtain asymptotic variance-covariance matrices: The SEM Algorithm. *Journal of the American Statistical Association*, *86*(416), 899–909.

Mossel, E., & Vigoda, E. (2006). Limitations of Markov chain Monte Carlo algorithms for Bayesian inference of phylogeny. *Annals of Applied Probability*, *16*(4), 2215–2234.

Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, *44*(4), 443–460.

Orchard, T., & Woodbury, M. A. (1972). A missing information principle: Theory and Applications. In *Proceedings of the sixth berkeley symposium on mathematical statistics and probability* (pp. 697–715).

Preacher, K. J., Curran, P. J., & Bauer, D. J. (2006). Computational tools for probing interactions in multiple linear regression, multilevel modeling, and latent curve analysis. *Journal of Educational and Behavioral Statistics*, *31*(3), 437–448.

R Core Team. (2021). R: A language and environment for statistical computing. Retrieved from `https://www.R-project.org/`

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2002). Reliable estimation of generalized linear mixed models using adaptive quadrature. *The Stata Journal*, *2*(1), 1–21.

Raghunathan, T. E. (2004). What do we do with missing data? Some options for analysis of incomplete data. *Annual Review of Public Health*, *25*, 99–117.

Robitzsch, A., & Lüdtke, O. (2021). mdmb: Model based treatment of missing data. Retrieved from `https://CRAN.R-project.org/package=mdmb` (R package version 1.5-8)

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592.

Sanderson, C., & Curtin, R. (2016). Armadillo: a template-based C++ library for linear algebra. *The Journal of Open Source Software*, *1*(2), 26.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. Boca Raton, FL: Chapman & Hall.

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*(46), 1–13.

Simonovits, M. (2003). How to compute the volume in high dimension? *Mathematical Programming*, *97*, 337–374.

Tucker, H. G. (1959). A generalization of the Glivenko-Cantelli theorem. *The Annals of Mathematical Statistics*, *30*(3), 828–830.

Volkow, N. D., Koob, G. F., Croyle, R. T., Bianchi, D. W., Gordon, J. A., Koroshetz, W. J., ... Weiss, S. R. B. (2018). The conception of the ABCD study : From substance use to a broad NIH collaboration. *Developmental Cognitive Neuroscience*, *32*, 4–7.

von Hippel, P. T. (2009). How to impute interactions, squares, and other transformed variables. *Sociological Methodology*, *39*(1), 265–291.

Wei, G. C., & Tanner, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, *85*(411), 699–704.

Wu, C. F. J. (1983). On the convergence properties of the EM algorithm. *The Annals of Statistics*, *11*(1), 95–103.

Zhang, Q., & Wang, L. (2017). Moderation analysis with missing data in the predictors.

Psychological Methods, *22*(4), 649–666.