


Article

# Consequences of Model Misspecification for Maximum Likelihood Estimation with Missing Data

Richard M. Golden <sup>1,\*</sup> , Steven S. Henley <sup>2,3,4</sup>, Halbert White <sup>5,†</sup> and T. Michael Kashner <sup>3,4,6</sup>

<sup>1</sup> School of Behavioral and Brain Sciences, GR4.1, 800 W. Campbell Rd., University of Texas at Dallas, Richardson, TX 75080, USA

<sup>2</sup> Martingale Research Corporation, 101 E. Park Blvd., Suite 600, Plano, TX 75074, USA

<sup>3</sup> Department of Medicine, Loma Linda University School of Medicine, Loma Linda, CA 92357, USA

<sup>4</sup> Center for Advanced Statistics in Education, VA Loma Linda Healthcare System, Loma Linda, CA 92357, USA

<sup>5</sup> Department of Economics, University of California San Diego, La Jolla, CA 92093, USA

<sup>6</sup> Office of Academic Affiliations (10X1), Department of Veterans Affairs, 810 Vermont Ave. NW, Washington, DC 20420, USA

\* Correspondence: golden@utdallas.edu; Tel.: +1-972-883-2423

† Halbert White sadly passed away before this article was published.

Received: 22 October 2018; Accepted: 29 August 2019; Published: 5 September 2019



**Abstract:** Researchers are often faced with the challenge of developing statistical models with incomplete data. Exacerbating this situation is the possibility that either the researcher's complete-data model or the model of the missing-data mechanism is misspecified. In this article, we create a formal theoretical framework for developing statistical models and detecting model misspecification in the presence of incomplete data where maximum likelihood estimates are obtained by maximizing the observable-data likelihood function when the missing-data mechanism is assumed ignorable. First, we provide sufficient regularity conditions on the researcher's complete-data model to characterize the asymptotic behavior of maximum likelihood estimates in the simultaneous presence of both missing data and model misspecification. These results are then used to derive robust hypothesis testing methods for possibly misspecified models in the presence of Missing at Random (MAR) or Missing Not at Random (MNAR) missing data. Second, we introduce a method for the detection of model misspecification in missing data problems using recently developed Generalized Information Matrix Tests (GIMT). Third, we identify regularity conditions for the Missing Information Principle (MIP) to hold in the presence of model misspecification so as to provide useful computational covariance matrix estimation formulas. Fourth, we provide regularity conditions that ensure the observable-data expected negative log-likelihood function is convex in the presence of partially observable data when the amount of missingness is sufficiently small and the complete-data likelihood is convex. Fifth, we show that when the researcher has correctly specified a complete-data model with a convex negative likelihood function and an ignorable missing-data mechanism, then its strict local minimizer is the true parameter value for the complete-data model when the amount of missingness is sufficiently small. Our results thus provide new robust estimation, inference, and specification analysis methods for developing statistical models with incomplete data.

**Keywords:** asymptotic theory; ignorable; Generalized Information Matrix Test; misspecification; missing data; nonignorable; sandwich estimator; specification analysis

## 1. Introduction

Researchers are often faced with the challenge of developing statistical models with incomplete data (Little and Rubin 2002; Molenberghs et al. 2014; Rubin 1976). Exacerbating this situation is the possibility

that the researcher's complete-data model or model of the missing-data mechanism is misspecified. The objective of this article is to formally explore the consequences of model misspecification in the presence of incomplete (missing) data for statistical models that utilize maximum likelihood estimation (MLE) (Fomby and Hill 2003).

**Missing Data Problem.** The missing data problem is prevalent throughout economics (Abrevaya and Donald 2017; Breunig 2019; Fomby and Hill 1998; McDonough and Millimet 2016; Miller 2010; Wooldridge 2004). Further, missing data is ubiquitous in other fields of science, engineering (Markovsky 2017), and machine learning (Leke and Marwala 2019). This includes clinical trials and health sciences analyses (e.g., Enders 2010; Little et al. 2012; Molenberghs and Kenward 2007; Zhou et al. 2014), survey data analysis (e.g., Gmel 2001; Troxel et al. 1998), regression analysis (e.g., Graham et al. 1997; Greenland and Finkle 1995), verification bias (e.g., Harel and Zhou 2006; Kosinski and Barnhart 2003a, 2003b), hierarchical modeling (e.g., Agresti 2002, chp. 12), and mixed modeling (e.g., Verbeke and Lesaffre 1997). Moreover, latent variable models arising in factor analysis and structural equation modeling contexts (e.g., Arminger and Sobel 1990; Gallini 1983), hidden Markov chain models (e.g., McLachlan and Krishnan 1997; Visser 2011), mixed Markov field models (e.g., Fridman 2003) and hidden Markov random field models (HMRF) (e.g., Ryden and Titterington 1998) are interpretable as missing data models where the "hidden states" correspond to the missing data. Additionally, unsupervised and temporal reinforcement learning methods relevant for building sophisticated behavioral learning process models are naturally represented as partially observable Markov decision processes (e.g., Littman 2009).

**Model Misspecification.** The problems of estimation and inference in the presence of model misspecification are important for several reasons. First, model misspecification may be present in many, if not most, situations; and so robust methods that address the assumption of correct specification are necessary (White 1980, 1982, 1994; Golden 1995, 1996, 2000, 2003). While a correctly specified model is always desirable, in many fields such as econometrics, medicine, and psychology, some degree of model misspecification may be inevitable despite the researcher's best efforts (e.g., White 1980, 1982, 1994). Thus, the development and application of robust methods (Golden et al. 2013, 2016; Henley et al. 2019) that address the challenges posed by model misspecification (e.g., White 1980, 1982, 1994) has been and continues to be an active area of research (e.g., see Fomby and Hill 2003; Hardin 2003, for relevant reviews). Second, situations arise where the Quasi-Maximum Likelihood Estimates (QMLE) converge to the true parameter value despite the presence of model misspecification. For example, the QMLE can be shown to be consistent to the true parameter value for both linear and nonlinear exponential family regression models even though only the conditional expectation of the response variable given the predictors (covariates) is correctly specified (e.g., Gourieroux et al. 1984; Royall 1986; Wedderburn 1974; Wei 1998; White 1994, Corollary 5.5, p. 67). Consistent parameter estimation of the true parameter values of the researcher's model in the complete data case may also occur for misspecified models where: (i) heteroscedasticity is present (e.g., Verbeek 2008, sec. 6.3), (ii) the random effects distribution is misspecified in linear hierarchical models (e.g., Verbeke and Lesaffre 1997), or (iii) correlations among dependent observations are misspecified (e.g., Hosmer and Lemeshow 2000, pp. 315–317; Liang and Zeger 1986; Wall et al. 2005; Vittinghoff et al. 2012). Third, in more complicated missing data situations, consistent estimation of the true parameter values is possible in linear structural equation models even though only the first two moments have been correctly specified (e.g., Arminger and Sobel 1990), and in longitudinal time-series modeling even though dependent observations are approximately modeled as independent (Parzen et al. 2006; Troxel et al. 1998; Zhao et al. 1996).

### 1.1. Maximum Likelihood Estimation for Models with Partially Observable Data

**Representing Partially Observable Data Generating Processes.** In the selection model framework for representing partially observable data generating processes (Rubin 1976; Little 1994; Molenberghs et al. 1998; Little and Rubin 2002), it is assumed that nature creates a complete-data record

(observation) by sampling from the complete-data Data Generating Process (DGP). The complete-data record containing the observation's values is then decimated by a pattern of missingness sampled from the missing-data mechanism, thus hiding those values in the complete-data record. Rubin (1976) defined three types of missing-data mechanisms. A missing-data mechanism is termed Missing At Random (MAR) when the probability distribution of the pattern of missingness is functionally dependent only upon the observed data. A special case of MAR, called Missing Completely at Random (MCAR), occurs when the probability distribution of the pattern of missingness is not functionally dependent on either observed or unobserved data. Missing data generating processes that are not MAR are termed Missing Not at Random MNAR (i.e., not MAR), also called NMAR. The probability distribution of the pattern of missingness for an MNAR missing-data mechanism is functionally dependent on unobservable data.

A strategy for representing the pattern of missing values in data sets that supports utilizing maximum likelihood estimation is to create a collection of  $d$  binary indicator variables,  $\mathbf{h}_i = [h_{i,1}, \dots, h_{i,d}]^T$  for the  $i$ th data record  $\mathbf{x}_i = [x_{i,1}, \dots, x_{i,d}]^T$  where the notation  $h_{i,k} = 1$  indicates that the  $k$ th element of  $\mathbf{x}_i$ ,  $x_{i,k}$ , is observable and the notation  $h_{i,k} = 0$  indicates that the  $k$ th element of  $\mathbf{x}_i$  is not observable. When  $h_{i,k} = 0$ ,  $x_{i,k}$  is typically set equal to a constant such as zero (Allison 2001; Groenwold et al. 2012). This method has been called the *missing-indicator method* (Groenwold et al. 2012) and has also been called the *dummy variable adjustment method* (Allison 2001). It provides a useful method for identifying the presence or absence of all the information in the data set and thus is applicable to representing MCAR, MAR, and MNAR missing-data mechanisms. In practice, the researcher often does not have specific knowledge for modeling the joint distribution of the complete-data representation and the missing data indicator variables, which further increases the likelihood of model misspecification. Groenwold et al. (2012) provides some explicit empirical examples illustrating the challenge of correctly applying this approach.

**Overview of Parameter Estimation in the Presence of Partially Observable Data.** If the missing-data mechanism is MCAR, then maximum likelihood estimation can be utilized by first applying *listwise deletion* (Allison 2001; King et al. 2001) also known as *complete-case analysis* (Little and Rubin 2002), which involves simply removing data records (observations) containing missing values from the data set (also see Groenwold et al. 2012). The resulting dataset, containing no missing values, is then used for statistical modeling. A problem with using listwise deletion to handle a MCAR missing-data mechanism is that the standard errors of the parameter estimates for the researcher's observable-data model may be larger because the information contained in records with missing values has been removed from the data set. However, a more serious issue with the listwise deletion method is that for MAR data the maximum likelihood estimates may be biased (e.g., Allison 2001; Ibrahim et al. 2005; King et al. 2001).

If the missing DGP has a MAR mechanism, then maximum likelihood estimates can be obtained by using Expectation Maximization (EM) (Allison 2001; Dempster et al. 1977; Little and Rubin 2002; McLachlan and Krishnan 1997) or multiple imputation (MI) (Efron 1994; Groenwold et al. 2012; Ibrahim et al. 2005; Robins and Wang 2000; Rubin 1996, 1987; Wang and Robins 1998). These algorithms allow for estimation and inference in the presence of missing data while only requiring that the researcher specify the complete-data model. In these situations, it can be shown that the MAR missing-data mechanism is safely "ignorable" in the sense that all of the available information in the data set is used for the purposes of computing unbiased maximum likelihood estimates without the possibility of inflated standard errors (e.g., Little and Rubin 2002, p. 19) that listwise deletion may cause.

In many important cases, the researcher may have a specific theory of how the complete-data was generated, but does not have a strong theory of how the complete-data was decimated by the missing-data mechanism. In such situations, it is not uncommon when developing models on incomplete data for researchers to assume an ignorable missing-data mechanism (MCAR, MAR). Further, a MNAR missing-data mechanism is not completely verifiable using only incomplete data

and so cannot be detected without additional assumptions. In addition, [Molenberghs et al. \(2008\)](#) showed that an MNAR model may be replaced with a MAR model that fits the observed data exactly. Further, while statistical tests exist for checking the MCAR mechanism on the data set (e.g., [Little 1988](#); see [Rhoads 2012](#) for a review), tests for the MAR assumption against the MNAR alternative require additional assumptions to be refutable ([Breunig 2019](#); [Jaeger 2006](#); [Rhoads 2012](#)). Nonetheless, in practice, much statistical modeling still relies on the assumption of an ignorable missing-data mechanism and thus robust methods that offer improved estimation and inference approaches to deal with models of ignorable missing-data mechanisms are of critical importance. Our theoretical framework provides the foundation for utilizing new model misspecification testing methods ([Golden et al. 2013, 2016](#); [Henley et al. 2019](#)) to improve statistical modeling on incomplete data in the presence of both ignorable and nonignorable missing data processes.

### 1.2. Prior Work on Misspecification in Missing Data Models

Although the consequences of model misspecification have been investigated for many years (e.g., [White 1982, 1994](#); [Fomby and Hill 2003](#); [Chen and Swanson 2013](#); [Golden et al. 2013, 2016](#)), a detailed investigation into the consequences of model misspecification for statistical models in the presence of missing data addressed by this article continues to be an open area of research. It is important to emphasize that when missing data is present, one must not only consider the possibility of misspecification in the complete-data model, but also the possibility of misspecification of the missing-data mechanism. For example, the complete-data model may be correctly specified, but the assumption that the missing-data mechanism is ignorable may be incorrect.

An important *robust* method for characterizing the asymptotic distribution of the QMLEs in the presence of model misspecification is the sandwich covariance matrix estimator (e.g., [Huber 1967](#); [White 1982, 1994](#)). For example, [Arminger and Sobel \(1990\)](#) used the sandwich covariance matrix estimator for the purpose of characterizing the asymptotic distribution of the QMLEs for linear structural equation models in the presence of missing data. [Robins and Wang \(2000, pp. 114–115\)](#) and [Sung and Geyer \(2007, pp. 991–92\)](#) also used the sandwich covariance matrix estimator as the basis for their analysis of the asymptotic behavior of multiple imputation estimation. [Yuan \(2009, pp. 1901–2\)](#) discusses the relation of the sandwich covariance matrix estimator with respect to missing data problems and model misspecification for Gaussian models. [Kashner et al. \(2010\)](#) used a misspecification-robust difference-in-differences binary logistic regression model with the sandwich covariance matrix estimator applied to “naturally” missing observational data from before–after study designs.

### 1.3. A Framework for Understanding Misspecification in Missing Data Models

This article provides a formal framework for characterizing models of missing data with ignorable missing-data mechanisms when either the complete-data model is misspecified or the missing-data mechanism is misspecified. First, we provide sufficient regularity conditions on the researcher’s complete-data model to characterize the asymptotic behavior of maximum likelihood estimates in the simultaneous presence of both missing data and model misspecification. These results are then used to derive robust hypothesis testing methods for possibly misspecified models in the presence of ignorable or nonignorable missing-data mechanisms. Second, a method for the detection of model misspecification in missing data problems is discussed using recently developed Generalized Information Matrix Tests (GIMT) ([Golden et al. 2013, 2016](#); also see [Cho and White 2014](#); [Cho and Phillips 2018](#); [Huang and Prokhorov 2014](#); [Ibragimov and Prokhorov 2017](#); [Prokhorov et al. 2019](#); [Schepsmeier 2015, 2016](#); [Zhu 2017](#)). Third, we provide regularity conditions for the Missing Information Principle (MIP) to hold in the presence of model misspecification in order to provide useful computational covariance matrix estimation formulas. Fourth, we provide regularity conditions that ensure the missing data expected negative log-likelihood function for models with ignorable missing-data mechanisms is convex on the parameter space, if and only if, the fraction of information loss function on the parameter space does not exceed one. Fifth, we provide regularity

conditions that ensure that when the researcher has: (i) correctly specified a probability model for partially observable data as a complete-data model with an ignorable missing-data mechanism, and (ii) the missing data expected negative log-likelihood is convex on the parameter space, then a strict local minimizer of the missing data expected negative log-likelihood is the unique true parameter value for the complete-data model.

To our knowledge, explicit regularity conditions for the new theorems presented here are not readily available in the published scientific literature. Further, methods for testing for model misspecification in the presence of missing values with the GIMT method have not been discussed in the published scientific literature. In the final section of this article, the key results of the stated theorems and their relevance for practical data analysis problems with the use of new model misspecification tests are presented. Sketches of the key proofs which are based upon conventional arguments may be found in the Appendix A.

## 2. Assumptions

In this section, we provide the assumptions of a formal framework for characterizing models of missing data with assumed ignorable missing-data mechanisms when either the complete-data model or the missing-data mechanism are possibly misspecified.

### 2.1. Data Generating Process Assumptions

**Assumption 1. I.I.D. Partially Observable Data Generating Process.** Let  $(\mathbf{X}_i, \mathbf{H}_i)$ ,  $i = 1, 2, \dots$  be a sequence of independent and identically distributed (i.i.d.) random vectors where  $(\mathbf{X}_i, \mathbf{H}_i)$  has a common Radon–Nikodým probability density  $p_{x,h} : \mathbb{R}^d \times \{0, 1\}^d \rightarrow [0, \infty)$  defined with respect to a sigma-finite measure  $\nu_{x,h}$ .

In regression modeling applications, the first element of the  $d$ -dimensional real vector  $\mathbf{x}_i$  (a realization of  $\mathbf{X}_i$ ) is a value of the outcome variable for a regression model associated with the  $i$ th data record while the remaining elements of  $\mathbf{x}_i$  are values for the predictor variables associated with the  $i$ th data record,  $i = 1, \dots, n$ . The  $i$ th observed data indicator record  $\mathbf{h}_i$  (a realization of  $\mathbf{H}_i$ ) is a  $d$ -dimensional binary vector defined so that its  $j$ th element is 1 if the  $j$ th element of  $\mathbf{x}_i$  is observable and the  $j$ th element of  $\mathbf{h}_i$  is 0 otherwise,  $i = 1, \dots, n$ . Let  $\mathbf{x}^n \equiv [\mathbf{x}_1, \dots, \mathbf{x}_n]$ ,  $\mathbf{x}^n \in \mathbb{R}^{dn}$ . Let  $\mathbf{h}^n \equiv [\mathbf{h}_1, \dots, \mathbf{h}_n]$ ,  $\mathbf{h}^n \in \mathbb{L}^{dn}$ . Given the full partially observable record  $(\mathbf{x}, \mathbf{h})$ , let the number of observable elements of  $\mathbf{x}$  be defined such that  $\rho(\mathbf{h}) \equiv (\mathbf{1}_d)^T \mathbf{h}$ ,  $\forall \mathbf{h} \in \mathbb{L}^d$  where  $\rho : \mathbb{L}^d \rightarrow \{0, 1, 2, \dots, d\}$  and the notation  $\mathbf{1}_d$  is used to denote a  $d$ -dimensional column vector of ones. For convenience, let  $\rho_{\mathbf{h}} \equiv \rho(\mathbf{h})$ . Also define the observable-data selection matrix  $\mathbf{s}(\mathbf{h})$  generated by  $\mathbf{h}$  as a matrix with  $\rho_{\mathbf{h}}$  rows and  $d$  columns such that the  $k$ th element of the  $j$ th row of  $\mathbf{s}(\mathbf{h})$  is equal to 1 if the  $j$ th non-zero element in  $\mathbf{h}$  is the  $k$ th element in  $\mathbf{h}$ ; and set the  $jk$ th element of  $\mathbf{s}(\mathbf{h})$  equal to 0 otherwise. Let  $\mathbf{y}_{\mathbf{h}} : \mathbb{R}^d \rightarrow \mathbb{R}^{\rho_{\mathbf{h}}}$  be defined such that:  $\mathbf{y}_{\mathbf{h}}(\mathbf{x}) = \mathbf{s}(\mathbf{h})\mathbf{x}$  for  $\mathbf{h} \in \mathbb{L}^d \setminus \mathbf{0}_d$ . The  $i$ th observable-data record component  $\mathbf{y}_i \equiv \mathbf{y}_{\mathbf{h}_i}(\mathbf{x}_i)$  is thus a  $\rho_{\mathbf{h}_i}$ -dimensional column vector generated from the realization of the full partially observable record  $(\mathbf{X}_i, \mathbf{H}_i)$ , for  $i = 1, \dots, n$ . Let  $\mathbf{Y}_i \equiv \mathbf{y}_{\mathbf{H}_i}(\mathbf{X}_i)$ ,  $i = 1, \dots, n$ . The sequence of random variables  $\mathbf{Y}_1, \dots, \mathbf{Y}_n$  is i.i.d. because  $(\mathbf{X}_1, \mathbf{H}_1), \dots, (\mathbf{X}_n, \mathbf{H}_n)$  are i.i.d. distributed by Assumption 1. Let  $\mathbf{y}^n \equiv \{\mathbf{y}_1, \dots, \mathbf{y}_n\}$  and  $\mathbf{h}^n \equiv \{\mathbf{h}_1, \dots, \mathbf{h}_n\}$  be realizations of  $\mathbf{Y}^n \equiv \{\mathbf{Y}_1, \dots, \mathbf{Y}_n\}$  and  $\mathbf{H}^n \equiv \{\mathbf{H}_1, \dots, \mathbf{H}_n\}$  respectively. Let  $(\mathbf{y}^n, \mathbf{h}^n)$  denote the observed data sample.

Now define the unobservable-data selection matrix  $\bar{\mathbf{s}}(\mathbf{h})$  generated by  $\mathbf{h}$  as a matrix with  $d - \rho_{\mathbf{h}}$  rows and  $d$  columns such that the  $k$ th element of the  $j$ th row of  $\bar{\mathbf{s}}(\mathbf{h})$  is equal to 1 if the  $j$ th zero element in  $\mathbf{h}$  is the  $k$ th element in  $\mathbf{h}$ ; and set the  $jk$ th element of  $\bar{\mathbf{s}}(\mathbf{h})$  equal to 0 otherwise. Let  $\mathbf{z}_{\mathbf{h}} : \mathbb{R}^d \rightarrow \mathbb{R}^{d-\rho_{\mathbf{h}}}$  be defined such that:  $\mathbf{z}_{\mathbf{h}}(\mathbf{x}) = \bar{\mathbf{s}}(\mathbf{h})\mathbf{x}$  for  $\mathbf{h} \in \mathbb{L}^d \setminus \mathbf{1}_d$ . Thus, the  $d - \rho_{\mathbf{h}_i}$ -dimensional column vector  $\mathbf{z}_i \equiv \mathbf{z}_{\mathbf{h}_i}(\mathbf{x}_i)$  contains the unobservable components associated with the  $i$ th observed data record,  $i = 1, \dots, n$ . Let  $\mathbf{Z}_i \equiv \mathbf{z}_{\mathbf{H}_i}(\mathbf{X}_i)$ ,  $i = 1, \dots, n$ .



Finally, note that the Radon–Nikodým density representation in Assumption 1 is used so that this theory is applicable to not only situations where the complete data vector  $\mathbf{X}_i$  consists of discrete or absolutely continuous random variables, but also for situations where the complete data vector  $\mathbf{X}_i$  includes both discrete and absolutely continuous random variables. In fact, the Radon–Nikodým density  $p_{x,h}$  is also applicable to situations where the elements of  $\mathbf{X}_i$  are constructed from combinations of both discrete and absolutely continuous random variables. In the special case, where  $\mathbf{X}_i$  is a vector consisting of only discrete random variables, then  $p_{x,h}$  may be interpreted as a probability mass function.

A common representational convention in the literature (e.g., Little and Rubin 2002; Orchard and Woodbury 1972, p. 699; Rubin 1976, p. 584; Schenker and Welsh 1988, p. 1553) is to assume that a realization  $(x_i, \mathbf{h}_i)$  of an observation can be represented as  $(z_i, \mathbf{y}_i)$ , which consists of an observable component vector  $\mathbf{y}_i \equiv \mathbf{y}_{\mathbf{h}_i}(x_i)$  of dimension  $\rho_{\mathbf{h}_i}$  and unobservable component vector  $\mathbf{z}_i \equiv \mathbf{z}_{\mathbf{h}_i}(x_i)$  of dimension  $d - \rho_{\mathbf{h}_i}$  leaving dependence on  $\mathbf{h}_i$  implicit ( $i = 1, \dots, n$ ). Our Assumption 1 makes the dependence on  $\mathbf{h}_i$  more explicit and clearly shows how one can apply standard *i.i.d.* asymptotic statistical theory to support the analysis of missing data problems.

Using  $p_{x,h}$ , define  $p_x(\cdot) \equiv \int p_{x,h}(\cdot, \mathbf{h}) dv_{\mathbf{h}}(\mathbf{h})$  and  $p_h(\cdot) \equiv \int p_{x,h}(\mathbf{x}, \cdot) dv_{\mathbf{x}}(\mathbf{x})$ . Let the *observable-data density*  $p_{y_{\mathbf{h}}}(\mathbf{y}_{\mathbf{h}}(\mathbf{x})) \equiv \int p_x(\mathbf{x}) dv_{z_{\mathbf{h}}}(\mathbf{z}_{\mathbf{h}}(\mathbf{x}))$ , which can be rewritten using a more implicit compact notation as  $p_{y_{\mathbf{h}}}(\mathbf{y}_{\mathbf{h}}) \equiv \int p_x(\mathbf{x}) dv_{z_{\mathbf{h}}}(\mathbf{z}_{\mathbf{h}})$ . The density  $p_{y_{\mathbf{h}}} : \mathbb{R}^{\rho_{\mathbf{h}}} \rightarrow [0, \infty)$  specifies the conditional probability distribution of the random vector  $\mathbf{y}_{\mathbf{h}}(\mathbf{X}) = \mathbf{s}(\mathbf{h})\mathbf{X}$  given a particular observed data indicator record  $\mathbf{h}$ . The *observable-data density*  $p_{y_{\mathbf{h}},h}(\mathbf{y}_{\mathbf{h}}(\mathbf{x}), \mathbf{h}) \equiv \int p_{x,h}(\mathbf{x}, \mathbf{h}) dv_{z_{\mathbf{h}}}(\mathbf{z}_{\mathbf{h}}(\mathbf{x}))$  specifies the joint probability distribution of the observed data record  $(\mathbf{Y}, \mathbf{H})$  that includes the pattern of missingness  $\mathbf{H}$  as well as the observable data component  $\mathbf{Y}$ .

Additionally, define the *missing-data mechanism*: density  $p_{h|x} \equiv p_{x,h}/p_x$ . The missingness-data mechanism is called a *MAR missing-data mechanism* if there exists a function  $p_{h|y_{\mathbf{h}}} : \mathbb{L}^d \times \mathbb{R}^{\rho_{\mathbf{h}}} \rightarrow [0, \infty)$  such that  $p_{h|x}(\mathbf{h}|\mathbf{x}) = p_{h|y_{\mathbf{h}}}(\mathbf{h}|\mathbf{y}_{\mathbf{h}}(\mathbf{x}))$  holds for all  $\mathbf{h} \in \mathbb{L}^d$  and for all  $\mathbf{x} \in \mathbb{R}^d$ . A *MAR missing-data mechanism*  $p_{h|x}$  is called an *MCAR missing-data mechanism* if  $p_{h|x}(\mathbf{h}|\mathbf{x}) = p_h(\mathbf{h})$  holds for all  $\mathbf{x} \in \mathbb{R}^d$  and for all  $\mathbf{h} \in \mathbb{L}^d$ . A conditional density  $p_{h|x}$  that is not a *MAR missing-data mechanism* is called a *MNAR missing-data mechanism*. These definitions are consistent with the discussion in Little and Rubin (2002, pp. 11–12; also see Rubin 1976), but are formulated for the specific *i.i.d.* case considered here.

## 2.2. Probability Model Assumptions

Let *supp*  $\mathbf{X}$  denote the support of  $\mathbf{X}$ .

**Assumption 2. Parametric Densities.** (i) Let  $\Theta$  be a compact and non-empty subset of  $\mathbb{R}^r$ ,  $r \in \mathbb{N}$ . (ii) Let  $f : \mathbb{R}^d \times \Theta \rightarrow [0, \infty)$ . For each  $\theta$  in  $\Theta$ ,  $f(\cdot; \theta)$  is a density with respect to  $v_x$  and, for each  $\mathbf{x} \in \text{supp } \mathbf{X}$ ,  $f(\mathbf{x}; \cdot)$  is continuous on  $\Theta$ . (iii)  $f(\mathbf{x}; \cdot)$  is continuously differentiable on  $\Theta$  for each  $\mathbf{x} \in \text{supp } \mathbf{X}$ . (iv)  $f(\mathbf{x}; \cdot)$  is twice continuously differentiable on  $\Theta$  for each  $\mathbf{x} \in \text{supp } \mathbf{X}$ .

The approximating complete-data density  $f(\cdot; \theta)$  specifies the likelihood of a data record in the case where no component of the data record is missing for each  $\theta$  in the parameter space  $\Theta$ . A set of complete-data densities indexed by the parameter vector  $\theta$  specifies the researcher’s complete data model:  $M_c \equiv \{f(\mathbf{x}; \theta) : \theta \in \Theta\}$ .

**Assumption 3. Ignorable Missing-Data Mechanism.** Let  $q_{h|x} : \mathbb{L}^d \times \mathbb{R}^d \rightarrow [0, \infty)$  be a measurable function. (i) For each  $\mathbf{x} \in \text{supp } \mathbf{X}$ ,  $q_{h|x}(\cdot|\mathbf{x})$  is a density with respect to  $v_h$ . (ii)  $q_{h|x}$  is *MAR*.

Note that the researcher’s approximation to the missing-data mechanism (specified by the density  $q_{h|x}$ ) is *MAR* so that  $q_{h|x}$  is not functionally dependent upon the unobservable-data record component  $\mathbf{z}_{\mathbf{h}}(\mathbf{x})$ . Furthermore,  $q_{h|x}$  is a constant on the parameter space  $\Theta$ . These are the two conditions that define an *ignorable missing-data mechanism* (e.g., Little and Rubin 2002, p. 119; also see Heitjan 1994; Kenward and Molenberghs 1998; Nielsen 1997; Rubin 1976). When  $q_{h|x}$  is *MAR*, it will be convenient to define the function  $q_{h|y_{\mathbf{h}}}$  such that  $q_{h|x}(\mathbf{h}|\mathbf{x}) = q_{h|y_{\mathbf{h}}}(\mathbf{h}|\mathbf{y}_{\mathbf{h}}(\mathbf{x}))$  for all  $(\mathbf{x}, \mathbf{h}) \in \mathbb{R}^d \times \mathbb{L}^d$ .

In contrast to Assumption 3, a non-ignorable missing-data mechanism corresponds to a situation where either: (i)  $q_{h|x}$  is functionally dependent upon the unobservable-data record component  $\mathbf{z}_h(\mathbf{x})$ , or (ii)  $q_{h|x}$  is not a constant on the complete-data model parameter space  $\Theta$ . Every model of an MNAR missing-data mechanism is non-ignorable. However, it is possible for a researcher to postulate a non-ignorable MAR or MCAR missing-data mechanism.

In order to specify a *missing-data probability model*, the researcher constructs a best approximating model of the DGP density  $p_{x,h}$  using the approximating complete-data density  $f(\cdot; \theta)$  for each  $\theta$  in  $\Theta$  together with the approximating missing-data mechanism  $q_{h|x}$ . In many practical missing data applications, it is common practice to only implicitly specify the missing-data probability model since the researcher explicitly provides only the complete-data density  $f(\cdot; \theta)$  and implicitly assumes an ignorable missing-data mechanism.

Let  $q_{x,h} \equiv f q_{h|x}$  specify the researcher's best approximating density of the DGP density  $p_{x,h}$ . Let  $H \subseteq \{0, 1\}^d$  denote the set of all permissible missing data patterns (i.e., the support of  $H_i$ ). Let  $q_h(\mathbf{h}; \theta) \equiv \int q_{x,h}(\mathbf{x}, \mathbf{h}; \theta) d\nu_x(\mathbf{x})$  and  $q_{y_h}(\mathbf{y}_h(\mathbf{x}); \theta) \equiv \int f(\mathbf{x}; \theta) d\nu_{z_h}(\mathbf{z}_h)$  for all  $\mathbf{h} \in H$ . The density  $q_{y_h}$  is intended to approximate the observable-data density  $p_{y_h}$ . The researcher's *observable-data model* of the data generating process is the set  $M_o \equiv \{q_{y_h}(\mathbf{y}_h(\mathbf{x}); \theta) : \theta \in \Theta, \mathbf{h} \in H\}$  for a given parameter space  $\Theta \subseteq \mathbb{R}^r$ .

The density  $q_{z_h|y_h,h} = (q_{h|x}/q_{h|y_h})(f/q_{y_h})$  specifies the researcher's model of the distribution of the missing data given the observable data when  $\mathbf{h} \in L^d \setminus \{0_d, 1_d\}$ . Let  $q_{z_h|y_h,h} \equiv 1$  for  $\mathbf{h} = 1_d$  (fully observable case where  $q_{h|x} = q_{h|y_h}$  and  $f = q_{y_h}$ ) and  $q_{z_h|y_h,h} \equiv f$  for  $\mathbf{h} = 0_d$  (fully unobservable case where  $q_{h|x} = q_{h|y_h}$  and  $q_{y_h} = 1$ ).

**Definition 1. Misspecified Model.** (i) The complete-data model  $M_c$  is called a *correctly specified complete-data model* if the complete-data DGP density  $p_x \in M_c$  holds  $v_x$ -a.e.; otherwise  $M_c$  is *misspecified complete data model*. (ii) The observable-data model  $M_o$  is called a *correctly specified observable-data model* if the observable-data DGP density  $p_{y_h} \in M_o$  holds  $v_{y_h}$ -a.e.; for all  $\mathbf{h} \in H$ ; otherwise  $M_o$  is a *misspecified observable data model*.

If  $p_x \in M_c$ , then this is a sufficient, but not necessary condition for ensuring  $M_c$  is correctly specified. In particular, the statement  $p_x \in M_c$  holds  $v_x$ -a.e. is a formal way of acknowledging that it is possible for another density  $\check{p}$  to have the property that its corresponding cumulative distribution function is exactly the same as the cumulative distribution function for the DGP density  $p_x$  even though  $\check{p}(\mathbf{x}) = p_x(\mathbf{x})$  only for all  $\mathbf{x}$  where the sigma-finite measure  $v_x$  vanishes. For example, let  $p_x(x) = 1$  for all  $|x| \leq 1$  with  $p_x(x) = 0$  for all  $|x| > 1$ . Let  $\check{p}(x) = 1$  for all  $|x| < 1$  with  $p_x(x) = 0$  for all  $|x| \geq 1$ . The cumulative distribution functions for  $p_x$  and  $\check{p}$  are identical even though  $p_x \neq \check{p}$ .

A missing-data probability model may be misspecified if either: (i) the complete-data model  $M_c$  is misspecified, (ii) the missing-data mechanism  $q_{h|x}$  is misspecified, or (iii) both the complete-data model  $M_c$  and the missing-data mechanism  $q_{h|x}$  are misspecified.

In regression modeling, a complete-data record  $\mathbf{x}$  is commonly partitioned such that  $\mathbf{x} = [R, \mathbf{u}]$  where  $R$  is the regression model response variable and  $\mathbf{u}$  is the predictor variables for the regression model. The complete-data probability model is specified by  $f$ . Typically,  $f(\mathbf{x}; \theta)$  is factored such that:  $f(\mathbf{x}; \theta) = f_{R|u}(R|\mathbf{u}; \theta_{R|u}) f_u(\mathbf{u}; \theta_u)$  where  $\theta = [\theta_{R|u}, \theta_u] \in \Theta_{R|u} \times \Theta_u$ . Thus, misspecification of the researcher's complete-data probability model in a regression modeling application may be due to either a misspecification of either (or both) the regression model and the conditional missing predictor variable model. In practice, the researcher's *conditional missing predictor model* is specified by densities of the form  $f_{u_{miss}|u_{obs}} \equiv f_u / f_{u_{obs}}$  where  $f_{u_{obs}}$  is the marginal distribution for the predictors that are fully observable according to the researcher's missing-data probability model. Additional discussion of conditional missing predictor models may be found in [Chen \(2004\)](#) and [Ibrahim et al. \(1999\)](#).

### 2.3. Likelihood Functions, Pseudo-True Parameter Values, and True Parameter Values

Note that the notation  $\log(q)$  will be used throughout this article to refer to the natural log of  $q$ .

**Definition 2. Complete-Data Likelihood Function.** Assume Assumptions 1, 2(i), and 2(ii) hold. Given a data sample  $\mathbf{x}^n$ , the complete-data likelihood function  $L_n^x : \Theta \times \mathbb{R}^{dn} \rightarrow [0, \infty)$  is defined such that:  $L_n^x(\theta; \mathbf{x}^n) = \prod_{i=1}^n f(\mathbf{x}_i; \theta)$  for all  $\theta \in \Theta$ . The complete-data negative average log-likelihood  $\bar{l}_n^x : \Theta \times \mathbb{R}^{dn} \rightarrow [0, \infty)$  is defined such that:  $\bar{l}_n^x(\theta; \mathbf{x}^n) = -n^{-1} \log L_n^x(\theta; \mathbf{x}^n)$  for all  $\theta \in \Theta$ . The complete-data expected negative average log-likelihood  $l^x : \Theta \rightarrow [0, \infty)$  is defined (when it exists) such that:  $l^x(\theta) = -\int p_x(\mathbf{x}) \log(f(\mathbf{x}; \theta)) dv_x(\mathbf{x})$ . The complete-data Kullback–Leibler Information Criterion (KLIC)  $\ddot{l}_n^x : \Theta \rightarrow [0, \infty)$  is defined (when it exists) such that:  $\ddot{l}_n^x(\theta) = l^x(\theta) + \int p_x(\mathbf{x}) \log(p_x(\mathbf{x})) dv_x(\mathbf{x})$ .

The complete-data likelihood function (e.g., White 1982, 1994; McCullagh and Nelder 1989; Dobson 2002; Little and Rubin 2002) is the usual likelihood function encountered in problems where no missing data is present. In many situations (e.g., when the complete-data probability model contains members of the linear exponential family), the complete-data expected negative log-likelihood  $l^x(\cdot) : \Theta \rightarrow [0, \infty)$  is convex on the parameter space  $\Theta$  (e.g., Kass and Voss 1997, pp. 14–19). In such situations, a strict local minimizer of  $l^x$  is the unique global minimizer on the parameter space. For more complicated probability models where  $l^x$  contains multiple strict local and global minimizers, the parameter space  $\Theta$  can sometimes be defined to contain exactly the strict local minimizer of  $l^x$ .

**Definition 3. Complete-Data True Parameter Value.** Assume that Assumptions 1, 2(i), and 2(ii) hold. A global minimizer of the complete-data negative average likelihood function  $\bar{l}_n^x(\cdot; \mathbf{X}^n) : \Theta \rightarrow [0, \infty)$  on the parameter space  $\Theta$  is called a complete-data quasi-maximum likelihood estimator. A global minimizer of the complete-data expected negative log-likelihood  $l^x(\cdot) : \Theta \rightarrow [0, \infty)$  is called a complete-data pseudo-true parameter value  $\theta_x^*$ . A parameter value  $\theta_0 \in \Theta$  defined such that for all  $\mathbf{x} \in \text{supp } \mathbf{X}$ :  $f(\mathbf{x}; \theta_0) = p_x(\mathbf{x})$  is called a complete-data true parameter value.

Note that when misspecification is present, it is possible that the complete-data true parameter value may not exist because the complete-data model is not capable of representing the complete-data DGP.

For a missing-data probability model, the missing-data likelihood function is:

$$L_n^{y,h}(\theta; \mathbf{y}^n, \mathbf{h}^n) \equiv \prod_{i=1}^n \left[ \int q_{h|x}(\mathbf{h}_i | \mathbf{x}_i) f(\mathbf{x}_i; \theta) dv_{z_{h_i}}(\mathbf{z}_{h_i}(\mathbf{x}_i)) \right] \tag{1}$$

when the researcher’s missing-data mechanism model specified by  $q_{h|x}(\mathbf{h} | \mathbf{x})$  is an ignorable missing-data mechanism model (i.e., see Assumption 3) so that  $q_{h|x}(\mathbf{h} | \mathbf{x}) = q_{h|y_h}(\mathbf{h} | \mathbf{y}_h(\mathbf{x}))$ , then (1) may be rewritten as:

$$L_n^{y,h}(\theta; \mathbf{y}^n, \mathbf{h}^n) \equiv \prod_{i=1}^n q_{h|y_{h_i}}(\mathbf{h}_i | \mathbf{y}_{h_i}(\mathbf{x}_i)) \prod_{i=1}^n \int f(\mathbf{x}_i; \theta) dv_{z_{h_i}}(\mathbf{z}_{h_i}(\mathbf{x}_i))$$

or equivalently as:

$$L_n^{y,h}(\theta; \mathbf{y}^n, \mathbf{h}^n) \equiv \prod_{i=1}^n q_{h|y_{h_i}}(\mathbf{h}_i | \mathbf{y}_i) \prod_{i=1}^n q_{y_{h_i}}(\mathbf{y}_i; \theta). \tag{2}$$

The likelihood  $L_n^{y,h}$  in (2) shows that when the researcher assumes an ignorable missing-data mechanism model, this implies that the global minimizers of  $L_n^{y,h}$  are only functionally dependent upon the observable data components  $\mathbf{y}^n \equiv [\mathbf{y}_1, \dots, \mathbf{y}_n]$  generated from combining the complete-data patterns  $\mathbf{x}^n \equiv [\mathbf{x}_1, \dots, \mathbf{x}_n]$  and missing-data patterns  $\mathbf{h}^n \equiv [\mathbf{h}_1, \dots, \mathbf{h}_n]$ . The global minimizers of  $L_n^{y,h}$  are not functionally dependent on either the researcher’s choice of missing-data mechanism  $q_{h|y_h}$  or an explicit representation of the specific patterns of missingness  $\mathbf{h}^n \equiv [\mathbf{h}_1, \dots, \mathbf{h}_n]$  (Rubin 1976; Schafer 1997, pp. 11–12; Little and Rubin 2002, p. 119). Furthermore, note that although the researcher has assumed an ignorable missing-data mechanism model for the likelihood  $L_n^{y,h}$  in (2), this assumption does not imply that the data generating process is actually MAR because the researcher’s assumption



of an ignorable missing-data mechanism model may be wrong. Thus, maximizing the likelihood  $L_n^{y,h}$  in (2) to estimate the parameter values  $\theta$  when the data generating process is MNAR is quasi-maximum likelihood estimation (e.g., White 1982, 1994). These remarks thus motivate the following definition of the observable-data likelihood function (e.g., Schafer 1997, pp. 11–12; Little and Rubin 2002, p. 119) that is central to the objectives of this article.

**Definition 4. Observable-Data Likelihood Function.** Assume that Assumptions 1, 2(i), 2(ii), and 3 hold. Let  $Y^n \equiv \prod_{i=1}^n \mathbb{R}^{p_{h_i}}$ . Given an observable data sample  $(\mathbf{y}^n, \mathbf{h}^n)$ , the observable-data likelihood function  $L_n^y : \Theta \times Y^n \times \mathbb{R}^{d_n} \rightarrow [0, \infty)$  is defined such that  $L_n^y(\theta; \mathbf{y}^n, \mathbf{h}^n) = \prod_{i=1}^n q_{y_{h_i}}(\mathbf{y}_i; \theta)$  where  $q_{y_h}(\mathbf{y}_h(\mathbf{x}); \theta) = \int f(\mathbf{x}; \theta) dv_{z_h}(\mathbf{z}_h(\mathbf{x}))$  for all  $\theta \in \Theta$ . The observable-data negative average log-likelihood  $\bar{l}_n : \Theta \times Y^n \times \mathbb{R}^{d_n} \rightarrow [0, \infty)$  is defined such that:

$$\bar{l}_n(\theta; \mathbf{y}^n, \mathbf{h}^n) = -n^{-1} \sum_{i=1}^n \log(q_{y_{h_i}}(\mathbf{y}_i; \theta)) \quad (3)$$

for all  $\theta \in \Theta$ . The observable-data expected negative average log-likelihood  $l : \Theta \rightarrow [0, \infty)$  is defined (when it exists) by:

$$l(\theta) = - \int p_{y_h, h}(\mathbf{y}_h, \mathbf{h}) \log(q_{y_h}(\mathbf{y}_h; \theta)) dv_{y_h, h}(\mathbf{y}_h, \mathbf{h}). \quad (4)$$

The observable-data negative average log likelihood  $\bar{l}_n$  in general is typically not convex on the parameter space  $\Theta$  even if the complete-data probability model specified by  $f(\mathbf{x}; \theta)$  is a member of the linear exponential family (Orchard and Woodbury 1972; Louis 1982). The non-convex property of  $\bar{l}_n$  is a consequence of the Missing Information Principle (see Louis 1982, and Theorem 3), which states that the Hessian of the observable-data negative average log-likelihood is the difference of two positive semidefinite matrices. McLachlan and Krishnan (1997, pp. 91–95; also see Schafer 1997, pp. 52–55; Murray 1977) provide some helpful empirical examples of non-convex missing-data likelihood functions. Thus, in the presence of missing data, it is not unusual for multiple local minimizers, ridges, or multiple global minimizers of the negative average observable data log-likelihood function to exist. To apply the asymptotic theory in such situations, it may be possible to choose the parameter space  $\Theta$  such that  $l : \Theta \rightarrow [0, \infty)$  is a convex function on  $\Theta$  and has a unique global minimizer on  $\Theta$ .

**Definition 5. Observable-Data True Parameter Value.** Assume that Assumptions 1, 2(i), and 2(ii) hold. A global minimizer of the observable-data negative average likelihood function  $\bar{l}_n$  on the parameter space  $\Theta$  is called an observable-data quasi-maximum likelihood estimator. A global minimizer of the observable-data expected negative log-likelihood  $l$  is called an observable-data pseudo-true parameter value  $\theta_0^*$ . A parameter value  $\theta_0^* \in \Theta$  defined such that for all  $\mathbf{y}_h \in \text{supp } \mathbf{Y}_h$ :  $q_{y_h}(\mathbf{y}_h; \theta_0^*) = p_{y_h}(\mathbf{y}_h)$  for each  $\mathbf{h} \in \mathbf{H}$  is called an observable-data true parameter value.

The observable-data quasi-maximum likelihood estimator is the parameter value that maximizes the likelihood of the observable data in terms of the assumptions associated with the researcher's proposed probability model. In addition (see Equations (1) and (2)), when the DGP missing-data mechanism is MAR the observable-data pseudo-true parameter value is semantically interpretable as identifying the probability distribution in the researcher's observable data probability model that is most similar to the probability distribution that generated the observed data (i.e., the distribution specified by the density  $p_{y_h, h}$ ) using the Kullback–Leibler Information Criterion (e.g., White 1982, 1994; Kullback and Leibler 1951). If the DGP missing-data mechanism is MAR, the complete-data probability model is correctly specified, and the observable-data expected negative log-likelihood is convex, then

the observable-data true parameter value and the complete-data true parameter value are identical (see Proposition 3(iii) of this article).

Since the researcher is assuming an ignorable missing-data mechanism, this means that the observable-data pseudo-true parameter value is calculated without incorporating knowledge of the observed patterns of missingness and without incorporating an explicit missing-data mechanism into the researcher's probability model. These assumptions correspond to potentially serious misspecification errors when the DGP missing-data mechanism is MNAR.

#### 2.4. Moment Assumptions

Let  $\nabla$  denote the gradient operator with respect to  $\theta \in \Theta$  yielding an  $r$ -dimensional column vector. Let  $\nabla^2$  denote the Hessian operator with respect to  $\theta \in \Theta$  yielding an  $r \times r$  symmetric matrix. Let the norm,  $\|\cdot\|$ , of an  $r$ -dimensional vector  $\theta$  be defined such that:  $\|\theta\|^2 \equiv \theta^T \theta$ . Let  $\phi_{y_h} : \mathbb{R}^{l_h} \times \Theta \rightarrow [0, \infty)$  be defined such that for all  $\mathbf{y} \in \text{supp } \mathbf{y}_h(\mathbf{X})$  and for all  $\theta \in \Theta$ :  $\phi_{y_h}(\mathbf{y}; \theta) = q_{y_h}(\mathbf{y}; \theta) / p_{y_h}(\mathbf{y}; \theta)$ ,  $\mathbf{h} \in \mathbb{L}^d$ . Let  $\phi_x : \mathbb{R}^d \times \Theta \rightarrow [0, \infty)$  be defined such that:  $\phi_x(\mathbf{x}; \theta) = f(\mathbf{x}; \theta) / p_x(\mathbf{x})$  for all  $\mathbf{x} \in \text{supp } X$  and for all  $\theta \in \Theta$ .

**Assumption 4. Domination Conditions.** For each  $\mathbf{h} \in \mathbf{H} \cup \{1_d\}$ :

- (i) (a)  $\log q_{y_h}$  is dominated on  $\Theta$  with respect to  $p_{y_h}$ ;
- (b) each element of  $\nabla \log q_{y_h}$  is dominated on  $\Theta$  with respect to  $p_{y_h}$ ;
- (c)  $\|\nabla \log q_{y_h}\|^2$  is dominated on  $\Theta$  with respect to  $p_{y_h}$ ;
- (d) each element of  $\nabla^2 \log q_{y_h}$  is dominated on  $\Theta$  with respect to  $p_{y_h}$ ; and
- (ii) there exists a finite positive number  $K$  such that for all  $\mathbf{x} \in \text{supp } X$  and for all  $\theta \in \Theta$ :  $f(\mathbf{x}; \theta) \leq K p_x(\mathbf{x})$ .

Assumption 4 holds under fairly general conditions. Assumption 4(i) corresponds to standard maximum likelihood regularity assumptions applied to the observable probability model representation (e.g., White 1982, 1994; Serfling 1980). Assumption 4(ii) is a relatively weak condition that states the likelihood of an environment events assigned by the researcher's complete data probability model must have a (generous) upper bound determined by the likelihood of that event in the environment.

Simple verifiable conditions for ensuring Assumption 4 holds are: (i) assume that the DGP is bounded (i.e.,  $|X_i| \leq K$  with probability one for some finite number  $K$ ), (ii)  $\log f(\mathbf{x}; \theta)$  is piecewise continuous in its first argument and a twice continuously differentiable function in its second argument, (iii) the parameter space  $\Theta$  is a closed and bounded set, and (iv) if for each possible realization  $\mathbf{x}_i$  in the researcher's model (i.e.,  $f(\mathbf{x}; \theta) > 0$  for all  $\theta \in \Theta$ ), the realization  $\mathbf{x}_i$  in the statistical environment should also be possible with likelihood greater than some positive number  $\varepsilon$  (i.e.,  $p_e(\mathbf{x}) > \varepsilon$ ).

#### 2.5. Solution Assumptions

**Assumption 5. Uniqueness.** (i) For some  $\theta^* \in \Theta$ ,  $l$  has a unique minimum at  $\theta^*$ . (ii)  $\theta^*$  is interior to  $\Theta$ .

The Assumption 5(i) is an identifiability assumption that is commonly made for the purposes of establishing consistency and asymptotic normality for maximum likelihood estimators in the completely observable case (e.g., White 1982). Assumption 5(i) will fail, for example, if the complete-data model contains redundant or irrelevant parameters (e.g., White 1982). However, the nature of the actual missing-data mechanism may also cause Assumption 5(i) to fail. For example, there may not be sufficient information in the observed data to uniquely specify how all predictor variables in a regression model covary.

To state our next assumption, let  $\mathbf{g}_{y_h}(y; \theta) \equiv -\nabla \log q_{y_h}(y; \theta)$  and write  $\mathbf{A}(\theta) \equiv \nabla^2 l(\theta)$  and  $\mathbf{B}(\theta) \equiv \int \mathbf{g}_{y_h}(\mathbf{y}_h; \theta) (\mathbf{g}_{y_h}(\mathbf{y}_h; \theta))^T p_{y_h, h}(\mathbf{y}_h, \mathbf{h}) d\nu_{y_h, h}(\mathbf{y}_h, \mathbf{h})$ . Let  $\mathbf{A}^* \equiv \mathbf{A}(\theta^*)$  and  $\mathbf{B}^* \equiv \mathbf{B}(\theta^*)$ .

**Assumption 6. Positive Definiteness.** (i)  $\mathbf{A}^*$  is positive definite. (ii)  $\mathbf{B}^*$  is positive definite.

Assumption 6(ii) is used in order to apply the Lindeberg–Levy Central Limit Theorem (e.g., White 1984, Theorem 5.2). Violations of Assumption 6(i) and Assumption 6(ii) may be interpreted as analogous to the presence of multicollinearity in the special case of linear regression modeling.

### 3. Theorems

In this section, the explicit regularity conditions developed in Section 2 are used to formulate and prove key theorems applicable to missing data models comprised of a possibly misspecified complete-data model and a missing-data mechanism that is possibly misspecified as ignorable. Many of the results are applicable to DGPs with either MAR or MNAR missing-data mechanisms.

Theorem 1 establishes conditions for the QMLE to converge to the pseudo-true parameter value  $\theta^*$  even if the complete-data model is misspecified. Theorem 2(i) establishes conditions for the distribution of the quasi-maximum likelihood estimates to have an asymptotic multivariate Gaussian distribution centered at  $\theta^*$  with a sandwich covariance matrix even if the complete-data model is misspecified. In addition, Theorem 2(ii) establishes that the Hessian and OPG covariance matrices will differ from each other and the sandwich covariance matrix when the observable-data model is misspecified. This latter result is important for two reasons. First, it suggests alternative methods for estimating the QMLE covariance matrix. Second, Theorem 2(ii) implies that when the Hessian and OPG covariance matrices are not equal that the complete-data model is misspecified regardless of whether or not the researcher has correctly specified the missing-data mechanism as ignorable even if the data is MNAR.

Theorem 3 provides explicit regularity conditions for the Orchard and Woodbury (1972); also see Louis (1982) Missing Information Principle to hold for ignorable missing-data mechanisms. In addition, Theorem 3 introduces several new forms of the Missing Information Principle that are important for interpreting and computing the covariance matrix of the missing-data maximum likelihood estimates in terms of the researcher's complete-data model. Proposition 2 establishes conditions for computationally tractable formulas for consistent gradient and covariance matrix estimators. Theorem 4 describes how the shape of the complete-data expected likelihood and the amount of missing information influence the convexity of the expected observable data negative log-likelihood function for both MAR and MNAR environments. Proposition 3 summarizes some key results regarding local and global identifiability of the observable-data pseudo-true parameter values, observable-data true parameter values, and complete-data true parameter values.

#### 3.1. Quasi-Maximum Likelihood Estimation for Possibly Misspecified Missing Data Models

Let the *missing-data gradient*  $\bar{\mathbf{g}}_n : \mathbb{R}^r \rightarrow \mathbb{R}^r$  be defined such that for all  $\theta \in \Theta$  and for all  $(\mathbf{y}^n, \mathbf{h}^n) \in \prod_{i=1}^n (\mathbb{R}^{p_{h_i}}) \times \mathbb{L}^{dn}$ :

$$\bar{\mathbf{g}}_n(\theta; \mathbf{y}^n, \mathbf{h}^n) \equiv n^{-1} \sum_{i=1}^n \mathbf{g}_{y_{h_i}}(\mathbf{y}_i; \theta). \quad (5)$$

**Proposition 1. Missing-Data Average Negative Log Likelihood Function and Gradient Estimation.** Assume that Assumptions 1, 2(i), 2(ii), 2(iii), 4(i)a, and 5 hold. Then as  $n \rightarrow \infty$ ,  $\bar{l}_n(\cdot; \mathbf{Y}^n, \mathbf{H}^n) \rightarrow l$  and  $\bar{\mathbf{g}}_n(\cdot; \mathbf{Y}^n, \mathbf{H}^n) \rightarrow \mathbf{g}$  uniformly on  $\Theta$  with probability one. In addition,  $l$  and  $\mathbf{g}$  are continuous on  $\Theta$ .

**Theorem 1. Estimator Consistency.** Assume that Assumptions 1, 2(i), 2(ii), 4(i)a, and 5 hold. Then as  $n \rightarrow \infty$ ,  $\hat{\theta}_n \rightarrow \theta^*$  with probability one.

Theorem 1 provides primitive conditions for ensuring the consistency of the missing-data quasi-maximum likelihood estimator  $\hat{\theta}_n$  for missing-data models involving assumed ignorable missing-data mechanisms in the possible presence of model misspecification.

### 3.2. QMLE Asymptotic Distribution for Possibly Misspecified Missing Data Models

Now consider the asymptotic distribution of  $\hat{\theta}_n$  which is a unique minimizer of  $\bar{l}_n$  on a closed and bounded parameter space  $\Theta$  that contains  $\theta^*$  in the interior of  $\Theta$ .

**Theorem 2. Asymptotic Distribution of Quasi-Maximum Likelihood Estimates.** Assume that Assumptions 1, 2, 4, 5, and 6 hold. (i) As  $n \rightarrow \infty$ ,  $\sqrt{n}(\hat{\theta}_n - \theta^*)$  converges in distribution to a zero-mean Gaussian random vector with non-singular covariance matrix  $\mathbf{C}^* \equiv (\mathbf{A}^*)^{-1}\mathbf{B}^*(\mathbf{A}^*)^{-1}$ . (ii) If, in addition, the observable-data probability model  $M_o$  is correctly specified, then  $\mathbf{A}^* = \mathbf{B}^*$ .

Theorem 2(i) (whose proof follows directly from the methods of Huber 1967; White 1982, 1994) provides explicit regularity conditions delivering the asymptotic distribution of the quasi-maximum likelihood estimator for possibly misspecified models in the presence of missing data. The covariance matrix of the quasi-maximum likelihood estimate is specified by the *missing-data robust covariance matrix* (Robins and Wang 2000; Clayton et al. 1998; Arminger and Sobel 1990):

$$\mathbf{C}^* = (\mathbf{A}^*)^{-1}\mathbf{B}^*(\mathbf{A}^*)^{-1}. \quad (6)$$

Wang and Robins (1998, p. 937) and Robins and Wang (2000, pp. 114–15) simply assume that the conclusions of Theorem 1 and Theorem 2 hold in their development of an asymptotic statistical theory of parameter estimation using both single and multiple imputation methods for possibly misspecified models with ignorable missing-data mechanisms. Thus, Theorems 1 and 2 also are useful for providing a set of primitive assumptions for the missing data asymptotic theory of imputation methods developed by Robins and Wang (2000) and Wang and Robins (1998).

Theorem 2(ii) provides explicit regularity conditions supporting the assertion that if the observable-data probability model is correctly specified (regardless of the correct specification of the researcher's missing-data mechanism), then the Information Matrix Equality ( $\mathbf{A}^* = \mathbf{B}^*$ ) holds so that  $\mathbf{C}^*$  can be replaced with either the *missing-data Hessian covariance matrix* formula (e.g., Little and Rubin 2002; Meng and Rubin 1991; Jamshidian and Jennrich 2000):

$$\mathbf{C}^* = [\mathbf{A}^*]^{-1} \quad (7)$$

or the *missing-data OPG (Outer-Product Gradient) covariance matrix* (e.g., Berndt et al. 1974; also see McLachlan and Krishnan 1997, p. 122) formula:

$$\mathbf{C}^* = [\mathbf{B}^*]^{-1}. \quad (8)$$

Note that the second part of Theorem 2, 2(ii), states that when (6), (7), and (8) are not equivalent (i.e.,  $\mathbf{C}^* \neq [\mathbf{A}^*]^{-1}$  and  $\mathbf{C}^* \neq [\mathbf{B}^*]^{-1}$ ), then the missing-data Hessian covariance matrix formula in (7) and the missing-data OPG covariance matrix formula in (8) are not correct. In this situation, the formula for the missing-data robust covariance matrix  $\mathbf{C}^*$  in (6) must be used instead of (7) and (8) in order to ensure reliable statistical inferences for possibly misspecified probability models whose parameters are estimated from missing data.

### 3.3. Validity of Missing Information Principles When Model Misspecification Is Present

The Uniform Law of Large Numbers (e.g., Jennrich 1969, Theorem 2) with Assumptions 1, 2, and 4, provides a variety of convenient ways to estimate  $\mathbf{A}^*$ ,  $\mathbf{B}^*$ , and  $\mathbf{C}^*$ . Let  $\hat{\mathbf{g}}_{y_h} = -\int (\nabla \log f(\mathbf{x}; \hat{\theta}_n)) \psi_{z_h|y_h}(z_h|y_h; \theta) dv_{z_h}(z_h)$ . Estimate  $(\mathbf{B}^*)^{-1}$  using the *missing-data OPG (Outer Product Gradient) covariance matrix estimator*:  $\hat{\mathbf{B}}_n^{-1} \equiv \left( n^{-1} \sum_{i=1}^n \hat{\mathbf{g}}_{y_{h_i}} (\hat{\mathbf{g}}_{y_{h_i}})^T \right)^{-1}$ . The *missing data Hessian covariance*

matrix estimator:  $\hat{\mathbf{A}}_n \equiv \nabla^2 \bar{l}_n(\hat{\boldsymbol{\theta}}_n; \mathbf{y}^n, \mathbf{h}^n)$ , is a convenient estimator for the missing-data robust covariance matrix  $\mathbf{C}^*$  given by the missing-data robust covariance matrix estimator:  $\hat{\mathbf{C}}_n \equiv (\hat{\mathbf{A}}_n)^{-1} \hat{\mathbf{B}}_n (\hat{\mathbf{A}}_n)^{-1}$ .

However, in practice, it is desirable to obtain computational formulas for estimating the asymptotic covariance matrix of the parameter estimates  $\mathbf{C}^*$  which are expressed in terms of the first and second derivatives of the complete-data probability model  $M_c \equiv \{f(\mathbf{x}; \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$  rather than the observable-data probability model  $M_o$ . In this section, such formulas will be developed for situations where the observable-data model may be misspecified following the classical derivation of the Missing Information Principle which does not explicitly assume that model misspecification may be present (Woodbury 1971; Orchard and Woodbury 1972; also see Louis 1982, eq. 3.1, and McLachlan and Krishnan 1997, p 100).

Let  $\psi_{z_h|y_h}(z_h(\mathbf{x})|y_h(\mathbf{x}); \boldsymbol{\theta}) \equiv \frac{f(\mathbf{x}; \boldsymbol{\theta})}{\int f(\mathbf{x}; \boldsymbol{\theta}) dv_{z_h}(z_h(\mathbf{x}))}$  for  $\mathbf{h} \in L^d \setminus \{0_d, \mathbf{1}_d\}$ ,  $\psi_{z_h|y_h}(z_h(\mathbf{x})|y_h(\mathbf{x}); \boldsymbol{\theta}) \equiv 1$  for  $\mathbf{h} = \mathbf{1}_d$ , and  $\psi_{z_h|y_h}(z_h(\mathbf{x})|y_h(\mathbf{x}); \boldsymbol{\theta}) \equiv f(\mathbf{x}; \boldsymbol{\theta})$  for  $\mathbf{h} = 0_d$ . Note that:  $\int \psi_{z_h|y_h}(z_h(\mathbf{x})|y_h(\mathbf{x}); \boldsymbol{\theta}) dv_{z_h}(z_h(\mathbf{x})) = 1$  for all  $\boldsymbol{\theta} \in \Theta$  and  $y_h(\mathbf{x}) \in \mathbb{R}^{p_h}$ .

**Proposition 2.** *Missing-Data Gradient Computation Formulas.*

Assume that Assumptions 1, 2, 3, and 4 hold. Then, (i)

$$\mathbf{g}_{y_h}(y_h(\mathbf{x}); \boldsymbol{\theta}) = - \int [\nabla \log f(\mathbf{x}; \boldsymbol{\theta})] \psi_{z_h|y_h}(z_h(\mathbf{x})|y_h(\mathbf{x}); \boldsymbol{\theta}) dv_{z_h}(z_h(\mathbf{x})), \tag{9}$$

and (ii)

$$q_{z_h|y_h, h}(z_h(\mathbf{x})|y_h(\mathbf{x}), \mathbf{h}; \boldsymbol{\theta}) = q_{z_h|y_h}(z_h(\mathbf{x})|y_h(\mathbf{x}); \boldsymbol{\theta}) = \psi_{z_h|y_h}(z_h(\mathbf{x})|y_h(\mathbf{x}); \boldsymbol{\theta}) \tag{10}$$

Proposition 2(i) provides an expression useful for computing the OPG missing-data covariance matrix estimator  $(\hat{\mathbf{B}}_n(\hat{\boldsymbol{\theta}}_n; \mathbf{Y}^n, \mathbf{H}^n))^{-1}$ . Moreover, when (9) is used to evaluate  $\bar{\mathbf{g}}_n(\boldsymbol{\theta}; \mathbf{y}^n, \mathbf{h}^n)$ , then we obtain Equation (2.13) from Orchard and Woodbury (1972; also see Woodbury 1971) and Equation (3.1) from Louis (1982). Proposition 2(ii) establishes that the conditional independence relation  $q_{z_h|y_h, h} = q_{z_h|y_h}$  holds and provides the convenient computational formula for an ignorable-type complete-data generation model:

$$q_{z_h|y_h, h}(z_h(\mathbf{x})|y_h(\mathbf{x}), \mathbf{h}; \boldsymbol{\theta}) = \frac{f(\mathbf{x}; \boldsymbol{\theta})}{\int f(\mathbf{x}; \boldsymbol{\theta}) dv_{z_h}(z_h(\mathbf{x}))}.$$

We now provide formal conditions for Orchard and Woodbury’s Equation (2.15) (Orchard and Woodbury 1972; also see Woodbury 1971) and Louis’s (Louis 1982) Equation (3.2) to hold by investigating the structure of  $\hat{\mathbf{A}}_n(\hat{\boldsymbol{\theta}}_n; \mathbf{Y}^n, \mathbf{H}^n)$ . When they exist, let

$$\tilde{\mathbf{A}}_n(\boldsymbol{\theta}; \mathbf{y}^n, \mathbf{h}^n) \equiv -n^{-1} \sum_{i=1}^n \int \nabla^2 \log f(\mathbf{x}_i; \boldsymbol{\theta}) \psi_{z_{h_i}|y_{h_i}}(z_{h_i}(\mathbf{x}_i)|y_{h_i}(\mathbf{x}_i); \boldsymbol{\theta}) dv_{z_{h_i}}(z_{h_i}(\mathbf{x}_i)) \tag{11}$$

and

$$\widehat{\mathbf{A}}_n(\boldsymbol{\theta}; \mathbf{y}^n, \mathbf{h}^n) \equiv -n^{-1} \sum_{i=1}^n \int \nabla^2 \log q_{z_{h_i}|y_{h_i}}(z_{h_i}(\mathbf{x}_i)|y_{h_i}(\mathbf{x}_i); \boldsymbol{\theta}) \psi_{z_{h_i}|y_{h_i}}(z_{h_i}(\mathbf{x}_i)|y_{h_i}(\mathbf{x}_i); \boldsymbol{\theta}) dv_{z_{h_i}}(z_{h_i}(\mathbf{x}_i)). \tag{12}$$

When Assumption 3 holds, then  $\tilde{\mathbf{A}}_n(\boldsymbol{\theta}; \mathbf{y}^n, \mathbf{h}^n)$  can be interpreted as the Hessian conditional complete-data information matrix and  $\widehat{\mathbf{A}}_n(\boldsymbol{\theta}; \mathbf{y}^n, \mathbf{h}^n)$  can be interpreted as the Hessian conditional missing information matrix. Note that in the special case where all records are complete:  $\widehat{\mathbf{A}}_n(\boldsymbol{\theta}; \mathbf{y}^n, \mathbf{h}^n)$  vanishes so the trace of  $\widehat{\mathbf{A}}_n(\boldsymbol{\theta}; \mathbf{y}^n, \mathbf{h}^n)$  may be interpreted as a measure of the amount of missing information.



When they exist, let  $\widetilde{\mathbf{A}}(\boldsymbol{\theta}) \equiv -\int \int \nabla^2 \log f(\mathbf{x}; \boldsymbol{\theta}) \psi_{z_h|y_h}(\mathbf{z}_h|\mathbf{y}_h; \boldsymbol{\theta}) dv_{z_h}(\mathbf{z}_h) p_{y_h, h}(\mathbf{y}_h, \mathbf{h}) dv_{y_h, h}(\mathbf{y}_h, \mathbf{h})$  and  $\widehat{\mathbf{A}}(\boldsymbol{\theta}) \equiv -\int \int \nabla^2 \log q_{z_h|y_h}(\mathbf{z}_h|\mathbf{y}_h; \boldsymbol{\theta}) \psi_{z_h|y_h}(\mathbf{z}_h|\mathbf{y}_h; \boldsymbol{\theta}) dv_{z_h}(\mathbf{z}_h) p_{y_h, h}(\mathbf{y}_h, \mathbf{h}) dv_{y_h, h}(\mathbf{y}_h, \mathbf{h})$ .

When it exists, let:

$$\widetilde{\mathbf{B}}_n(\boldsymbol{\theta}; \mathbf{y}^n, \mathbf{h}^n) \equiv n^{-1} \sum_{i=1}^n \int \nabla \log f(\mathbf{x}_i; \boldsymbol{\theta}) (\nabla \log f(\mathbf{x}_i; \boldsymbol{\theta}))^T \psi_{z_{h_i}|y_{h_i}}(\mathbf{z}_{h_i}|\mathbf{y}_{h_i}; \boldsymbol{\theta}) dv_{z_{h_i}}(\mathbf{z}_{h_i}), \tag{13}$$

which is referred to as the *OPG conditional complete-data information matrix* when A3 holds. Inspecting (13), we see that  $\widetilde{\mathbf{B}}_n(\boldsymbol{\theta}; \mathbf{y}^n, \mathbf{h}^n)$  is positive semidefinite for all  $\boldsymbol{\theta}$ . Let

$$\widehat{\mathbf{g}}_h(\mathbf{x}; \boldsymbol{\theta}) \equiv \nabla \log f(\mathbf{x}; \boldsymbol{\theta}) - \mathbf{g}_{y_h}(\mathbf{y}_h(\mathbf{x}); \boldsymbol{\theta}). \tag{14}$$

Then, when it exists,

$$\widehat{\mathbf{B}}_n(\boldsymbol{\theta}; \mathbf{y}^n, \mathbf{h}^n) \equiv n^{-1} \sum_{i=1}^n \int \widehat{\mathbf{g}}_{h_i}(\mathbf{x}_i; \boldsymbol{\theta}) (\widehat{\mathbf{g}}_{h_i}(\mathbf{x}_i; \boldsymbol{\theta}))^T \psi_{z_{h_i}|y_{h_i}}(\mathbf{z}_{h_i}|\mathbf{y}_{h_i}; \boldsymbol{\theta}) dv_{z_{h_i}}(\mathbf{z}_{h_i}), \tag{15}$$

which is referred to as the *OPG conditional missing information matrix* when Assumption 3 holds. The trace of the positive semidefinite matrix  $\widehat{\mathbf{B}}_n(\cdot; \mathbf{y}^n; \mathbf{h}^n)$  defined in (15) may be interpreted as a measure of the amount of missing data. In the special case where there are no missing data, the trace of (15) vanishes.

Let  $\widetilde{\mathbf{B}}(\boldsymbol{\theta}) \equiv \int \int \nabla \log f(\mathbf{x}; \boldsymbol{\theta}) (\nabla \log f(\mathbf{x}; \boldsymbol{\theta}))^T \psi_{z_h|y_h}(\mathbf{z}_h|\mathbf{y}_h; \boldsymbol{\theta}) dv_{z_h}(\mathbf{z}_h) p_{y_h, h}(\mathbf{y}_h, \mathbf{h}) dv_{y_h, h}(\mathbf{y}_h, \mathbf{h})$  and  $\widehat{\mathbf{B}}(\boldsymbol{\theta}) \equiv \int \int \widehat{\mathbf{g}}_h(\mathbf{x}; \boldsymbol{\theta}) (\widehat{\mathbf{g}}_h(\mathbf{x}; \boldsymbol{\theta}))^T \psi_{z_h|y_h}(\mathbf{z}_h|\mathbf{y}_h; \boldsymbol{\theta}) dv_{z_h}(\mathbf{z}_h) p_{y_h, h}(\mathbf{y}_h, \mathbf{h}) dv_{y_h, h}(\mathbf{y}_h, \mathbf{h})$

**Theorem 3. Missing Information Principle.**

Assume that Assumptions 1, 2, 4, and 5 hold. Then (i)

$$\widetilde{\mathbf{A}}_n(\boldsymbol{\theta}; \mathbf{y}^n, \mathbf{h}^n) = \widetilde{\mathbf{A}}_n(\boldsymbol{\theta}; \mathbf{y}^n, \mathbf{h}^n) - \widehat{\mathbf{A}}_n(\boldsymbol{\theta}; \mathbf{y}^n, \mathbf{h}^n). \tag{16}$$

and  $\mathbf{A}(\boldsymbol{\theta}) = \widetilde{\mathbf{A}}(\boldsymbol{\theta}) - \widehat{\mathbf{A}}(\boldsymbol{\theta})$ . In addition, (ii)

$$\widetilde{\mathbf{B}}_n(\boldsymbol{\theta}; \mathbf{y}^n, \mathbf{h}^n) = \widetilde{\mathbf{B}}_n(\boldsymbol{\theta}; \mathbf{y}^n, \mathbf{h}^n) - \widehat{\mathbf{B}}_n(\boldsymbol{\theta}; \mathbf{y}^n, \mathbf{h}^n), \tag{17}$$

$$\widehat{\mathbf{A}}_n(\boldsymbol{\theta}; \mathbf{y}^n, \mathbf{h}^n) = \widehat{\mathbf{B}}_n(\boldsymbol{\theta}; \mathbf{y}^n, \mathbf{h}^n), \text{ and} \tag{18}$$

$$\mathbf{B}(\boldsymbol{\theta}) = \widetilde{\mathbf{B}}(\boldsymbol{\theta}) - \widehat{\mathbf{B}}(\boldsymbol{\theta}).$$

Theorem 3 provides explicit regularity conditions for the Missing Information Principle (MIP) to hold for the case where either (or both) the missing data model is possibly misspecified as ignorable or the researcher’s complete-data model is misspecified. The formula  $\mathbf{A} = \widetilde{\mathbf{A}} - \widehat{\mathbf{A}}$  in Equation (16) corresponds to the MIP presented in Equation 2.15 from Orchard and Woodbury (1972; also see Jank and Booth 2003; McLachlan and Krishnan 1997, pp. 101–3, 111–13; Meng and Rubin 1991). Substituting the relation in (18) into Equation (16) yields an alternative form of the MIP discussed by Louis (1982, eq. 3.2), which is:  $\widetilde{\mathbf{A}}_n = \widetilde{\mathbf{B}}_n - \widehat{\mathbf{B}}_n$ . Both of these forms of the MIP are valid in the presence of an observable-data model which is possibly misspecified.

If the observable-data model is correctly specified, then the results of Theorem 2(ii) imply that  $\widetilde{\mathbf{A}}_n = \widetilde{\mathbf{B}}_n$  which may be combined with the results of Theorem 3 to obtain two additional MIPs that are valid for the case when the observable-data model is correctly specified:  $\widetilde{\mathbf{A}}_n = \widetilde{\mathbf{B}}_n - \widehat{\mathbf{B}}_n$  and

$\hat{\mathbf{A}}_n = \tilde{\mathbf{B}}_n - \hat{\mathbf{A}}_n$ . To our knowledge, this discussion of these four specific forms of the MIP and their validity in the presence of model misspecification has not been discussed in the literature.

### 3.4. Detection of Model Misspecification in the Presence of Missing Data

It is important to note that contrapositive of Theorem 2(ii) may be used as the basis for the detection of model misspecification in the observable-data model since Theorem 2(ii) implies that a failure of the Information Matrix Equality indicates the presence of model misspecification in the researcher's observable-data model. Thus, Theorem 2(ii) suggests a new approach to checking for model misspecification in the presence of missing data for models with assumed ignorable missing-data mechanisms. If the missing-data Hessian covariance matrix estimator  $(\hat{\mathbf{A}}_n)^{-1}$  and the missing-data OPG covariance matrix estimator  $(\hat{\mathbf{B}}_n)^{-1}$  are asymptotically different, then this indicates the presence of model misspecification in the observable-data model.

More specifically, the general theoretical framework for developing Generalized Information Matrix Tests (GIMTs) for the detection of model misspecification (Golden et al. 2013, 2016) may be applied to construct a wide range of entirely new misspecification tests for detecting model misspecification in observable-data models that assume an ignorable decimation mechanism. Further, this method is valid for both MAR and MNAR environments regardless of whether the missing-data mechanism has been correctly specified as ignorable.

In practice, researchers are often interested in whether the complete-data model, rather than the observable-data model, is misspecified. However, misspecification of the observable-data model when an ignorable missing-data mechanism is postulated implies that either the complete-data model is misspecified or the missing-data mechanism is MNAR. Thus, Theorem 2(ii) in conjunction with the GIMT Framework (Golden et al. 2013, 2016) provides a method for the detection of misspecification of complete-data probability models and also provides a method for detecting the presence of an MNAR data generating process in situations where the complete-data model is known, in fact, to be correctly specified.

Nonetheless, it is important to emphasize that the GIMT method is only capable of detecting some types of misspecification of the researcher's complete-data model. For example, suppose that the complete-data model was misspecified and the missing-data mechanism was correctly specified as ignorable. Situations may exist where the presence of the missingness "hides" misspecification in the observable data model and thus the complete-data model is not identified as misspecified. This occurs because the missing-data mechanism sometimes renders the presence of model misspecification unobservable. Further, this method for detecting model misspecification by checking the Information Matrix Equality cannot directly detect misspecification of the missing-data mechanism because the Information Matrix Equality is not functionally dependent upon the missing-data mechanism. However, misspecification of the missing-data mechanism as ignorable may be indirectly detected by the GIMT method in the special case where the complete-data model is *known* to be correctly specified and misspecification is detected in the observable-data model. In this situation, the correctly specified complete-data model (alternative) serves to provide the necessary distribution assumption (Jaeger 2006; Molenberghs et al. 2008; Rhoads 2012) to test for the presence of a nonignorable missing-data mechanism.

### 3.5. Estimating the Fraction of Missing Information with Possible Model Misspecification

**Definition 6. Fraction of Information Loss Functions.** (i) The Hessian fraction of information loss function  $\xi_{\mathbf{A}} : \Theta \rightarrow \mathbb{R}$  is defined such that for all  $\theta \in \Theta$ :  $\xi_{\mathbf{A}}(\theta) = \lambda_{\max} \left[ (\tilde{\mathbf{A}}(\theta))^{-1} \hat{\mathbf{A}}(\theta) \right]$  when  $(\tilde{\mathbf{A}}(\theta))^{-1}$  exists. The quantity  $\xi_{\mathbf{A}}^* \equiv \xi_{\mathbf{A}}(\theta^*)$  is called the Hessian fraction of information loss. (ii) The OPG fraction of information loss function  $\xi_{\mathbf{B}} : \Theta \rightarrow \mathbb{R}$  is defined such that for all  $\theta \in \Theta$ :  $\xi_{\mathbf{B}}(\theta) = \lambda_{\max} \left[ (\tilde{\mathbf{B}}(\theta))^{-1} \hat{\mathbf{B}}(\theta) \right]$  when  $(\tilde{\mathbf{B}}(\theta))^{-1}$

exists. The quantity  $\xi_B^* \equiv \xi_B(\theta^*)$  is called the OPG fraction of information loss. (iii) The robust fraction of information loss function  $\xi_C : \Theta \rightarrow R$  is defined such that for all  $\theta \in \Theta$ :  $\xi_C(\theta) = \lambda_{\max}[\widetilde{C}(\theta)\widehat{C}(\theta)]$  where  $\widehat{C}(\theta) \equiv \widetilde{C}^{-1}(\theta) - \widetilde{C}^{-1}(\theta)$  when  $\widehat{C}(\theta)$  exists. The quantity  $\xi_C^* \equiv \xi_C(\theta^*)$  is called the robust fraction of information loss.

Dempster et al. (1977; also see McLachlan and Krishnan 1997; and Little and Rubin 2002, p. 177) discuss the Hessian fraction of information loss. In particular, they define the fraction of information loss as the largest eigenvalue of  $\widetilde{A}^{-1}\widehat{A}$  evaluated at the true parameter values. The OPG and robust fractions of information loss have not been previously discussed in the literature to our knowledge. The Missing Information Principle presented in Theorem 3 provides an interpretation of the Hessian, OPG, and robust fraction of information loss functions, which measure the magnitude of a matrix relative deviation between  $\widetilde{A}, \widetilde{B}$  and  $\widetilde{C}^{-1}$  and  $A, B$  and  $C^{-1}$  relative to  $\widetilde{A}, \widetilde{B}$  and  $\widetilde{C}^{-1}$ . This interpretation is valid for environments where the researcher may have misspecified the missing data model as ignorable or misspecified the complete-data model.

Let  $\ddot{\xi}_A(\theta) = \lambda_{\max}[(\widetilde{A}(\theta))^{-1}\widehat{B}(\theta)]$  be an alternative representation of  $\xi_A(\theta)$  that follows directly from the Louis (1982, eq. 3.2) MIP  $A = \widetilde{A} - \widehat{B}$ .

The fraction of information loss function will typically take on non-negative values that are no greater than one on the parameter space because the parameter space is usually chosen so that the expected negative observed-data log-likelihood function is convex on the parameter space. However, in regions of the parameter space where the expected negative observed-data log-likelihood is not convex (e.g., saddle points), the fraction of information loss function can take on values that are greater than one (Dempster et al. 1977, p. 10; McLachlan and Krishnan 1997, p. 107).

**Theorem 4.** *Fraction of Information Loss and Missing-Data Likelihood Convexity. Assume that Assumptions 1, 2, and 4 hold.*

- (i) *Let  $\theta^+$  be a point in the interior of  $\Theta$ . Assume that  $\widetilde{A}(\theta^+)$  is positive definite. Both  $\xi_A(\theta^+) \leq 1$  and  $\ddot{\xi}_A(\theta^+) \leq 1$  if and only if there exists a non-empty open convex subset  $\Gamma$  of  $\Theta$  which contains  $\theta^+$  such that  $l$  is convex on  $\Gamma$ . In addition, the range of  $\xi_A$  and  $\ddot{\xi}_A$  on  $\Gamma$  is the set of non-negative real numbers.*
- (ii) *Assume that  $\widetilde{A}$  is positive definite on a non-empty open convex subset  $\Gamma$  of  $\Theta$ . Both  $\xi_A(\theta) \leq 1$  or  $\ddot{\xi}_A(\theta) \leq 1$  for all  $\theta \in \Gamma$  if and only if  $l$  is convex on  $\Gamma$ . In addition, the range of  $\xi_A$  and  $\ddot{\xi}_A$  on  $\Gamma$  is the set of non-negative real numbers.*

The assumption that  $\widetilde{A}$  is positive definite on the parameter space is not very restrictive in practice. Under typically assumed regularity conditions,  $\widetilde{A}$  will be positive definite on the parameter space when the complete-data negative log-likelihood is strictly convex on the parameter space. For example, standard regularity conditions will typically ensure the complete-data negative log-likelihood is strictly convex when the complete-data model is a member of the linear exponential family (Kass and Voss 1997, pp. 14–19). The additional condition  $\xi_A \leq 1$  may be semantically interpreted to mean that the observable-data expected negative log-likelihood will be convex if and only if the fraction of information loss function (i.e., amount of “missingness” in the data) is not too large over the parameter space.

**Proposition 3. Identifiability.** Assume that Assumptions 1, 2(i), 2(ii), and 4(i)(a) hold. Let  $\Gamma$  be a non-empty open subset of the parameter space  $\Theta$ . Assume that the observable-data expected negative log-likelihood  $l$  is a convex function on  $\Gamma$ . Let  $\theta^* \in \Gamma$  be a strict local minimizer of  $l$ . Then the following assertions hold.

- (i) The minimizer  $\theta^*$  is the unique global minimizer of  $l$  on  $\Gamma$ .
- (ii) If the missing-data mechanism  $p_{h|x}$  is MAR and the observable-data model is correctly specified on  $\Gamma$ , then the unique global minimizer  $\theta^*$  is the unique observable-data true parameter value for  $l$  on  $\Gamma$ .
- (iii) If the missing-data mechanism  $p_{h|x}$  is MAR and the complete-data model is correctly specified on  $\Gamma$ , then the unique global minimizer  $\theta^*$  is both the observable-data true, and complete-data true parameter value for  $l$  on  $\Gamma$ .

The assumption that  $\theta^* \in \Gamma$  is a strict local minimizer of  $l$  means that Assumption 5 holds for  $\theta^*$ . In summary, Theorem 4 provides regularity conditions that ensure the observable-data expected negative log-likelihood is convex on the parameter space provided the complete-data expected negative log-likelihood is strictly convex and the amount of missingness is sufficiently small. Given that the observable-data expected negative log-likelihood is strictly convex, then it follows (with some additional regularity conditions) from Proposition 3(iii) that for MAR data any strict local minimizer of the observable-data expected negative log-likelihood is the unique global minimizer and that global minimizer corresponds to the unique complete-data true parameter value.

#### 4. Summary and Conclusions

In practice, it is challenging to develop probability models involving missing data where either the researcher's complete-data model or the missing-data mechanism may be misspecified. To directly address this issue, a formal theoretical framework that explicitly discusses the consequences of model misspecification has been introduced, which encompasses robust estimation, inference, and specification analysis methods for models involving missing data. The main results of our theoretical framework, summarized in Table 1, follow from the key assumption of an *i.i.d.* partially observable data generating process (see Assumption 1) that may be either of type MAR or MNAR. Another key assumption is that the probability model of the *i.i.d.* missing DGP has been modeled by the researcher as a complete-data model with a postulated ignorable missing-data mechanism. Thus, our framework is specifically designed to investigate the consequences of misspecification in cases that also include the situation where the missing DGP is a MNAR mechanism, but the researcher assumes that it is MAR mechanism.

**Table 1.** Key theoretical results for probability models with assumed ignorable missing-data mechanisms.

Result	Description
Consistency Theorem T1	QMLE is a consistent estimator of the pseudo-true parameter values for observable-data probability models with an assumed ignorable missing-data mechanism in the presence of a missing DGP specified by a MAR or MNAR missing-data mechanism.
Asymptotic Distribution Theorem T2(i)	The asymptotic distribution of the QMLE is Gaussian with covariance matrix $\mathbf{C}^* = (\mathbf{A}^*)^{-1} = (\mathbf{B}^*)^{-1}$ for observable-data probability models with an assumed ignorable missing-data mechanism in the presence of a missing DGP specified by a MAR or MNAR missing-data mechanism.
Misspecification Detection Theorem T2(ii)	A GIMT may be used to detect the presence of misspecification in the observable-data probability model with an assumed ignorable missing-data mechanism in the presence of a missing DGP that is a MAR or MNAR missing-data mechanism. If this observable-data probability model is misspecified, this implies the complete-data probability model is misspecified when the missing-data mechanism is possibly misspecified but correctly specified as ignorable.
Missing Information Principles Theorem T3	Let $\bar{l}_n(\theta) = -n^{-1} \sum_{i=1}^n \log(q_{y_{n_i}}(\mathbf{y}_i; \theta))$ denote the observable-data negative average log-likelihood. The Hessian of $\bar{l}_n(\theta)$ in the presence of possible model misspecification may be estimated using either: $\mathbf{A}_n = \tilde{\mathbf{A}}_n - \hat{\mathbf{A}}_n$ and $\mathbf{A}_n = \tilde{\mathbf{A}}_n - \hat{\mathbf{B}}_n$ . If, in addition, either observable-data or complete-data model is correctly specified, then the Hessian of $\bar{l}_n(\theta)$ may be estimated using either: $\mathbf{A}_n = \tilde{\mathbf{B}}_n - \hat{\mathbf{B}}_n$ and $\mathbf{A}_n = \tilde{\mathbf{B}}_n - \hat{\mathbf{A}}_n$ .
Identifiability Proposition P3	Assume that the observable-data negative log-likelihood is convex on a convex region, $\Gamma$ , of the parameter space with a unique global minimizer in the interior of $\Gamma$ . Assume that the observable-data model is correctly specified and the missing-data mechanism is correctly specified as ignorable. Then assume that global minimizer is the observable-data model true parameter value. If, in addition, the complete-data model is correctly specified on $\Gamma$ , then the unique global minimizer on $\Gamma$ , is the complete-data model true parameter value.
Fraction of Information Loss Theorem T4	If the amount of missing data as measured by the Fraction of Information Loss is small and the complete-data model negative log-likelihood is strictly convex on a convex region of the parameter space, then with appropriate regularity conditions the observable data negative log-likelihood will be convex on that convex region of the parameter space.

**Estimation.** Theorem 1 establishes that the quasi-maximum likelihood estimator (QMLE) is a consistent estimator of the pseudo-true parameter values of the observable-data model with an assumed ignorable missing-data mechanism (MCAR, MAR). Further, QMLEs are shown to be consistent for the observable-data model in the presence of an MNAR missing-data mechanism. Our framework not only characterizes the asymptotic behavior of the quasi-maximum likelihood estimators in the presence of model misspecification and missing data, but also provides conditions for those estimators to converge to the true parameter values for the complete-data model. When the amount of missing



data as measured by the Fraction of Information Loss is small and the complete-data model negative log-likelihood is strictly convex on a convex region of the parameter space, then the observable data negative log-likelihood will be convex on the same region of the parameter space (Theorem 4). This key result supports the Identifiability Proposition 3 that shows when the observable-data model or the complete-data model true parameter values may be estimated.

**Inference.** In our framework, the correct specification of the complete-data model always implies correct specification of the observable-data model. Therefore, a key theoretical result provided in Theorem 2(i) is that if the complete-data probability model is correctly specified and the missing-data mechanism is correctly specified as ignorable, then either the robust missing-data sandwich covariance matrix estimator  $\hat{C}_n^{-1}$ , missing-data Hessian covariance matrix estimator  $\hat{A}_n^{-1}$ , or missing-data OPG covariance matrix estimator  $\hat{B}_n^{-1}$  may be used for the purposes of estimating the covariance matrix of the observable data pseudo-true parameter estimates. However, in general, only the missing-data sandwich covariance matrix estimator  $\hat{C}_n^{-1}$  can be used to obtain unbiased estimates of the covariance matrix of the observable data pseudo-true parameter estimates. Thus, for this reason, it is recommended that researchers always use the *robust* missing-data sandwich covariance matrix estimator instead of the missing-data Hessian covariance matrix estimator or the missing-data OPG covariance matrix estimator.

In addition, the Missing Information Principle (MIP) provides a computationally useful way of expressing the gradient and Hessian of the observable-data likelihood in terms of the gradient and Hessian of the complete-data likelihood. In practice, the gradient and Hessian of the complete-data likelihood are usually more available. We show that MIP as described by Louis (1982) and the MIP formula described by Dempster et al. (1977) are equivalent and valid for the special case where the researcher correctly postulates an ignorable missing-data mechanism but may possibly misspecify the complete-data model. We also provide two additional new MIPs which are valid when the researcher's observable-data probability model is correctly specified. These results are summarized in Theorem 3.

**Specification Analysis.** Our theory supports a new approach for the detection of model misspecification in missing data problems using the results of Theorem 2(ii) with the Generalized Information Matrix Test (GIMT) methods of Golden et al. (2013, 2016). Under the assumption that the missing-data mechanism is possibly misspecified, but postulated as ignorable (MAR), the GIMT method for specification testing can be used to detect the presence of model misspecification in the observable-data model. In practice, these results serve to elucidate the consequences of how ignorable and nonignorable missing-data mechanisms may affect a complete-data model, which may be either possibly correctly specified or misspecified (White 1982, 1994), when the researcher has postulated an ignorable mechanism. Table 2 depicts the relationships of missing-data mechanisms to the specification of the complete-data model when the observable-data model has been determined to be misspecified and the researcher's model of missing-data mechanism is postulated as ignorable. As shown, it can be concluded that the complete-data model is misspecified when the observable-data model is misspecified in the presence of a MAR mechanism. Notably, in the special case where the complete-data model is *known* to be correctly specified, the presence of a MNAR missing-data mechanism may be detected within this framework. This result follows as a consequence of an ignorable missing-data mechanism not affecting the specification of the observable-data model. Thus, rejecting the null hypothesis for a specification test on the observable-data model evidences the presence of a nonignorable missing-data mechanism (MNAR). Such a test may also be viewed as testing the missing at random (MAR) hypothesis (Jaeger 2006; Lu and Copas 2004; Rhoads 2012). Finally, when the complete-data model may be possibly misspecified the determination that the observable-data is misspecified based on specification testing leads to the conclusion that either the complete-data model is misspecified OR the missing-data mechanism is MNAR.

**Table 2.** Consequences of missing-data mechanism specification: detecting and interpreting misspecification in the observable-data model.

Missing-Data Mechanism <sup>1</sup>	Complete-Data Model	Conclusion when Observable-Data Model <sup>2</sup> is misspecified <sup>3</sup>
MAR	Possibly Misspecified	Complete-Data Model is Misspecified.
MNAR or MAR	Correctly Specified	Missing-Data mechanism is MNAR <sup>4</sup> .
MNAR or MAR	Possibly Misspecified	Either the Complete-Data Model is Misspecified OR the Missing-Data Mechanism is MNAR.

<sup>1</sup> Researcher's model of missing-data mechanism is postulated as ignorable. <sup>2</sup> Maximum likelihood estimates obtained by minimizing observable data likelihood (Equation (3)). <sup>3</sup> Generalized Information Matrix Tests (GIMT) (Golden et al. 2013, 2016) may be applied to detect misspecification in the observable-data model (Theorem 2(ii)). <sup>4</sup> GIMT provides a statistical test for detecting if the partially observable DGP has an MNAR missing-data mechanism in situations where the complete-data model is *known* to be correctly specified.

In conclusion, our framework provides a unified theory, which formalizes methods that provide insights into the robustness of maximum likelihood estimation, inference, and specification analysis when misspecification is present in the researcher's complete-data model or the missing data ignorability assumption is incorrect. These results will assist researchers to more explicitly understand the consequences of the inevitable occurrence of model misspecification in missing data analysis problems, as well as the use of appropriate robust methods, when pursuing the goal of developing correctly specified models.

**Author Contributions:** The mathematical framework was developed by R.M.G. and H.W. in collaboration with S.S.H. and T.M.K., R.M.G., and S.S.H. developed the missing data algorithms. H.W. did not have the opportunity to review the final version of this manuscript due to his untimely passing. H.W. was a great friend and colleague who is very much missed.

**Funding:** This research was made possible by grants from the National Institute on Drug Abuse (NIDA) (R43DA047224, PI: S.S. Henley), National Institute of General Medical Sciences (NIGMS) (R43GM123831, PI: S.S. Henley; R43GM114899, PI: S.S. Henley; R43GM106465, PI: S.S. Henley), National Institute of Mental Health (NIMH) (R43MH105073, PI: S.S. Henley), National Cancer Institute (NCI) (R44CA139607, PI: S.S. Henley), and the National Institute on Alcohol Abuse and Alcoholism (NIAAA) (R43/R44AA013768, PI: S.S. Henley; R43/44AA013351, PI: S.S. Henley) under the Small Business Innovation Research (SBIR) program.

**Acknowledgments:** The authors wish to gratefully acknowledge the support of the NIDA, NIGMS, NIMH, NCI, NIAAA, and the Department of Veterans Affairs. This paper reflects the authors' views and not necessarily the opinions or views of the NIDA, NIGMS, NIMH, NCI, NIAAA, or the Department of Veterans Affairs.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Appendix A Proofs of Theorems and Propositions

**Proof of Proposition 1.** Follows by verifying the conditions for the Strong Uniform Law of Large Numbers (White 1994, p. 351, Theorem A.2.1; Jennrich 1969, Theorem 2). □

**Proof of Theorem 1.** Follows by verifying the conditions for White's (White 1994, p. 16) Theorem 2.12 to hold and then verifying the conditions for White (1994, p. 28) Theorem 3.4 to hold. □

**Proof of Theorem 2.** (i) Follows by verifying the conditions for White (1994, pp. 89–90) Theorem 6.2 to hold. See Golden et al. (2016) for additional details. (ii) Follows by noting that:  $\int q_{y_h}(\mathbf{y}; \theta) d\nu_{y_h}(\mathbf{y}) = 1$

gives for each  $\theta \in \Theta$  and then differentiating twice to obtain:  $\nabla^2 \int q_{y_h}(\mathbf{y}; \theta) dv_{y_h}(\mathbf{y}) = 0_{r \times r}$  which after some algebra becomes:

$$\int \left( \nabla \log q_{y_h}(\mathbf{y}; \theta) (\nabla \log q_{y_h}(\mathbf{y}; \theta))^T + \nabla^2 \log q_{y_h}(\mathbf{y}; \theta) \right) q_{y_h}(\mathbf{y}; \theta) dv_{y_h}(\mathbf{y}) = 0_{r \times r} \tag{A1}$$

If the observable-data probability model is correctly specified, then a  $\theta_0^*$  exists such that:  $q_{y_h}(\mathbf{y}_h(\mathbf{x}); \theta_0^*) = p_{y_h}(\mathbf{y}_h(\mathbf{x}))$  (a.e. -  $v_{y_h}$ ) for each  $\mathbf{h} \in H$ . It then follows that:  $q_{y_h}(\mathbf{y}_h(\mathbf{x}); \theta_0^*) = p_{y_h}(\mathbf{y}_h(\mathbf{x}))$  may be substituted into (A1) to obtain:  $\mathbf{B}^* - \mathbf{A}^* = 0_{r \times r}$ .  $\square$

**Proof of Proposition 2.** (i) The gradient of  $\bar{l}_n(\theta; \mathbf{y}^n, \mathbf{h}^n)$ ,  $\bar{\mathbf{g}}_n(\theta; \mathbf{y}^n, \mathbf{h}^n) \equiv \nabla \bar{l}_n(\theta; \mathbf{y}^n, \mathbf{h}^n)$ , is:  $\bar{\mathbf{g}}_n(\theta; \mathbf{y}^n, \mathbf{h}^n) = n^{-1} \sum_{i=1}^n \mathbf{g}_{y_{h_i}}(\mathbf{y}_i; \theta)$  where  $\mathbf{g}_{y_h}(\mathbf{y}_h(\mathbf{x}); \theta) = \frac{-\int \nabla f(\mathbf{x}; \theta) dv_{z_h}(\mathbf{z}_h(\mathbf{x}))}{\int f(\mathbf{x}; \theta) dv_{z_h}(\mathbf{z}_h(\mathbf{x}))}$  and the result follows. (ii) Using the definitions of  $q_{z_h|y_h, h}$  and  $q_{y_h, h}$  in Section 2.2, substitute  $q_{y_h, h}(\mathbf{y}_h(\mathbf{x}), \mathbf{h}; \theta) = \int q_{h|x}(\mathbf{h}|\mathbf{x}) f(\mathbf{x}; \theta) dv_{z_h}(\mathbf{z}_h(\mathbf{x}))$  into

$$q_{z_h|y_h, h}(\mathbf{z}_h(\mathbf{x})|\mathbf{y}_h(\mathbf{x}), \mathbf{h}; \theta) = \frac{q_{h|x}(\mathbf{h}|\mathbf{x}) f(\mathbf{x}; \theta)}{q_{y_h, h}(\mathbf{y}_h(\mathbf{x}), \mathbf{h}; \theta)}$$

and use the ignorability Assumption 3  $q_{h|x}(\mathbf{h}|\mathbf{x}) = \eta_{h|y_h}(\mathbf{h}|\mathbf{y}_h(\mathbf{x}))$  to obtain:

$$q_{z_h|y_h, h}(\mathbf{z}_h(\mathbf{x})|\mathbf{y}_h(\mathbf{x}), \mathbf{h}; \theta) = \frac{\eta_{h|y_h}(\mathbf{h}|\mathbf{y}_h(\mathbf{x})) f(\mathbf{x}; \theta)}{\eta_{h|y_h}(\mathbf{h}|\mathbf{y}_h(\mathbf{x})) \int f(\mathbf{x}; \theta) dv_{z_h}(\mathbf{z}_h(\mathbf{x}))} = \frac{f(\mathbf{x}; \theta)}{\int f(\mathbf{x}; \theta) dv_{z_h}(\mathbf{z}_h(\mathbf{x}))}$$

Thus,  $q_{z_h|y_h, h}(\mathbf{z}_h(\mathbf{x})|\mathbf{y}_h(\mathbf{x}), \mathbf{h}; \theta) = \psi_{z_h|y_h}(\mathbf{z}_h(\mathbf{x})|\mathbf{y}_h(\mathbf{x}); \theta)$ .  $\square$

**Proof of Theorem 3.** (i) Note  $\log q_{y_{h_i}}(\mathbf{y}_{h_i}; \theta) = \log f(\mathbf{x}_i; \theta) - \log q_{z_{h_i}|y_{h_i}}(\mathbf{z}_{h_i}|\mathbf{y}_{h_i}; \theta)$ . Using Assumption 2,  $-n^{-1} \sum_{i=1}^n \nabla^2 \log q_{y_{h_i}}(\mathbf{y}_{h_i}; \theta) = -n^{-1} \sum_{i=1}^n \nabla^2 \log f(\mathbf{x}_i; \theta) + n^{-1} \sum_{i=1}^n \nabla^2 \log q_{z_{h_i}|y_{h_i}}(\mathbf{z}_{h_i}|\mathbf{y}_{h_i}; \theta)$  and then integrating gives:

$$\begin{aligned} & -n^{-1} \sum_{i=1}^n \int \nabla^2 \log q_{y_{h_i}}(\mathbf{y}_{h_i}; \theta) \psi_{z_{h_i}|y_{h_i}}(\mathbf{z}_{h_i}|\mathbf{y}_{h_i}; \theta) dv_{z_{h_i}}(\mathbf{z}_i) \\ & = -n^{-1} \sum_{i=1}^n \int \nabla^2 \log f(\mathbf{x}_i; \theta) \psi_{z_{h_i}|y_{h_i}}(\mathbf{z}_{h_i}|\mathbf{y}_{h_i}; \theta) dv_{z_{h_i}}(\mathbf{z}_i) \\ & + n^{-1} \sum_{i=1}^n \int \nabla^2 \log q_{z_{h_i}|y_{h_i}}(\mathbf{z}_{h_i}|\mathbf{y}_{h_i}; \theta) \psi_{z_{h_i}|y_{h_i}}(\mathbf{z}_{h_i}|\mathbf{y}_{h_i}; \theta) dv_{z_{h_i}}(\mathbf{z}_i) \end{aligned} \tag{A2}$$

To show the final part of (i), take expectations and use Dominated Convergence Theorem (e.g., Bartle 1966, Corollary 5.7, p. 45) with Assumption 4 to ensure the expectations exist and obtain:  $\mathbf{A} = \widetilde{\mathbf{A}} - \widehat{\mathbf{A}}$ .

(ii) Using Assumption 1, 2(i), 2(ii), 2(iii), 4, and Proposition 2(i):

$$\begin{aligned} \bar{\mathbf{A}}_n(\theta; \mathbf{y}^n, \mathbf{h}^n) & = \nabla \bar{\mathbf{g}}_n(\theta; \mathbf{y}^n, \mathbf{h}^n) = \nabla \left( n^{-1} \sum_{i=1}^n \mathbf{g}_{y_{h_i}}(\mathbf{y}_{h_i}; \theta) \right) \\ & = -\nabla \left( n^{-1} \sum_{i=1}^n \int (\nabla \log f(\mathbf{x}_i; \theta)) \psi_{z_{h_i}|y_{h_i}}(\mathbf{z}_{h_i}(\mathbf{x}_i)|\mathbf{y}_{h_i}(\mathbf{x}_i); \theta) dv_{z_{h_i}}(\mathbf{z}_{h_i}(\mathbf{x}_i)) \right) \end{aligned} \tag{A3}$$

Since Assumption 4 holds, the gradient and integral operator in (A3) can be exchanged using Bartle’s (Bartle 1966, p. 46) Corollaries 5.8 and 5.9 to obtain:

$$\begin{aligned} \bar{\mathbf{A}}_n(\boldsymbol{\theta}; \mathbf{y}^n, \mathbf{h}^n) = & -n^{-1} \sum_{i=1}^n \int \nabla^2 \log f(\mathbf{x}_i; \boldsymbol{\theta}) \psi_{z_{h_i}|y_{h_i}}(\mathbf{z}_{h_i}(\mathbf{x}_i) | \mathbf{y}_{h_i}(\mathbf{x}_i); \boldsymbol{\theta}) dv_{z_{h_i}}(\mathbf{z}_{h_i}(\mathbf{x}_i)) \\ & -n^{-1} \sum_{i=1}^n \int \nabla \log f(\mathbf{x}_i; \boldsymbol{\theta}) \nabla \psi_{z_{h_i}|y_{h_i}}(\mathbf{z}_{h_i}(\mathbf{x}_i) | \mathbf{y}_{h_i}(\mathbf{x}_i); \boldsymbol{\theta}) dv_{z_{h_i}}(\mathbf{z}_{h_i}(\mathbf{x}_i)) \end{aligned} \tag{A4}$$

$$\begin{aligned} \nabla \psi_{z_{h_i}|y_{h_i}}(\mathbf{z}_{h_i}(\mathbf{x}) | \mathbf{y}_{h_i}(\mathbf{x}); \boldsymbol{\theta}) &= \left( \frac{\nabla f(\mathbf{x}; \boldsymbol{\theta})}{f(\mathbf{x}; \boldsymbol{\theta})} \right) \left( \frac{f(\mathbf{x}; \boldsymbol{\theta})}{q_{y_{h_i}}(\mathbf{y}_{h_i}(\mathbf{x}); \boldsymbol{\theta})} \right) - \frac{f(\mathbf{x}; \boldsymbol{\theta}) \nabla q_{y_{h_i}}(\mathbf{y}_{h_i}(\mathbf{x}); \boldsymbol{\theta})}{(q_{y_{h_i}}(\mathbf{y}_{h_i}(\mathbf{x}); \boldsymbol{\theta}))^2} \\ &= \left( \nabla \log f(\mathbf{x}; \boldsymbol{\theta}) - \nabla \log q_{y_{h_i}}(\mathbf{y}_{h_i}(\mathbf{x}); \boldsymbol{\theta}) \right) \psi_{z_{h_i}|y_{h_i}}(\mathbf{y}_{h_i}(\mathbf{x}); \boldsymbol{\theta}) \end{aligned} \tag{A5}$$

Substituting (A5) into (A4) and using Assumption 4 and Bartle’s (Bartle 1966, p. 46) Corollaries 5.8 and 5.9 to exchange gradient and integral operators gives:

$$\begin{aligned} \bar{\mathbf{A}}_n(\boldsymbol{\theta}; \mathbf{y}^n, \mathbf{h}^n) &= \widetilde{\mathbf{A}}_n(\boldsymbol{\theta}; \mathbf{y}^n, \mathbf{h}^n) - \widetilde{\mathbf{B}}_n(\boldsymbol{\theta}; \mathbf{y}^n, \mathbf{h}^n) \\ &+ n^{-1} \sum_{i=1}^n \int \nabla \log f(\mathbf{x}_i; \boldsymbol{\theta}) \nabla \log q_{y_{h_i}}(\mathbf{y}_{h_i}(\mathbf{x}_i); \boldsymbol{\theta}) \psi_{z_{h_i}|y_{h_i}}(\mathbf{y}_{h_i}(\mathbf{x}_i); \boldsymbol{\theta}) dv_{z_{h_i}}(\mathbf{z}_{h_i}(\mathbf{x}_i)). \end{aligned} \tag{A6}$$

Using Assumptions 1, 2(i), 2(ii), 2(iii), 4, and Proposition 2(i), (A6) then becomes:

$$\bar{\mathbf{A}}_n(\boldsymbol{\theta}; \mathbf{y}^n, \mathbf{h}^n) = \widetilde{\mathbf{A}}_n(\boldsymbol{\theta}; \mathbf{y}^n, \mathbf{h}^n) - \widetilde{\mathbf{B}}_n(\boldsymbol{\theta}; \mathbf{y}^n, \mathbf{h}^n) + \bar{\mathbf{B}}_n(\boldsymbol{\theta}; \mathbf{y}^n, \mathbf{h}^n). \tag{A7}$$

Subtract (16) from (A7) to obtain:

$$\widehat{\mathbf{A}}_n(\boldsymbol{\theta}; \mathbf{y}^n, \mathbf{h}^n) = \widetilde{\mathbf{B}}_n(\boldsymbol{\theta}; \mathbf{y}^n, \mathbf{h}^n) - \bar{\mathbf{B}}_n(\boldsymbol{\theta}; \mathbf{y}^n, \mathbf{h}^n). \tag{A8}$$

Substituting (14) into (15) and using the definitions of  $\widetilde{\mathbf{B}}_n$  and  $\bar{\mathbf{B}}_n$  gives Equation (17). Combining Equations (16) and (A7) gives Equation (18) which establishes the first part of (ii).

To show the final part of (ii), take expectations of (17) and use Assumption 4 with the Dominated Convergence Theorem (e.g., Bartle 1966, Corollary 5.7, p. 45) to ensure the expectations exist and obtain:  $\mathbf{B} = \widetilde{\mathbf{B}} - \bar{\mathbf{B}}$ . □

**Proof of Theorem 4.** Let the vector-valued function  $\lambda : \mathbb{R}^{r \times r} \rightarrow \mathbb{R}^r$  be defined such that  $\lambda(\mathbf{Q})$  is an ordered list of the real eigenvalues of a real symmetric matrix  $\mathbf{Q}$ . Two  $r$ -dimensional square matrices  $\mathbf{Q}$  and  $\mathbf{R}$  will satisfy  $\lambda(\mathbf{Q}) = \lambda(\mathbf{R})$  if there exists a non-singular matrix  $\mathbf{T}$  such that  $\mathbf{T}^{-1} \mathbf{Q} \mathbf{T} = \mathbf{R}$  (Franklin 1968, Theorem 1, p. 76). Theorem 4(ii) will be proved first and then used to prove Theorem 4(i).

Since Assumptions 1, 2, and 4 hold then Theorem 3 implies  $\mathbf{A} = \widetilde{\mathbf{A}} - \widehat{\mathbf{A}}$  and  $\widehat{\mathbf{A}} = \widehat{\mathbf{B}}$  so that:  $\mathbf{A} = \widetilde{\mathbf{A}} - \widehat{\mathbf{B}}$ . The matrix-function  $(\widetilde{\mathbf{A}})^{-1}$  exists and is continuous on  $\Gamma$  since  $\widetilde{\mathbf{A}}$  is positive definite on  $\Gamma$  by assumption. Pre-multiply and post-multiply  $\mathbf{A} = \widetilde{\mathbf{A}} - \widehat{\mathbf{B}}$  by  $\widetilde{\mathbf{A}}^{-1/2}$  to obtain  $\widetilde{\mathbf{A}}^{-1/2} \mathbf{A} \widetilde{\mathbf{A}}^{-1/2} = \mathbf{I}_r - \widetilde{\mathbf{A}}^{-1/2} \widehat{\mathbf{B}} \widetilde{\mathbf{A}}^{-1/2}$  which implies:

$$\lambda\left(\widetilde{\mathbf{A}}^{-1/2} \mathbf{A} \widetilde{\mathbf{A}}^{-1/2}\right) = 1_r - \lambda\left(\widetilde{\mathbf{A}}^{-1/2} \widehat{\mathbf{B}} \widetilde{\mathbf{A}}^{-1/2}\right). \tag{A9}$$

Since  $\widetilde{\mathbf{A}}^{-1/2}$  is non-singular, Theorem 1 of Franklin (1968) implies:  $\lambda\left(\widetilde{\mathbf{A}}^{-1/2} \widehat{\mathbf{B}} \widetilde{\mathbf{A}}^{-1/2}\right) = \lambda\left(\widetilde{\mathbf{A}}^{-1} \widehat{\mathbf{B}}\right)$  which when substituted into (A9) gives:

$$\lambda\left(\widetilde{\mathbf{A}}^{-1/2} \mathbf{A} \widetilde{\mathbf{A}}^{-1/2}\right) = 1_r - \lambda\left(\widetilde{\mathbf{A}}^{-1} \widehat{\mathbf{B}}\right) = 1_r - \lambda\left(\widetilde{\mathbf{A}}^{-1} \mathbf{A}\right). \tag{A10}$$

where the relation  $\widehat{\mathbf{A}} = \widehat{\mathbf{B}}$  was used to obtain the right-hand side of (A10). Since  $\widetilde{\mathbf{A}}^{-1/2}$  is non-singular, the matrix  $\mathbf{A}$  is positive semidefinite if and only if  $\widetilde{\mathbf{A}}^{-1/2} \mathbf{A} \widetilde{\mathbf{A}}^{-1/2}$  is positive semidefinite. Thus, using (A10) implies that the matrix  $\mathbf{A}$  is positive semidefinite if and only if all eigenvalues of  $\widetilde{\mathbf{A}}^{-1} \widehat{\mathbf{B}}$  (or  $\widetilde{\mathbf{A}}^{-1} \widehat{\mathbf{A}}$ ) are less than or equal to one. Finally note that the Hessian of  $l$ ,  $\mathbf{A}$ , is positive semidefinite on the non-empty open convex set  $\Gamma$  if and only if  $l$  is convex on  $\Gamma$  (see Proposition 5 of Luenberger 1984, p. 180). This establishes the first part of Theorem 4(ii).

To show the second part of Theorem 4(ii), note that the matrix  $\widetilde{\mathbf{A}}^{-1/2} \widehat{\mathbf{B}} \widetilde{\mathbf{A}}^{-1/2}$  is real symmetric positive semidefinite because  $\widehat{\mathbf{B}}$  is real symmetric positive semidefinite. Since  $\widetilde{\mathbf{A}}^{-1/2}$  is non-singular, Theorem 1 of Franklin (1968) implies:  $\lambda(\widetilde{\mathbf{A}}^{-1/2} \widehat{\mathbf{B}} \widetilde{\mathbf{A}}^{-1/2}) = \lambda(\widetilde{\mathbf{A}}^{-1} \widehat{\mathbf{B}})$ . Thus, the eigenvalues of  $\widetilde{\mathbf{A}}^{-1} \widehat{\mathbf{B}}$  are non-negative and real. Since  $\widehat{\mathbf{A}} = \widehat{\mathbf{B}}$ , the eigenvalues of  $\widetilde{\mathbf{A}}^{-1} \widehat{\mathbf{A}}$  are also non-negative and real.

Since  $\widetilde{\mathbf{A}}$  is continuous on  $\Theta$  and Assumptions 1, 2, and 4 hold, it follows that there exists a sufficiently small non-empty open convex set  $\Gamma$  containing  $\theta^\dagger$  such that  $\widetilde{\mathbf{A}}$  takes on values as close to  $\widetilde{\mathbf{A}}(\theta^\dagger)$  as desired. Thus,  $\Gamma$  can be chosen so that for all  $\mathbf{u} \in U$ : If  $\mathbf{u}^T \widetilde{\mathbf{A}}(\theta^\dagger) \mathbf{u} > 0$ , then  $\mathbf{u}^T \widetilde{\mathbf{A}}(\theta) \mathbf{u} > 0$  for all  $\theta \in \Gamma$ . Use the result Theorem 4(ii) to then show that Theorem 4(i).  $\square$

**Proof of Proposition 3.** Assumptions 1, 2(i), 2(ii), and 4(i)(a) and the Dominated Convergence Theorem (e.g., Bartle 1966, Corollary 5.7, p. 45) ensure  $l$  is finite.

- (i) Since  $l$  is convex on the non-empty open convex set  $\Gamma$ , and  $\theta^*$  is a strict local minimizer of  $l$  on  $\Gamma$  then  $\theta^*$  is the unique global minimizer of  $l$  on  $\Gamma$  (Bazarrar et al. 2006, pp. 125–26).
- (ii) If the observable-data model is correctly specified on  $\Gamma$ , then the observable-data true parameter value is in  $\Gamma$ . Since the missing DGP density is MAR, every observable-data true parameter value is a global minimizer of  $l$  on  $\Gamma$ . By Proposition 3(i),  $\theta^*$  is the unique global minimizer of  $l$  on  $\Gamma$  which implies the global minimizer  $\theta^*$  is the unique observable-data true parameter value.
- (iii) If the complete-data model is correctly specified on  $\Gamma$ , then the complete-data true parameter value is in  $\Gamma$ . If there exists a complete-data true parameter value  $\theta_0$  so that  $f(\mathbf{x}; \theta_0) = p_x(\mathbf{x})$  (a.e.  $-v_x$ ) then  $\int f(\mathbf{x}; \theta_0) dv_{z_h}(\mathbf{z}_h) = \int p_x(\mathbf{x}) dv_{z_h}(\mathbf{z}_h)$  and thus  $q_{y_h}(\mathbf{y}_h(\mathbf{x}); \theta_0) = p_{y_h}(\mathbf{y}_h(\mathbf{x}))$  for all  $\mathbf{x}$  in the support of  $\mathbf{X}$  and for all  $\mathbf{h} \in H$ . Thus, correct specification of the complete-data model on  $\Gamma$  implies correct specification of the observable-data model on  $\Gamma$ . By the assumption that the missing DGP density is MAR, and the correct specification of the observable-data model, Proposition 3(i), and Proposition 3(ii), it follows that  $\theta_0$  is the unique global minimizer  $\theta^*$  of  $l$  on  $\Gamma$ .  $\square$

## References

- Abrevaya, Jason, and Stephen G. Donald. 2017. A GMM approach for dealing with missing data on regressors. *The Review of Economics and Statistics* 99: 657–662. [CrossRef]
- Agresti, Alan. 2002. *Categorical Data Analysis*, 2nd ed. New York: Wiley.
- Allison, Paul D. 2001. *Missing Data*. Sage University Papers Series on Quantitative Applications in the Social Sciences, 07–136; Thousand Oaks: Sage.
- Arminger, Gerhard, and Michael E. Sobel. 1990. Pseudo-maximum likelihood estimation of mean and covariance structure with missing data. *Journal of the American Statistical Association* 85: 195–203. [CrossRef]
- Bartle, Robert G. 1966. *The Elements of Integration*. New York: Wiley.
- Bazarrar, Mokhtar S., Hanif D. Sherali, and C. M. Shetty. 2006. *Nonlinear Programming: Theory and Algorithms*. Hoboken: Wiley.
- Berndt, Ernst K., Bronwyn H. Hall, Robert E. Hall, and Jerry A. Hausman. 1974. Estimation and inference in nonlinear structural models. *Annals of Economic and Social Measurement* 3: 653–65.
- Breunig, Christoph. 2019. Testing Missing at Random Using Instrumental Variables. *Journal of Business & Economic Statistics* 2017: 223–34.



- Chen, Hua Yun. 2004. Nonparametric and semiparametric models for missing covariates in parametric regression. *Journal of the American Statistical Association* 99: 1176–89. [\[CrossRef\]](#)
- Chen, Xiaohong, and Norman R. Swanson. 2013. *Recent Advances and Future Directions in Causality, Prediction, and Specification Analysis*. New York: Springer.
- Cho, Jin Seo, and Halbert White. 2014. Testing the Equality of Two Positive-Definite Matrices with Application to Information Matrix Testing. In *Advances in Econometrics: Essays in Honor of Peter C. B. Phillips*. Edited by Yoosoon Chang, Thomas B. Fomby and Joon Park. West Yorkshire: Emerald Group Publishing Limited, vol. 33, pp. 491–556.
- Cho, Jin Seo, and Peter C.B. Phillips. 2018. Pythagorean generalization of testing the equality of two symmetric positive definite matrices. *Journal of Econometrics* 202: 45–56. [\[CrossRef\]](#)
- Clayton, David, David Spiegelhalter, Graham Dunn, and Andrew Pickles. 1998. Analysis of longitudinal binary data from multiphase sampling. *Journal of the Royal Statistical Society Series B* 60: 71–87. [\[CrossRef\]](#)
- Dempster, Arthur. P., Nan. M. Laird, and Donald. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B* 39: 1–38. [\[CrossRef\]](#)
- Dobson, Annette J. 2002. *An Introduction to Generalized Linear Models*. New York: CRC Press.
- Efron, Bradley. 1994. Missing data, imputation, and the bootstrap. *Journal of the American Statistical Association* 89: 463–75. [\[CrossRef\]](#)
- Enders, Craig K. 2010. *Applied Missing Data Analysis*, 1st ed. New York: The Guilford Press.
- Fomby, Thomas B., and R. Carter Hill. 2003. *Maximum Likelihood Estimation of Misspecified Models: Twenty Years Later*. New York: Elsevier.
- Fomby, Thomas B., and R. Carter Hill, eds. 1998. *Messy Data—Missing Observations, Outliers, and Mixed-Frequency Data (Advances in Econometrics)*. Advances in Econometrics, No. 13. Bingley: Emerald Group Publishing Limited.
- Franklin, Joel N. 1968. *Matrix Theory*. Upper Saddle River: Prentice-Hall.
- Fridman, Arthur. 2003. Mixed Markov models. *Proceedings of the National Academy of Sciences of the United States of America* 100: 8092–96. [\[CrossRef\]](#)
- Gallini, Joan. 1983. Misspecifications that can result in path analysis structures. *Applied Psychological Measurement* 7: 125–37. [\[CrossRef\]](#)
- Gmel, Gerhard. 2001. Imputation of missing values in the case of a multiple item instrument measuring alcohol consumption. *Statistics in Medicine* 20: 2369–81. [\[CrossRef\]](#)
- Golden, Richard M. 1995. Making correct statistical inferences using a wrong probability model. *Journal of Mathematical Psychology* 39: 3–20. [\[CrossRef\]](#)
- Golden, Richard M. 1996. *Mathematical Methods for Neural Network Analysis and Design*. Cambridge: MIT Press.
- Golden, Richard M. 2000. Statistical tests for comparing possibly misspecified and nonnested models. *Journal of Mathematical Psychology* 44: 153–70. [\[CrossRef\]](#)
- Golden, Richard M. 2003. Discrepancy risk model selection test theory for comparing possibly misspecified or nonnested models. *Psychometrika* 68: 165–332. [\[CrossRef\]](#)
- Golden, Richard M., Steven S. Henley, Halbert White, and T. Michael Kashner. 2013. New directions in information matrix testing: Eigenspectrum tests. In *Recent Advances and Future Directions in Causality, Prediction, and Specification Analysis*. Edited by Xiaohong Chen and Norman R. Swanson. New York: Springer, pp. 145–77.
- Golden, Richard M., Steven S. Henley, Halbert White, and T. Michael Kashner. 2016. Generalized information matrix tests for detecting model misspecification. *Econometrics* 4: 46. [\[CrossRef\]](#)
- Gourieroux, Christian S., Alain Monfort, and Alain Trognon. 1984. Pseudo-maximum likelihood methods: Theory. *Econometrica* 52: 681–700. [\[CrossRef\]](#)
- Graham, John W., Scott M. Hofer, Stewart I. Donaldson, David P. MacKinnon, and Joseph L. Schafer. 1997. Analysis with missing data in prevention research. In *New Methodological Approaches to Alcohol Prevention Research*. Edited by Kendall J. Bryant, Michael Windle and Stephen G. West. Washington, DC: American Psychological Association.
- Greenland, Sander, and William D. Finkle. 1995. A critical look at methods for handling missing covariates in epidemiologic regression analyses. *American Journal of Epidemiology* 142: 1255–64. [\[CrossRef\]](#)
- Groenwold, Rolf H.H., Ian R. White, A. Rogier T. Donders, James R. Carpenter, Douglas G. Altman, and Karel G.M Moons. 2012. Missing covariate data in clinical research: when and when not to use the missing-indicator method for analysis. *CMAJ* 184: 1265–69. [\[CrossRef\]](#)

- Hardin, James W. 2003. The sandwich estimate of variance. In *Maximum Likelihood Estimation of Misspecified Models: Twenty Years Later*. Edited by Thomas B. Fomby and R. Carter Hill. New York: Elsevier, pp. 45–73.
- Harel, Ofer, and Xiao-Hu Zhou. 2006. Multiple imputation for correcting verification bias. *Statistics in Medicine* 25: 3769–86. [[CrossRef](#)]
- Heitjan, Daniel F. 1994. Ignorability in general incomplete-data models. *Biometrika* 81: 701–8. [[CrossRef](#)]
- Henley, Steven S., Richard M. Golden, and T. Michael Kashner. 2019. Statistical Modeling Methods: Challenges and Strategies. *Biostatistics & Epidemiology*, 1–35. [[CrossRef](#)]
- Hosmer, David W., and Stanley Lemeshow. 2000. *Applied Logistic Regression*, 2nd ed. New York: Wiley.
- Huang, Wanling, and Artem Prokhorov. 2014. A Goodness-of-Fit Test for Copulas. *Econometric Reviews* 33: 751–71. [[CrossRef](#)]
- Huber, Peter J. 1967. The behavior of maximum likelihood estimates under non-standard conditions. In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley: University of California Press, vol. 1, pp. 221–33.
- Ibragimov, Rustam, and Artem Prokhorov. 2017. *Heavy Tails And Copulas: Topics In Dependence Modelling In Economics and Finance*. Hackensack: World Scientific Publishing.
- Ibrahim, Joseph G., Chen Ming-Hui, Stuart R. Lipsitz, and Amy H. Herring. 2005. Missing-Data Methods for Generalized Linear Models: A Comparative Review. *Journal of the American Statistical Association* 100: 332–46. [[CrossRef](#)]
- Ibrahim, Joseph G., Stuart R. Lipsitz, and Ming-Hui Chen. 1999. Missing covariates in generalized linear models when the missing data mechanism is nonignorable. *Journal of The Royal Statistical Society Series B* 61: 173–90. [[CrossRef](#)]
- Jaeger, Manfred. 2006. On Testing the Missing at Random Assumption. In *Machine Learning: ECML 2006. Lecture Notes in Computer Science*. Edited by Johannes Fürnkranz, Tobias Scheffer and Myra Spiliopoulou. Berlin/Heidelberg: Springer, vol. 4212, pp. 671–78.
- Jamshidian, Mortaza, and Robert I. Jennrich. 2000. Standard errors for EM estimation. *Journal of The Royal Statistical Society Series B* 62: 257–70. [[CrossRef](#)]
- Jank, Wolfgang, and James Booth. 2003. Efficiency of Monte Carlo EM and Simulated Maximum Likelihood in Two-Stage Hierarchical Models. *Journal of Computational and Graphical Statistics* 12: 214–29. [[CrossRef](#)]
- Jennrich, Robert I. 1969. Asymptotic properties of nonlinear least squares estimators. *Annals of Mathematical Statistics* 40: 633–43. [[CrossRef](#)]
- Kashner, T. Michael, Steven S. Henley, Richard M. Golden, John M. Byrne, Sheri A. Keitz, Grant W. Cannon, Barbara K. Chang, Gloria J. Holland, David C. Aron, Elaine A. Muchmore, and et al. 2010. Studying the Effects of ACGME duty hours limits on resident satisfaction: Results from VA Learner’s Survey. *Academic Medicine* 85: 1130–39. [[CrossRef](#)]
- Kass, Robert E., and Paul W. Voss. 1997. *Geometric Foundations of Asymptotic Inference*. New York: Wiley.
- Kenward, Michael G., and Geert Molenberghs. 1998. Likelihood based frequentist inference when data are missing at random. *Statistical Science* 13: 236–47. [[CrossRef](#)]
- King, Gary, James Honaker, Anne Joseph, and Kenneth Scheve. 2001. Analyzing incomplete political science data: An alternative algorithm for multiple imputation. *American Political Science Association* 95: 49–69. [[CrossRef](#)]
- Kosinski, Andrzej S., and Huiman X. Barnhart. 2003a. A global sensitivity analysis of performance of a medical diagnostic test when verification bias is present. *Statistics in Medicine* 22: 2711–21. [[CrossRef](#)]
- Kosinski, Andrzej S., and Huiman X. Barnhart. 2003b. Accounting for Nonignorable Verification Bias in Assessment of Diagnostic Tests. *Biometrics* 59: 163–71. [[CrossRef](#)]
- Kullback, Solomon, and Richard A. Leibler. 1951. On information and sufficiency. *Annals of Mathematical Statistics* 22: 79–86. [[CrossRef](#)]
- Leke, Collins Achepeh, and Tshilidzi Marwala. 2019. *Deep Learning and Missing Data in Engineering Systems*, 1st ed. Cham: Springer Nature Switzerland.
- Liang, Kung-Yee, and Scott L. Zeger. 1986. Longitudinal data analysis using generalized linear models. *Biometrika* 73: 13–22. [[CrossRef](#)]
- Little, Roderick J. A., and Donald B. Rubin. 2002. *Statistical Analysis with Missing Data*, 2nd ed. New York: Wiley.

- Little, Roderick J., Ralph D'Agostino, Michael L. Cohen, Kay Dickersin, Scott S. Emerson, John T. Farrar, Constantine Frangakis, Joseph W. Hogan, Geert Molenberghs, Susan A. Murphy, and et al. 2012. The prevention and treatment of missing data in clinical trials. *The New England Journal of Medicine* 367: 1355–60. [[CrossRef](#)]
- Little, Roderick J.A. 1994. A class of pattern-mixture models for multivariate incomplete data. *Biometrika* 81: 471–83. [[CrossRef](#)]
- Little, Roderick J.A. 1988. A Test of Missing Completely at Random for Multivariate Data with Missing Values. *Journal of the American Statistical Association* 83: 1198–202. [[CrossRef](#)]
- Littman, Michael L. 2009. A tutorial on partially observable Markov decision processes. *Journal of Mathematical Psychology* 53: 119–25. [[CrossRef](#)]
- Louis, Thomas A. 1982. Finding the Observed Information Matrix when Using the EM Algorithm. *Journal of The Royal Statistical Society Series B* 44: 226–33. [[CrossRef](#)]
- Luenberger, David G. 1984. *Linear and Nonlinear Programming*, 2nd ed. Massachusetts: Addison-Wesley.
- . Lu, Guobing, and John B. Copas. 2004. Missing at Random, Likelihood Ignorability and Model Completeness. *The Annals of Statistics* 32: 754–65.
- Markovskiy, Ivan. 2017. A Missing Data Approach to Data-Driven Filtering and Control. *IEEE Transactions on Automatic Control* 62: 1972–78. [[CrossRef](#)]
- McCullagh, P., and John A. Nelder. 1989. *Generalized Linear Models*. New York: Chapman and Hall.
- McDonough, Ian K., and Daniel L. Millimet. 2016. *Missing Data, Imputation, and Endogeneity*. Bonn: IZA Institute of Labor Economics.
- McLachlan, Geoffrey, and Thriyambakam Krishnan. 1997. *The EM Algorithm and Extensions*. New York: Wiley.
- Meng, Xiao-Li, and Donald B. Rubin. 1991. Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. *Journal of the American Statistical Association* 86: 899–909. [[CrossRef](#)]
- Miller, J. 2010. Isaac 2010. Cointegrating regressions with messy regressors and an application to mixed-frequency series. *Journal of Time Series Analysis* 31: 255–77.
- Molenberghs, Geert, and Michael Kenward. 2007. *Missing Data in Clinical Studies*. New York: Wiley.
- Molenberghs, Geert, Bart Michiels, Michael G. Kenward, and P.J. Diggle. 1998. Missing data mechanisms and pattern-mixture models. *Statistica Neerlandica* 52: 153–61. [[CrossRef](#)]
- Molenberghs, Geert, Caroline Beunckens, and Cristina Sotito. 2008. Every missingness not at random model has a missingness at random counterpart with equal fit. *Journal of The Royal Statistical Society Series B* 70: 371–88. [[CrossRef](#)]
- Molenberghs, Geert, Garrett Fitzmaurice, Michael G. Kenward, Anastasios Tsiatis, and Geert Verbeke. 2014. *Handbook of Missing Data Methodology*, 1st ed. London: Chapman & Hal, Boca Raton: CRC.
- Murray, Gordon D. 1977. Contribution to the discussion of paper by A. P. Dempster, N. M. Laird, and D. B. Rubin. *Journal of The Royal Statistical Society Series B* 39: 27–28.
- Nielsen, Søren Feodor. 1997. Inference and missing data: Asymptotic results. *Scandinavian Journal of Statistics* 24: 261–74. [[CrossRef](#)]
- Orchard, Terence, and Max A. Woodbury. 1972. A missing information principle: Theory and applications. *Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability* 1: 697–715.
- Parzen, Michael, Stuart R. Lipsitz, Garrett M. Fitzmaurice, Joseph G. Ibrahim, and Andrea Troxel. 2006. Pseudo-likelihood methods for longitudinal binary data with nonignorable missing responses and covariates. *Statistics in Medicine* 25: 2784–96. [[CrossRef](#)]
- Prokhorov, Artem, Ulf Schepsmeier, and Yajing Zhu. 2019. Generalized Information Matrix Tests for Copulas. *Econometric Reviews* 25: 1024–54. [[CrossRef](#)]
- Rhoads, Christopher H. 2012. Problems with Tests of the Missingness Mechanism in Quantitative Policy Studies. *Statistics, Politics, and Policy* 3: 6. [[CrossRef](#)]
- Rubins, James M., and Naisiyin Wang. 2000. Inference for imputation estimators. *Biometrika* 87: 113–24. [[CrossRef](#)]
- Royall, Richard M. 1986. Model robust confidence intervals using maximum likelihood estimators. *International Statistical Review* 54: 221–26. [[CrossRef](#)]
- Rubin, Donald B. 1976. Inference and missing data. *Biometrika* 63: 581–92. [[CrossRef](#)]
- Rubin, Donald B. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.
- Rubin, Donald B. 1996. Multiple imputation after 18+ years. *Journal of the American Statistical Association* 91: 473–89. [[CrossRef](#)]

- Ryden, Tobias, and D. M. Titterton. 1998. Computational Bayesian analysis of hidden Markov models. *Journal of Computational and Graphical Statistics* 7: 194–211.
- Schafer, Joseph L. 1997. *Analysis of Incomplete Multivariate Data*. New York: Chapman and Hall.
- Schenker, Nathaniel, and A. H. Welsh. 1988. Asymptotic results for multiple imputation. *Annals of Statistics* 16: 1550–66. [[CrossRef](#)]
- Schepsmeier, Ulf. 2015. Efficient information based goodness-of-fit tests for vine copula models with fixed margins: A comprehensive review. *Journal of Multivariate Analysis* 138: 34–52. [[CrossRef](#)]
- Schepsmeier, Ulf. 2016. A goodness-of-fit test for regular vine copula models. *Econometric Reviews* 38: 25–46. [[CrossRef](#)]
- Serfling, Robert J. 1980. *Approximation Theorems of Mathematical Statistics*, 2nd ed. New York: Wiley-Interscience.
- Sung, Yun Ju, and Charles J. Geyer. 2007. Monte Carlo likelihood inference for missing data models. *The Annals of Statistics* 35: 990–1011. [[CrossRef](#)]
- Troxel, Andrea B., Diane L. Fairclough, Desmond Curran, and Elizabeth A. Hahn. 1998. Statistical Analysis of Quality of Life with Missing Data in Cancer Clinical Trials. *Statistics in Medicine* 17: 653–66. [[CrossRef](#)]
- Troxel, Andrea B., Stuart R. Lipsitz, and David P. Harrington. 1998. Marginal models for the analysis of longitudinal measurements with nonignorable nonmontone missing data. *Biometrika* 85: 661–72. [[CrossRef](#)]
- Verbeek, Marno. 2008. *A Guide to Modern Econometrics*. New York: Wiley.
- Verbeke, Geert, and Emmanuel Lesaffre. 1997. The effect of misspecifying the random-effects distribution in linear mixed models for longitudinal data. *Computational Statistics & Data Analysis* 23: 541–56.
- Visser, Ingmar. 2011. Seven things to remember about hidden Markov models: A tutorial on Markovian models for time series. *Journal of Mathematical Psychology* 55: 403–15. [[CrossRef](#)]
- Vittinghoff, Eric, David V. Glidden, Stephen C. Shiboski, and Charles E. McCulloch. 2012. *Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models*, 2nd ed. New York: Springer.
- Wall, Melanie M., Yu Dai, and Lynn E. Eberly. 2005. GEE estimation of a misspecified time-varying covariate: An example with the effect of alcoholism treatment on medical utilization. *Statistics in Medicine* 24: 925–39. [[CrossRef](#)]
- Wang, Naisyin, and James M. Robins. 1998. Large-sample theory for parametric multiple imputation procedures. *Biometrika* 85: 935–48. [[CrossRef](#)]
- Wedderburn, Robert William Maclagan. 1974. Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika* 61: 439–47.
- Wei, Bo-Cheng. 1998. *Exponential Family Nonlinear Models*. New York: Springer.
- White, Halbert. 1980. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* 48: 817–38. [[CrossRef](#)]
- White, Halbert. 1982. Maximum likelihood estimation of misspecified models. *Econometrica* 50: 1–25. [[CrossRef](#)]
- White, Halbert. 1984. *Asymptotic Theory for Econometricians*. New York: Academic Press.
- White, Halbert. 1994. *Estimation, Inference and Specification Analysis*. New York: Cambridge University Press.
- Woodbury, Max. A. 1971. Contribution to the discussion of “The analysis of incomplete data” by Herman. O. Hartley and Ronald. R. Hocking. *Biometrics* 27: 808–13.
- Wooldridge, Jeffrey M. 2004. *Inverse Probability Weighted Estimation for General Missing Data Problems*. Cemmap Working Paper, No. CWP05/04. London: Centre for Microdata Methods and Practice (cemmap).
- Yuan, Ke-Hai. 2009. Normal distribution based pseudo ML for missing data: With applications to mean and covariance structure analysis. *Journal of Multivariate Analysis* 100: 1900–18. [[CrossRef](#)]
- Zhao, Lue Ping, Lipsitz Stuart, and Danika Lew. 1996. Regression analysis with missing covariate data using estimating equations. *Biometrics* 52: 1165–82. [[CrossRef](#)]
- Zhou, Xiao-Hua, Chuan Zhou, Danping Lui, and Xaiobo Ding. 2014. *Applied Missing Data Analysis in the Health Sciences*, 1st ed. Statistics in Practice. New York: Wiley.
- Zhu, Yajing. 2017. Dependence Modelling and Testing: Copula and Varying Coefficient Model with Missing Data. Ph.D. thesis, Concordia University, Montreal, QC, Canada.

