Title

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

Author

Kim, Dale S

Publication Date

2025

DOI

10.3102/10769986241304015

Article

Journal of Educational and Behavioral Statistics
Vol. 00, No. 0, pp. 1–22
DOI: 10.3102/10769986241304015

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

Dale S. Kim
University of California, Los Angeles

We study an Expectation-Maximization (EM) algorithm for estimating product-term regression models with missing data. The study of such problems in the frequentist 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 three empirical experiments using both simulated and real data. We demonstrate that our algorithm provides greater estimation accuracy, exhibits robustness to distributional violations, and confers higher power to detect interaction effects. We conclude with a discussion of extensions and topics of further research.

Keywords: *EM algorithm; numerical integration; interactions; missing data*

We consider the problem of missing data in regression models with product-term predictors. In the educational and behavioral 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 in 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 et al., 2018; Preacher et al., 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 a 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 et al., 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 Gaussian random variables cannot itself be Gaussian. While some studies have shown that there may be some conditions under which this method is reasonable (Enders et al., 2014; Seaman et al., 2012), it is not guaranteed to provide unbiased estimates in general (Bartlett et al., 2015; Lüdtke et al., 2019; Zhang & Wang, 2017).

To address these issues, correctly specified methods have been developed. This is typically accomplished by 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 (Ibrahim, 1990). Hence, this technique can ensure compatibility between the substantive model of interest, and the overall joint distribution of the data (J. Liu et al., 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 et al., 2020).

The application of this technique however, has been largely focused on multiple imputation methods with Markov Chain Monte Carlo under a Bayesian paradigm (Kim et al., 2015; Lüdtke et al., 2020; Zhang & Wang, 2017). On the other hand, research in the frequentist framework 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 nonlinear models, numerical integration is known to suffer in accuracy and computational complexity as the number of dimensions increases (Hinrichs et al., 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.

The premise of the current research is to 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 we will demonstrate how to use analytic solutions in their place. These exact solutions will yield more accurate estimates relative to their approximate counterparts. Therefore, this research has two main goals: (a) to develop the theoretical motivation of the hybrid EM algorithm and (b) to empirically study the benefits of analytic solutions in practical data scenarios.

Model and Notation

Let $\mathbf{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ 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:

$$Y = \mathbf{d}(\mathbf{X})^T \boldsymbol{\beta} + \epsilon, \tag{1}$$

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

$$\mathbf{d}(\mathbf{X}) = \begin{bmatrix} 1 & \mathbf{X}^{\mathrm{T}} & \overrightarrow{X_j X_k}^T \end{bmatrix}^T, \tag{2}$$

where $\overrightarrow{g(\mathbf{X})}$ denotes the vector of all unique permutations of $g(\mathbf{X})$ 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, $\mathbf{d}(\mathbf{X})$ is a vector that augments $\mathbf{X}$ 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 $\mathbf{X}$ 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 $\mathbf{X}$. 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.

Missing Data Assumptions

To recast the data from a predictor-outcome distinction to a missing-observed distinction, we use the following notation. Denote an augmented data vector as $\mathbf{U} = \begin{bmatrix} Y & \mathbf{X}^T \end{bmatrix}^T$, which can be reordered as $\begin{bmatrix} \mathbf{U}_O^T & \mathbf{U}_M^T \end{bmatrix}^T$, where O is the index set of observed variables and M is the index set of missing variables. Further, the probability density/mass function of $\mathbf{U}$ is denoted $f(\mathbf{U})$ and parameterized generally by a vector $\boldsymbol{\theta}$, for which we write $f_{\boldsymbol{\theta}}(\mathbf{U})$. Then, let $\mathbf{R} \in \{0, 1\}^{p+1}$ be a binary random vector, which indicates whether the elements of $\mathbf{U}$ are observed, and has a probability distribution parameterized by the vector $\boldsymbol{\zeta}$. It is generally assumed that no variable in $\mathbf{U}$ will be completely missing in the sample.

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):

$$\mathbb{P}_{\boldsymbol{\zeta}}(\mathbf{R}|\mathbf{U}) = \mathbb{P}_{\boldsymbol{\zeta}}(\mathbf{R}|\mathbf{U}_O). \tag{3}$$

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

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 et al., 1977). The steps are as follows:

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

$$Q_{\boldsymbol{\theta}^{(t)}}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}^{(t)}}[\log f_{\boldsymbol{\theta}}(\mathbf{U})|\mathbf{U}_O] = \int_{\mathbf{u}_M} \log f_{\boldsymbol{\theta}}(\mathbf{U}) f_{\boldsymbol{\theta}^{(t)}}(\mathbf{u}_M|\mathbf{u}_O)\, d\mathbf{u}_M. \tag{4}$$

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

$$\boldsymbol{\theta}^{(t+1)} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, Q_{\boldsymbol{\theta}^{(t)}}(\boldsymbol{\theta}), \tag{5}$$

where we use the integration symbol with respect to a vector as shorthand for multiple integration or summation with respect to all elements of the vector, depending on if the random variable is continuous or discrete (e.g., $\int_{\mathbf{z}} f(\mathbf{z})\, d\mathbf{z} = \int_{z_1} \cdots \int_{z_p} f(\mathbf{z})\, dz_p \cdots dz_1$, for $\mathbf{z} \in \mathbb{R}^p$). 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). Further, standard errors can be obtained by numerically differentiating the EM iterations (Meng & Rubin, 1991) or the Fisher score function (Jamshidian & Jennrich, 2000).

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:

$$f(\mathbf{U}) = f(Y|\mathbf{X})f(\mathbf{X}), \tag{6}$$

where

$$\begin{aligned} f(\mathbf{X}) &= \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ f(Y|\mathbf{X}) &= \mathcal{N}\left(\mathbf{d}(\mathbf{x})^T\boldsymbol{\beta}, \sigma_\epsilon^2\right). \end{aligned} \tag{7}$$

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

$$f(\mathbf{U}) = \exp\left[\boldsymbol{\eta}(\boldsymbol{\theta})^T\mathbf{T}(\mathbf{U}) - A(\boldsymbol{\theta})\right], \tag{8}$$

which yields a Q-function of

$$\begin{aligned} Q_{\boldsymbol{\theta}^{(t)}}(\boldsymbol{\theta}) &= \mathbb{E}_{\boldsymbol{\theta}^{(t)}}[\log f_{\boldsymbol{\theta}}(\mathbf{U})|\mathbf{U}_O] \\ &= \mathbb{E}_{\boldsymbol{\theta}^{(t)}}\left[\boldsymbol{\eta}(\boldsymbol{\theta})^T\mathbf{T}(\mathbf{U}) - A(\boldsymbol{\theta})|\mathbf{U}_O\right] \\ &= \boldsymbol{\eta}(\boldsymbol{\theta})^T\mathbb{E}_{\boldsymbol{\theta}^{(t)}}[\mathbf{T}(\mathbf{U})|\mathbf{U}_O] - A(\boldsymbol{\theta}), \end{aligned} \tag{9}$$

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

$$\mathbf{T}(\mathbf{U}) = \\ \left[Y \quad Y\overrightarrow{X_j^T} \quad Y^2 \quad Y\overrightarrow{X_jX_k^T} \quad \overrightarrow{X_j^T} \quad \overrightarrow{X_jX_k^T} \quad \overrightarrow{X_j^{2T}} \quad \overrightarrow{X_jX_kX_l^T} \quad \overrightarrow{X_j^2X_k^T} \quad \overrightarrow{X_jX_kX_lX_m^T} \quad \overrightarrow{X_j^2X_kX_l^T} \quad \overrightarrow{X_j^2X_k^{2T}}\right]^T. \tag{10}$$

The derivation of $\mathbf{T}(\mathbf{U})$ has been relegated to Supplemental Appendix A (available in the online version of this article). We also derive the maximizers of the Q-function in Supplemental Appendix B (available in the online version of this article).

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 $\mathbf{T}(\mathbf{U})$ may appear complex and the possible missing data patterns for $\mathbb{E}[\mathbf{T}(\mathbf{U})|\mathbf{U}_O]$ 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 $\mathbf{X}$ has any missingness pattern.
- **MDP 2:** Y is observed and $\mathbf{X}$ is patterned such that no product terms are fully missing.
- **MDP 3:** Y is observed and $\mathbf{X}$ is patterned such that one or more product terms are fully missing.

We will provide the methods of calculating $\mathbb{E}[\mathbf{T}(\mathbf{U})|\mathbf{U}_O]$ 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.

Missing Data Pattern 1

MDP 1 is concerned with the case when Y is missing, and $\mathbf{X}$ can take on any missingness pattern. We will show that all elements of $\mathbb{E}[\mathbf{T}(\mathbf{U})|\mathbf{U}_O]$ 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 $\mathbf{X}$ and do not have a Y term. In Equation 10 these are the latter 8 (of 12) entries of $\mathbf{T}(\mathbf{U})$. Note that these entries are all products of the elements of $\mathbf{X}$ (e.g., $X_jX_kX_lX_m$ or $X_j^2X_k^2$). Further, Y is missing in this MDP, so we have $\mathbf{U}_O = \mathbf{X}_O$. Thus, under MDP 1, we can more generally express the elements of $\mathbb{E}[\mathbf{T}(\mathbf{U})|\mathbf{U}_O]$ that only depend on $\mathbf{X}$ as

$$\mathbb{E}\left[\prod_{i\in M} X_i^{a_i} |\mathbf{X}_O\right], \tag{11}$$

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 $\mathbf{X}_M|\mathbf{X}_O$ 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 $\mathbf{X}$ be a Gaussian random vector distributed as $\mathbf{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then any product moment of the form $\mathbb{E}\left[\prod_{i=1}^{p} X_i^{a_i}\right]$ 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

$$M_{\mathbf{X}}(\mathbf{t}) = \exp\left(\mathbf{t}^T\boldsymbol{\mu} + \frac{1}{2}\mathbf{t}^T\boldsymbol{\Sigma}\mathbf{t}\right). \tag{12}$$

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

$$\frac{\partial^a}{\prod_{i=1}^{p} \partial t_i^{a_i}} M_{\mathbf{X}}(\mathbf{t})|_{\mathbf{t}=0} = \mathbb{E}\left[\prod_{i=1}^{p} X_i^{a_i}\right], \tag{13}$$

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

From here, the expectation in the form of Equation 11 can be obtained by seeing that the parameters of $\mathbf{X}_M | \mathbf{X}_O \sim \mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$ are

$$\begin{aligned} \boldsymbol{\mu}_c &= \boldsymbol{\mu}_M + \boldsymbol{\Sigma}_{MO}\boldsymbol{\Sigma}_O^{-1}(\mathbf{x}_O - \boldsymbol{\mu}_O) \\ \boldsymbol{\Sigma}_c &= \boldsymbol{\Sigma}_M - \boldsymbol{\Sigma}_{MO}\boldsymbol{\Sigma}_O^{-1}\boldsymbol{\Sigma}_{OM}, \end{aligned} \tag{14}$$

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

Among the remaining four 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 $\mathbb{E}[YX_j^aX_k^b|\mathbf{X}_O]$ for $a, b \in \{0, 1\}$. Then we can re-write this quantity as

$$\mathbb{E}[YX_j^aX_k^b|\mathbf{X}_O] = \mathbb{E}[\mathbf{d}(\mathbf{X})^T\boldsymbol{\beta}X_j^aX_k^b|\mathbf{X}_O], \tag{15}$$

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

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

$$\begin{aligned}
\mathbb{E}\left[Y^2|\mathbf{X}_O\right] &= \mathbb{E}\left[\mathbb{E}\left[Y^2|\mathbf{X}_M,\mathbf{X}_O\right]|\mathbf{X}_O\right] \\
&= \mathbb{E}\left[\mathrm{Var}(Y|\mathbf{X}) + \mathbb{E}[Y|\mathbf{X}]^2|\mathbf{X}_O\right] \\
&= \mathbb{E}\left[\sigma_\epsilon^2 + \left(\boldsymbol{\beta}^T\mathbf{d}(\mathbf{X})\right)^2|\mathbf{X}_O\right] \\
&= \sigma_\epsilon^2 + \mathbb{E}\left[\boldsymbol{\beta}^T\mathbf{d}(\mathbf{X})\mathbf{d}(\mathbf{X})^T\boldsymbol{\beta}|\mathbf{X}_O\right] \\
&= \sigma_\epsilon^2 + \mathbb{E}\left[\sum_{i,j}\beta_i\beta_j\left(\mathbf{d}(\mathbf{X})\mathbf{d}(\mathbf{X})^T\right)_{ij}|\mathbf{X}_O\right],
\end{aligned} \tag{16}$$

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

Missing Data Pattern 2

MDP 2 considers the scenario where Y is observed and $\mathbf{X}$ is patterned such that no product terms are fully missing. Equivalently, we can say that $\mathbf{X}$ is patterned such that at least one X_j is observed in every product term. In this situation, $\mathbf{X}_M|Y,\mathbf{X}_O$ takes on a multivariate Gaussian distribution, and thus $\mathbb{E}[\mathbf{T}(\mathbf{U})|\mathbf{U}_O]$ can be completely solved analytically. To see why this is the case, let us re-write the analytical model in Equation 1 under the assumptions of MDP 2. First, note that we can separate terms by observed variables and missing variables:

$$\begin{aligned}
Y &= \mathbf{d}(\mathbf{X})^T\boldsymbol{\beta} + \epsilon \\
&= \beta_0 + \sum_{j=1}^{p}\beta_j X_j + \sum_{j\neq k}\beta_{jk}X_jX_k + \epsilon \\
&= \beta_0 + \sum_{j\in O}\beta_j X_j + \sum_{j\in M}\beta_j X_j + \sum_{j,k\in O}\beta_{jk}X_jX_k + \sum_{j\in M,\,k\in O}\beta_{jk}X_jX_k + \epsilon.
\end{aligned} \tag{17}$$

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

$$\begin{aligned}
\tilde{\beta}_0 &:= \beta_0 + \sum_{j\in O}\beta_j X_j + \sum_{j,k\in O}\beta_{jk}X_jX_k \\
\tilde{\beta}_j &:= \beta_j + \sum_{k\in O}\beta_{jk}X_k, \text{for } j\in M.
\end{aligned} \tag{18}$$

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

$$Y = \tilde{\beta}_0 + \sum_{j \in M} \tilde{\beta}_j X_j + \epsilon, \tag{19}$$

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

$$X_m = \frac{Y - \tilde{\beta}_0 - \sum_{j \in M \setminus m} \tilde{\beta}_j X_j - \epsilon}{\tilde{\beta}_m}. \tag{20}$$

Thus, any X_m is a linear combination of other Gaussian random variables, therefore must be Gaussian itself. Hence, $\mathbf{X}_M | Y, \mathbf{X}_O$ follows a multivariate Gaussian distribution. The derivation of the exact density $f(\mathbf{X}_M | Y, \mathbf{X}_O)$ can be found in Supplemental Appendix D (available in the online version of this article).

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

Missing Data Pattern 3

MDP 3 concerns the case where Y is observed and $\mathbf{X}$ is patterned such that product terms are fully missing. In this situation, the entries of $\mathbb{E}[\mathbf{T}(\mathbf{U}) | \mathbf{U}_O]$ may be difficult to derive analytically, or admit no closed form. We will show this by showing how the distribution of $\mathbf{X}_m | Y, \mathbf{X}_O$ would be characterized. As in our argument for MDP 2, we will separate the terms of the regression model by observed variables, missing variables, and terms with one of each:

$$\begin{aligned} Y &= \mathbf{d}(\mathbf{X})^T \boldsymbol{\beta} + \epsilon \\ &= \beta_0 + \sum_{j \in M} \beta_j X_j + \sum_{j \in O} \beta_j X_j + \sum_{j,k \in M} \beta_{jk} X_j X_k + \sum_{j \in M, k \in O} \beta_{jk} X_j X_k + \sum_{j,k \in O} \beta_{jk} X_j X_k + \epsilon. \end{aligned} \tag{21}$$

Then we treat the observed variables as constants and absorb them into the $\boldsymbol{\beta}$ coefficients as follows:

$$\begin{aligned} \tilde{\beta}_0 &:= \beta_0 + \sum_{j \in O} \beta_j X_j + \sum_{j,k \in O} \beta_{jk} X_j X_k \\ \tilde{\beta}_j &:= \beta_j + \sum_{k \in O} \beta_{jk} X_k, \text{for } j \in M, \end{aligned} \tag{22}$$

then for any $m \in M$ the model can be re-written as

$$
\begin{aligned}
Y &= \tilde{\beta}_0 + \sum_{j\in M} \tilde{\beta}_j X_j + \sum_{j,k\in M} \beta_{jk} X_j X_k + \epsilon \\
&= \tilde{\beta}_0 + \left(\tilde{\beta}_m + \sum_{j\in M\backslash m} \beta_{jm} X_j \right) X_m + \sum_{j\in M\backslash m} \tilde{\beta}_j X_j + \sum_{j,k\in M\backslash m} \beta_{jk} X_j X_k + \epsilon,
\end{aligned}
\tag{23}
$$

which implies that

$$
X_m = \frac{Y - \tilde{\beta}_0 - \sum_{j\in M\backslash m} \tilde{\beta}_j X_j - \sum_{j,k\in M\backslash m} \beta_{jk} X_j X_k - \epsilon}{\tilde{\beta}_m + \sum_{j\in M\backslash m} \beta_{jm} X_j}. \tag{24}
$$

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

$$
\begin{aligned}
\mathbb{E}[\mathbf{T}(\mathbf{U})|\mathbf{U}_O] &= \frac{\int_{\mathbf{u}_M} \mathbf{T}(\mathbf{u}_M, \mathbf{u}_O) f(\mathbf{u}_M|\mathbf{u}_O)\, d\mathbf{u}_M}{\int_{\mathbf{u}_M} f(\mathbf{u}_M|\mathbf{u}_O)\, d\mathbf{u}_M} \\
&= \frac{f(\mathbf{u}_O)}{f(\mathbf{u}_O)} \frac{\int_{\mathbf{u}_M} \mathbf{T}(\mathbf{u}_M, \mathbf{u}_O) f(\mathbf{u}_M, \mathbf{u}_O)\, d\mathbf{u}_M}{\int_{\mathbf{u}_M} f\left(\mathbf{u}_{Mg}, \mathbf{u}_O\right) d\mathbf{u}_M} \\
&\approx \frac{\sum_{g=1}^{G} \mathbf{T}\left(\mathbf{u}_{Mg}, \mathbf{u}_O\right) f\left(\mathbf{u}_{Mg}, \mathbf{u}_O\right)}{\sum_{g=1}^{G} f\left(\mathbf{u}_{Mg}, \mathbf{u}_O\right)},
\end{aligned}
\tag{25}
$$

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

Summary of Results

We propose to construct a hybrid EM method that uses the analytic results derived in this section for MDPs 1 and 2, and numerical integration for MDP 3. To incorporate these results into a hybrid EM algorithm, first consider the Q-function from the perspective of a sample. Let a case index be denoted from $i = 1, \ldots, n$. Then the Q-function can be written as

$$
Q_{\boldsymbol{\theta}^{(t)}}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{\theta}^{(t)}}[f_{\boldsymbol{\theta}}(\mathbf{u}_i)|\mathbf{u}_{iO}]. \tag{26}
$$

The key observation of this characterization is that the conditional expectation can be taken in a case-wise manner. Thus the calculation of the conditional

Algorithm 1. The Hybrid EM Algorithm

```
Input : Start values $\boldsymbol{\theta}^{(0)}$, observed data $\mathbf{U}_O$, model specification
Determine MDP. Categorize each $\mathbf{u}_{iO}$ into MDPs, 1, 2, or 3 by comparing its
missingness pattern to the model specification;
Set $t \leftarrow 0$;
repeat
    Hybrid E-Step. Calculate $\mathbb{E}_{\boldsymbol{\theta}^{(t)}}[\mathbf{T}(\mathbf{u}_i)|\mathbf{u}_{iO}]$ for $i = 1, \ldots, n$:
        if $\mathbf{u}_{iO}$ is MDP 1 then
            Apply Equations 15 and 16 for expectations with $Y$;
            Apply Lemma 1 with parameters from Equation 14 for expectations
            without $Y$;
        if $\mathbf{u}_{iO}$ is MDP 2 then
            Apply Lemma 1 with parameters from Equation D7 for all expectations;
        if $\mathbf{u}_{iO}$ is MDP 3 then
            Apply Equation 25 for all expectations;
    M-Step. Set $\boldsymbol{\theta}^{(t+1)} \leftarrow \operatorname{argmax}_{\boldsymbol{\theta}} Q_{\boldsymbol{\theta}^{(t)}}(\boldsymbol{\theta})$, see Supplemental Appendix B for
    closed-form maximizers;
    if $\max|\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^{(t)}| \leqslant \varepsilon$, a small convergence criterion then break repeat;
    else Set $t \leftarrow t + 1$;
Output: Parameter Estimates $\hat{\boldsymbol{\theta}} \leftarrow \boldsymbol{\theta}^{(t+1)}$
```

expectation for each case can differ depending on the missingness pattern, which dictates if numerical integration is necessary or not. That is, the analytic solutions described earlier can be used if case *i* has MDPs 1 or 2, and numerical integration can be used if it has MDP 3. A more formal description of the complete algorithm is described in Algorithm 1.

Empirical Studies

Given that the hybrid EM algorithm minimizes the use of numerical approximations, we now investigate the impact this has on data analysis. We do this over three empirical studies: (a) a basic simulation study varying several characteristics of the data, (b) a simulation study using real data with behavioral measures, and (c) a study on power.

Basic Simulation Study

For a basic simulation study, we sought to study estimator performance over several settings:

- Estimation methods: hybrid EM (HYB), full numerical integration (NI), and complete data least squares (CD) as a baseline.

- Sample size (n): 100, 250, 500, and 1,000.
- Proportion of missingness (φ_{MIS}): 0.10, 0.20, and 0.30.
- $r = 100$ replications per condition combination.

Our primary interest was to study the effect of analytic EM iterations versus NI. As such, we specifically generated missingness patterns from MDP 1 and MDP 2, which resulted in the HYB method using only analytic integration. The NI method uses numerical integration regardless of missing data pattern, and we used the Riemann midpoint method following Robitzsch and Lüdtke (2021) and used 40 grid points spread uniformly between ± 4 standard deviations from the marginal means. The least-squares estimates under listwise deletion were used as the start values for both the HYB and NI methods. Note that for completeness, we also investigated the effect of varying the prevalence of MDP 3 (φ_{MDP3}). These results can be found in Supplemental Appendix E (available in the online version of this article).

Data Generation

The parameters for **X** were generated in the following way:

$$\begin{aligned} &\boldsymbol{\mu} \sim \mathcal{U}_p(-3, 3) \\ &\boldsymbol{\Sigma} = \mathbf{DCD}, \end{aligned} \tag{27}$$

where **D** is a diagonal matrix of standard deviations, with the diagonal distributed as $\mathcal{U}_p\left(1, \sqrt{3}\right)$ and **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, **X** was sampled from $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. To reflect a higher number of predictors that is more common in real data we set $p = 7$.

For the regression model, three of the seven predictors were chosen at random to form product terms in $\mathbf{d}(\mathbf{X})$, for a total of $d = 10$ design variables. Then, an adjusted R^2 parameter and $\boldsymbol{\beta}$ vector were simulated using

$$\begin{aligned} &R_a^2 \sim \mathcal{U}_1(0.1, 0.5) \\ &\boldsymbol{\beta} \sim \mathcal{U}_d(-3, 3). \end{aligned} \tag{28}$$

Given a sample of **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 Supplemental Appendix F in the online version of the journal for details). This allows us to draw error terms with $\epsilon \sim \mathcal{N}\left(0, \sigma_\epsilon^2\right)$ and calculate the outcome with $Y = \mathbf{d}(\mathbf{X})^T \boldsymbol{\beta} + \epsilon$.

Once **X** and Y were generated, the observed data indicator **R** was generated under a MAR mechanism. This was done by randomly selecting a non-product variable in **X** to serve as an always-observed auxiliary variable designated X_a, which determined missingness in all other variables. This ensured the MAR assumption was always met. Using an intermediate latent propensity variable based on X_a, we determined the cases that were designated to contain missing values according to φ_{MIS}. Then half of these cases were allocated to MDP 1 and the other half to MDP 2. The exact mathematical details of this procedure can be found in Supplemental Appendix G (available in the online version of this article).

Performance Metrics

To evaluate performance, we calculated bias and mean square error (MSE) quantities aggregated within coefficient vectors as follows:

$$\begin{aligned} \text{Aggregated Bias} &:= \sum_{j=1}^{d} \frac{\left(\hat{\beta}_j - \beta_j\right)}{d} \\ \text{Aggregated MSE} &:= \sum_{j=1}^{d} \frac{\left(\hat{\beta}_j - \beta_j\right)^2}{d}. \end{aligned} \tag{29}$$

We then examined these quantities averaged over $r = 100$ replicated datasets, per simulation condition.

Results

Plots of aggregated bias and aggregated MSE are displayed in Figure 1. Across all conditions, the HYB method had a lower MSE than NI. While the MSE increased with φ_{MIS} for both methods, the gap between HYB and NI also increased, indicating that the HYB method is more robust to an increase in missing data. As would be expected with maximum likelihood theory, the MSE of CD, HYB, and NI decreased with *n* and are generally unbiased across all conditions. While we did not formally study the elapsed computation time, we note that in our implementation both the HYB and NI methods are inconsequentially fast. Across all conditions, both methods averaged below one-tenth of a second, with maximum run times of 1.31 s for HYB and 2.19 s for NI.

Real Data Study

In this study, we compared the practical use of the HYB and NI methods by using data from real behavioral measures. Educational and behavioral data may be discrete or skewed, contrary to the multivariate Gaussian assumption that we utilize for the predictors. The Gaussian assumption is typically not required when the data are complete, but utilized here to accommodate missingness in

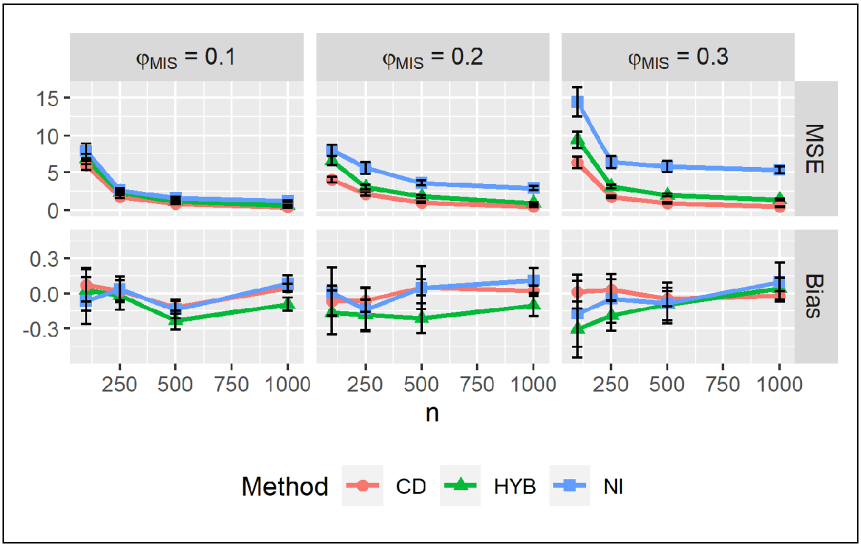

FIGURE 1. *Aggregated MSE and bias by* n, φ_{MIS}, *and method. Error bars indicate* ± *1 standard error.*

Note. MSE = mean square error.

the predictors. Therefore, we study the robustness of the multivariate normality assumption when the predictors are instead discretized or skewed as real data would be. We do this by taking a set of real, complete, behavioral data and setting it as our population of interest. We then sample from this population by non-parametric bootstrap and then artificially insert missingness to the bootstrapped datasets. The performance of the HYB and NI methods are then evaluated in their ability to recover the regression parameters from these bootstrapped datasets.

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 to provide the basis of our data (n = 1,011).

We considered a linear model of conduct disorder as a function of attention deficit hyperactivity disorder (ADHD), depression (DEP), their product-term (ADHD × DEP), controlled for by anxiety (ANX), and oppositional defiant disorder (ODD). Previous work has shown comorbidity among these variables

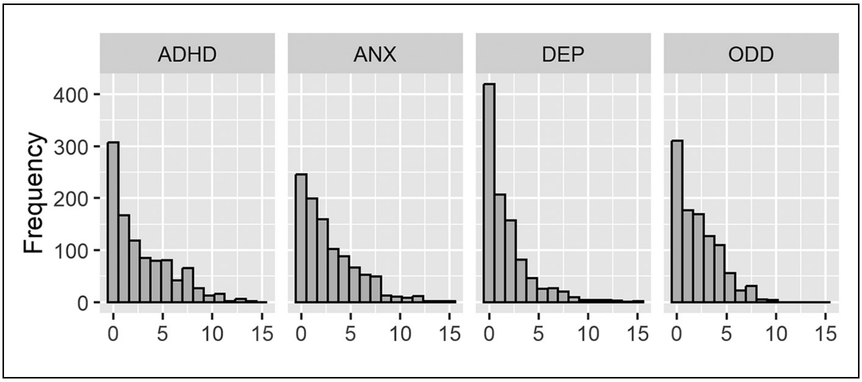

FIGURE 2. *Histograms of the predictors variables for the real data study.*
Note. MSE = mean square error.

(Angold et al., 1999; Jensen et al., 1997). Measures were taken using summary scores of the child behavior checklist (Achenbach & Rescorla, 2001). These scores are fairly skewed and discrete, residing on the integers 0 to 15. We display histograms of these scores in Figure 2.

For each bootstrapped data set, missingness was inserted using the same MAR generating procedure as the previous basic simulation study (equal proportions of MDPs 1 and 2) but fixed $\varphi_{MIS} = 0.2$. For performance metrics, we evaluated the empirical bias and empirical MSEs per coefficient. That is we computed

$$\begin{aligned}\text{Empirical Bias} &:= \hat{\beta}_j - \beta_j \\ \text{Empirical MSE} &:= \left(\hat{\beta}_j - \beta_j\right)^2.\end{aligned} \tag{30}$$

These metrics were evaluated over $r = 100$ bootstrapped repetitions. The least-squares estimates of the original complete data set were considered the true parameters. These estimates are displayed in Table 1.

Boxplots of the results are displayed in Figure 3. On average, the biases from all methods were close to zero across all parameter estimates, with the only exception being the NI method underestimating the β_{ODD} parameter by a factor of 1.38. The HYB method had lower MSE than the NI method across all parameters except for β_{DEP}, and was substantially lower for the intercept (β_0) and β_{ODD}. These results highlight the improved performance of the HYB method in real data scenarios and its robustness to violating the Gaussian assumption over the predictors in practice.

TABLE 1.
Estimates, Standard Errors, t-*Values, and* p-*Values of the Original Data Least-Squares Coefficients for the Real Data Study*

Coefficient	Estimate	*SE*	*t*-Value	*p*-Value
β_0	−0.170	0.088	−1.936	.053
β_{ODD}	0.573	0.033	17.583	.000
β_{ANX}	−0.033	0.024	−1.396	.163
β_{ADHD}	0.103	0.026	4.033	.000
β_{DEP}	0.014	0.045	0.312	.755
$\beta_{ADHD \times DEP}$	0.019	0.006	3.428	.001

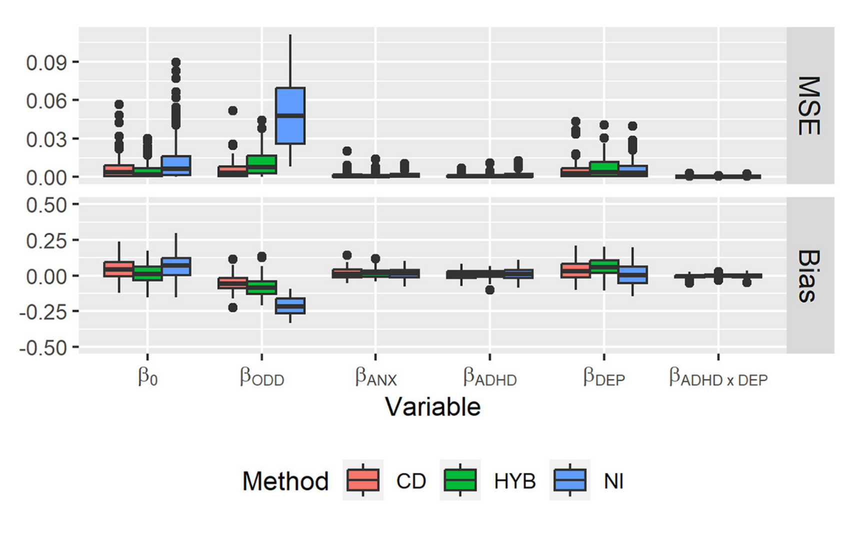

FIGURE 3. *Empirical MSE and bias by method over all regression parameters.*
Note. MSE = mean square error.

Power Study

Given that analytic iterations confer improved MSE as demonstrated by the basic simulation study and the real data study, we examine how this may translate to increased power. We used the same dataset as the real data study as a basis, and once again set the least-squares estimates of the complete data as the true coefficient vector. Then we used a parametric bootstrap to generate error terms such that the alternative hypothesis will be true on the interaction coefficient before missingness is inserted with an approximate power of 0.80 (in a one-tailed t-test with $\alpha = 0.05$). That is, we generated complete datasets using

$$\begin{aligned} &\tilde{\epsilon}_i \sim \mathcal{N}\left(0, \sigma^2_{\tilde{\epsilon}}\right) \\ &\tilde{Y}_i = \mathbf{d}(\mathbf{x}_i)^T \beta_{\mathrm{LS}} + \tilde{\epsilon}_i, \end{aligned} \tag{31}$$

for all $i = 1, \ldots, n$, and where $\mathbf{x}_i$ are the original data predictors and β_{LS} are the original least-squares estimates. The parameter $\sigma^2_{\tilde{\epsilon}}$ is set to a value such that the power on the interaction coefficient is .80 (further details are available in Supplemental Appendix H in the online version of the journal). This allowed us to generate complete datasets where the hypothesized model is true with the desired amount of power determined by $\sigma^2_{\tilde{\epsilon}}$. Missingness was then inserted using the same scheme as the previous real data study. For each method, t-statistics were calculated with

$$\mathrm{t} = \frac{\hat{\beta}_{\mathrm{ADHD} \times \mathrm{DEP}}}{\sqrt{\widehat{\mathrm{Var}}\left(\hat{\beta}_{\mathrm{ADHD} \times \mathrm{DEP}}\right)}}, \tag{32}$$

where we calculated the standard errors for the HYB and NI by numerically differentiating the Fisher score function (Jamshidian & Jennrich, 2000). Then the estimated power was taken to be the proportion of times that the t-statistic exceeded a critical t-value based on $\alpha = 0.05$ and $\mathrm{df} = n - d$. Thus we examined, over 100 replications, how much will the HYB and NI methods recover the original power after missingness is inserted.

We display the estimated power of each method in Figure 4a. Complete data least-squares estimates are included as a control comparison. The estimated power was 0.82 for the complete data, 0.77 for HYB, and 0.42 for NI. As is expected by the insertion of missingness, both the HYB and NI methods show reduced power. However, the HYB method was substantially more robust, only having a reduced power of 0.05, whereas the NI method had a reduced power of 0.40 relative to the complete data power. Consistent with the results of the previous two studies, this is attributable to the increased estimation variance of the NI method, which can be seen in Figure 4b. These box plots display magnitudes of $\hat{\beta}_{\mathrm{ADHD} \times \mathrm{DEP}}$ for each method. We show that the complete data and HYB methods have similar variances, whereas the NI method is visibly larger.

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

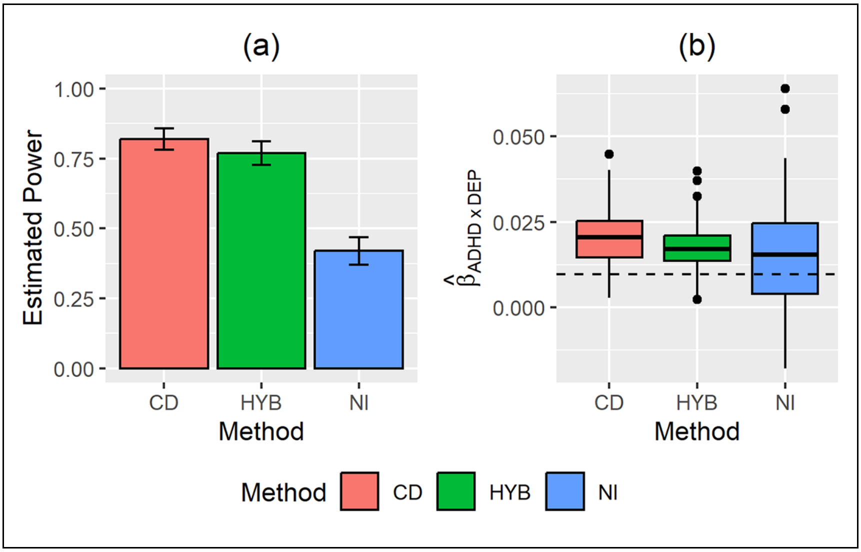

FIGURE 4. *Estimated power of the three methods (a) and the box plots of* $\hat{\beta}_{ADHD \times DEP}$ *(b). Error bars on estimated power denote* ± 1 *standard error. The dotted line denotes the critical value of* $\hat{\beta}_{\text{ADHD} \times \text{DEP}}$.

algorithm, where analytic *E*-steps were used whenever possible and NI was used otherwise.

Comparing the hybrid method to NI, several themes arose across the three simulation studies. First, both methods showed very little bias, even when the predictors were non-normal. Second, the hybrid method outperformed NI primarily in terms of estimation variability, in both normal and non-normal scenarios. And third, this reduction in variability can translate to substantial increases in power.

Tempering these promising results is the fact that the hybrid method used in this study is specific only to product-term regression models. Certain aspects of this research may generalize well toward other models, for example, our analysis of MDP 1 may readily generalize into polynomial predictors. For other regression designs, such as the generalized linear model and/or discrete predictors, further study will be required to parse their idiosyncratic characteristics. In contrast, the NI method remains general and readily applicable.

Another avenue for future research is to investigate other missingness mechanisms. In the current study, we focused on an ignorable missingness mechanism. However recently, frameworks have been developed to determine situations where consistent estimates can be obtained when the missingness mechanism is non-ignorable (Mohan & Pearl, 2021; Rabe-Hesketh & Skrondal, 2023). Also, the robustness of these methods to distributional violations could be more systematically studied through more carefully designed experiments that vary aspects such as discreteness and skewness.

From a computational perspective, the scalability of these methods can be further investigated by varying the number of variables. At higher proportions of missingness, a more principled strategy for start values may also be called for. Additionally, the hybrid approach may also be incorporated into Monte Carlo methods of integration, including Gibbs and Metropolis-Hasting variants (Levine & Casella, 2001; Wei & Tanner, 1990). The NI method may also be improved via other approximating functions (Q. Liu & Pierce, 1994) and/or by adaptive methods (Rabe-Hesketh et al., 2002). In the context of the current research, these topics are amenable directions for further study in missing data methods for regression models, and may also result in higher power to detect effects.

Declaration of Conflicting Interests

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Dale S. Kim https://orcid.org/0000-0002-5365-3748

References

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

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

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.

Clark, D. B., Fisher, C. B., Bookheimer, S., Brown, S. A., Evans, J. H., Hopfer, C., Hudziak, J., Montoya, I., Murray, M., Pfefferbaum, A., & 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.

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.

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.

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

Hayes, A. F. (2018). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach* (2nd ed.). 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., & 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*. 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*, *55*(3), 361–381.

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.

Mohan, K., & Pearl, J. (2021). Graphical models for processing missing data. *Journal of the American Statistical Association*, *116*(534), 1023–1037. https://doi.org/10.1080/01621459.2021.1874961

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.

Rabe-Hesketh, S., & Skrondal, A. (2023). Ignoring non-ignorable missingness. *Psychometrika*, *88*(1), 31–50. https://doi.org/10.1007/s11336-022-09895-1

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* (R package version 1.5-8). https://CRAN.R-project.org/package=mdmb

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

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. 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.

Volkow, N. D., Koob, G. F., Croyle, R. T., Bianchi, D. W., Gordon, J. A., Koroshetz, W. J., Pérez-stable, E. J., Riley, W. T., Bloch, M. H., Conway, K., Deeds, B. G., Dowling, G. J., Grant, S., Howlett, K. D., Matochik, J. A., Morgan, G. D., Murray, M. M., Noronha, A., Spong, C. Y., … 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.

Author

DALE S. KIM is an Assistant Professor in the Department of Statistics & Data Science at University of California, Los Angeles, Math Sciences Building 8125, 520 Portola Plaza, Los Angeles, CA 90095, e-mail: daleskim@stat.ucla.edu. His research interests are factor analysis, missing data, and computational statistics.

Manuscript received May 3, 2024
Revision received September 27, 2024
Accepted November 7, 2024