The effect of sample size and missingness on inference with missing data

Julian Morimoto
Harvard University, Cambridge, Massachusetts, USA

ABSTRACT
When are inferences (whether Direct-Likelihood, Bayesian, or Frequentist) obtained from partial data valid? This article answers this question by offering a new asymptotic theory about inference with missing data that is more general than existing theories. It proves that as the sample size increases and the extent of missingness decreases, the average-loglikelihood function generated by partial data and that ignores the missingness mechanism will converge in probability to that which would have been generated by complete data; and if the data are Missing at Random, this convergence depends only on sample size. Thus, inferences from partial data, such as posterior modes, confidence intervals, likelihood ratios, test statistics, and indeed, all quantities or features derived from the partial-data loglikelihood function, will be consistently estimated. Additionally, the missing data mechanism has asymptotically no effect on parameter estimation and hypothesis testing if the data are Missing at Random. This adds to previous research which has only proved the consistency and asymptotic normality of the posterior mode. Practical implications are discussed, and the theory is illustrated through simulation using a previous study of International Human Rights Law.

ARTICLE HISTORY
Received 2 February 2022
Accepted 23 November 2022

KEYWORDS
Incomplete data; sample size and missing data mechanism; partial likelihood; asymptotic inference with missing data

1. Introduction

Missing data are a big problem in scientific research, and it is not always clear when inferences drawn from incomplete datasets will be similar to those that would have been drawn from complete ones. In the social sciences, on average, about 50% of survey respondents will fail to provide complete answers (King et al. 2001). Missing phenotype data is also often an inevitable and major hurdle in genetic experiments where quantitative traits are not observed or certain expressions are not realized (Nielsen and Mørup 2014; Bobb et al. 2011). In astronomy, unfavorable weather, man-made radio emissions, and emissions from the sun can cause missing-data (George, Ambika, and Misra 2015). Consequently, researchers have tried to find ways to draw valid inferences from partially complete datasets. One technique is to draw inferences from the likelihood function generated by partial data (Little and Rubin 2002; Little, Rubin, and Zangeneh 2017), which has proven quite successful in many contexts (Murray 2018; Shin, Davison, and
Long 2017; von Hippel 2016; Honaker, King, and Blackwell 2011; Nielsen 2007). This approach involves approximating the loglikelihood function generated by the partially observed dataset by “integrating” over the missing values (this process is shown in more detail in the proofs of Theorems 1 and 2). One can then use the resulting function to draw Direct-Likelihood, Bayesian, or Frequentist inferences from the partial data (Seaman et al. 2013).

It is important to note that this function ignores the missingness mechanism (Little and Rubin 2002). As a terminological point, whenever this article refers to average-loglikelihood/loglikelihood/likelihood functions generated by partial data, this article means those functions that focus only on the distribution of the underlying data (whether observed or missing) and that do not take into account the missingness mechanism. Focusing on these functions is consistent with the approach of other scholars who have done so given that it is difficult, if not impossible, to know anything about the missingness mechanism in practice (Little and Rubin 2002; Little, Rubin, and Zangeneh 2017).

Despite the popularity of the partial-likelihood approach, much is still unknown about when this method yields asymptotically valid results. Little and Rubin (2002) note that:

In one high-level sense, there is no formal difference between [likelihood] or Bayes inferences for incomplete data and [likelihood] or Bayes inference for complete data? Asymptotic standard errors obtained from the information matrix are somewhat more questionable with missing data, however, because the observed data do not generally constitute an iid sample, and simple results that imply the large sample normality of the likelihood function do not immediately apply. Other complications arise from dealing with the process that creates missing data.

Thus, it is not immediately obvious whether (and which) inferences drawn from the partial likelihood function will be asymptotically valid. Current asymptotic theories are limited in that they have only proved the consistency and asymptotic normality of the partial-likelihood function’s posterior mode when the data are Missing at Random (Little and Rubin 2002; Little, Rubin, and Zangeneh 2017; Seaman et al. 2013; Takai and Kano 2013; Nielsen 1997). This leaves many important unanswered questions. Will other quantities derived from the loglikelihood such as uncertainty estimates, confidence intervals, likelihood ratios, and test statistics be consistently estimated? What impact does the extent of missingness have, if any, on inference beyond consistency and asymptotic normality, and why? How does sample size interact with missingness? Will inferences from Missing not at Random data be analogously valid under certain conditions? When are inferences about subvectors of parameters consistent (Little, Rubin, and Zangeneh 2017)? What effect does the missingness mechanism have on parameter estimation and hypothesis testing? Much is still unknown about whether and when inferences obtained from partial data will be asymptotically valid.

This article fills in this knowledge gap by offering a new asymptotic theory about how sample size and the missingness mechanism affect the validity of inferences from missing data. It proves two Theorems that explain the effect of sample size and missingness on inference with missing data. Theorem 1 proves that the average-loglikelihood function (ignoring the missingness mechanism) generated by partial data will converge in probability to the function that would have been generated by complete data as the sample size increases and the extent of missingness decreases. Theorem 2 proves that
for data that are Missing at Random (technically “(Almost) Everywhere Missing at Random”, which is different from “Realized Missing at Random” (Seaman et al. 2013), but this point will be further discussed below in Theorem 2), this convergence depends only on sample size. This suggests that inferences obtained from partial datasets will approximate those that would have been obtained from complete ones.

These two theorems also have three important practical implications. First, inferences drawn from partial datasets will be close to their true values given a sufficiently large sample size and a sufficiently small amount of missing data. Second, when the data are Missing at Random, this approximation depends only on the sample size and not the proportion of missing data. This means that a researcher could have a dataset with a large amount of missing data and still be able to draw valid inferences if that dataset is sufficiently large. The missingness mechanism will also have no asymptotic effect on parameter estimation and hypothesis testing. Third, it suggests that Maximum Likelihood inference is better than Multiple Imputation when working with partially-observed datasets with many observations.

The rest of this article is structured as follows. Section 2 states and proves the two Theorems. After Theorem 1 is proved, the first practical implication mentioned above is discussed in more detail. The remaining two are discussed after the proof of Theorem 2. These Theorems are then verified via simulation. To do so, this article simulates data using a dataset from a previous study on the impact of International Human Rights Law on public support for human rights abuses. Through this simulation, this article verifies that (1) the average-loglikelihood function (ignoring the missingness mechanism) generated by partial data will approximate that which would have been generated by complete data as sample size increases and missingness decreases, (2) this approximation depends only on sample size for Missing at Random data, (3) this leads to valid inferences from missing data, which will tend to be better than those obtained through Multiple Imputation.

This article offers a new theory about the effect of sample size and missingness on inference with missing data. This theory sheds light on the factors that affect the validity of Direct-Likelihood, Bayesian, and Frequentist inferences obtained from missing data that ignore the missingness mechanism, and proves a more general result than has been demonstrated in previous literature. It yields three practical implications that help researchers know when inferences drawn from partial datasets will approximate those obtained from complete ones, the impact (or lack thereof, or “ignorability”) of the proportion of missing data in large samples, and the benefits of using Maximum Likelihood inference instead of Multiple Imputation.

2. Inference with missing data: formal explanation

This section proves that as the extent of missingness decreases and the sample size increases, the average-loglikelihood function (ignoring the missingness mechanism) generated by a partial dataset will converge in probability to the average-loglikelihood function generated by a complete one. Moreover, it shows that if the data are Missing at Random (MAR), then this convergence depends only on the sample size. This result adds to the literature by proving that all possible inferences, such as posterior modes,
uncertainty estimates, confidence intervals, likelihood ratios, test statistics, and any possible features or quantities that one can draw from a loglikelihood function generated by partial data will approximate their true values (what they would have been without missing data), whereas previous results have shown only that the posterior mode of the partial loglikelihood will tend to approximate its analogue in a complete dataset (Little and Rubin 2002; Little, Rubin, and Zangeneh 2017; Seaman et al. 2013; Takai and Kano 2013; Nielsen 1997). The results also imply that if the data are Missing at Random, then the missing data mechanism will have asymptotically no effect on parameter estimation and hypothesis testing. After the proof of each theorem, practical implications are discussed.

### 2.1. Formal setup

The setup, including relevant concepts, terminology, and notation, for the formal results in this section is as follows. Let

1. \( X : \Omega \rightarrow \mathcal{X} \) be a multivariate random variable on some probability space, \((\Omega, \mathcal{F}, P)\), onto some compact subset of \( \mathbb{R}^m \), \( \mathcal{X} \).
2. \( X \sim \mathcal{D}(\theta^\prime) \) where \( \mathcal{D} \) is a continuous probability density function with parameter \( \theta^\prime \in \mathbb{R}^r \).
3. \( \{P_k\}_{k=1}^\infty \) is a sequence of probability measures on \( \{0, 1\}^m \) that depend on values of \( (X, \psi) \in \mathbb{R}^m \times \mathcal{K}_\psi \), where \( \mathcal{K}_\psi \) is some space, and \( \{\psi_k\}_{k=1}^\infty \) is a corresponding sequence of parameters in \( \mathcal{K}_\psi \) such that \( \forall X \in \mathbb{R}^m, M \in \{0, 1\}^m \),

\[
\lim_{k \to \infty} P_k(M|X, \psi_k) = 1_{\{1\}^m}(M)
\]

Note that this is a translation of the phrase “as the extent of missingness decreases” into mathematical language. This translation considers a sequence of probability measures that characterize the distribution of the different possible missingness vectors. “As the extent of missingness decreases” centers around this sequence converging pointwise to the indicator function that takes on 1 if the missingness vector is entirely 1’s (indicating that the data are fully observed), and 0 otherwise. One can think of this as a kind of weak convergence of measures result. Note also that so far, no assumptions are made about whether the data are Missing at Random. Theorem 2 discusses what happens when the Missing at Random assumption holds.

1. \( O(X, M) \) denotes the “subvector” of \( X \in \mathbb{R}^m \) that represents values that are observed when given a missingness vector \( M \in \{0, 1\}^m \) (this notation is also used in Seaman et al. (2013)). In particular, \( O(X, M) \) is a function of \( X \) and \( M \) that returns a vector whose \( j \)th entry is the \( j \)th entry of \( X \) if the \( j \)th entry of \( M \) is 1, and a “placeholder”, \(*\), otherwise. For example, \( O([1, 2, 3], [1, 0, 1]) = [1, *, 3] \), where \( * \) is a placeholder for the missing value. No vector space operations are performed on \( O(X, M) \) directly. Rather, the only significance of \( O(X, M) \) is that it is used as an argument of the partial likelihood function, discussed further below in item (7).
2. \( \{F_i\}_{i=1}^{\infty} \) is a sequence of randomly sampled independent outcomes in \((\Omega, \mathcal{F}, P)\) of non zero probability.

3. \( \{M_i\}_{i=1}^{\infty} \) is a corresponding sequence in \(\{0,1\}^m\).

4. \( f \) is the corresponding loglikelihood function that ignores the missing data mechanism and only depends on \( \theta \in \mathbb{R}^r \), a compact subset of \( \mathbb{R}^r \). For a complete realization of \( X \), \( f(X|\theta) \) is equal to the probability of observing \( X \) when given \( X \sim D(\theta) \). To evaluate \( f \) for vectors with missing observations, the following conditional expectation is taken: \( f(O(X,M)|\theta) = \mathbb{E}[f(X^*|\theta)|O(X,M) = O(X^*, M)] \) (Little and Rubin 2002; Seaman et al. 2013).

Note that in this setup, the parameter is arbitrary. Therefore, Theorem 1 will hold if one considers only a subvector of a parameter of interest, and Theorem 2 will hold if the data are “Partially Missing at Random” (Little, Rubin, and Zangeneh 2017).

In this article, the subscript on the expectation operator indicates the variable with respect to which the expectation is taken. The Theorems assume random variables with continuous distributions, but they can be easily generalized to discrete ones by way of simple functions.

### 2.2. Formal results

With this setup, 2 important lemmas hold. After stating these lemmas, this article will provide Theorem 1. All proofs are contained in the Appendix.

**Lemma 1.** \( \lim_{k \to \infty} \mathbb{E}_{M \in \{0,1\}^m} [f(O(X,M)|\theta)|X, \psi_k] = f(X|\theta) \) almost uniformly on \( K_X \times K_\theta \).

**Lemma 2.** There exists a sequence of simple functions \( \{\phi_n\}_{n=1}^{\infty} \) of the form \( \sum_{j=1}^{n} z_j 1_{E_j} \) such that \( \lim_{n \to \infty} \phi_n = X \) uniformly. Moreover,

1. \( f(O(X,M)|\theta) = f(O(\phi_n,M)|\theta) + \varepsilon(n) \) uniformly over \( K_X \times K_\theta \) for all \( M \in \{0,1\}^m \), and
2. \( f(X|\theta) = f(\phi_n|\theta) + \varepsilon(n) \) uniformly over \( K_X \times K_\theta \).

**Theorem 1** will now be stated. The main intuition behind Theorem 1 is this. The average-loglikelihood function (ignoring the missingness mechanism) generated by partial data can be viewed as a summation of some function of different realizations of a random variable that are partially observed due to the introduction of a missingness vector. As the sample size increases, more and more of each realization is added, and one can apply the Law of Large Numbers to all the terms of the summation that correspond to that particular realization of the random variable. Essentially, one is splitting up the terms of the summation into different “classes”, and applying the Law of Large Numbers to each class. Through this application, one shows that this summation converges to some vector. One can apply the Portmanteau Theorem (i.e., Lemma 1) to show that the value of the resulting vector depends on the missingness mechanism, and that this vector converges to the function of the realization of the random variable that is completely observed as the extent of missingness decreases.
Theorem 1.

\[
\lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} f(O(X(F_i), M_i)|\theta)
= \lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} f(X(F_i)|\theta) + \frac{1}{N} \sum_{i=1}^{N} \epsilon_i(k) + \frac{2}{N} \sum_{i=1}^{N} \epsilon_i(n)
\]

in probability on \(K_X \times K_\theta\), where \(\{\psi_k\}_{k \in \mathbb{N}}\) is a sequence of parameters generating the \(\{M_i\}_{i \in \mathbb{N}}\), and \(\{\phi_n\}_{n=1}^{\infty}\) is a sequence of simple functions such that \(\lim_{n \to \infty} \phi_n = X\) uniformly over \(K_\theta\).

Prior to discussing the implications of Theorem 1, it is worth explaining the role of the simple function, and \(n\) above. The main idea here is that the simple functions capture the idea of approximation, so the term \(\epsilon(n)\) can be treated as 0 for practical purposes. One can approximate a random variable \(X\) to some degree using simple functions (since no machine can ever really achieve an approximation of infinite precision, this process accurately reflects how most approximation algorithms work). The loglikelihood functions that are generated from a partial dataset will in some sense “approximate an approximation” (by way of simple functions) of the loglikelihood functions generated by a complete dataset. When one fixes the precision with which they approximate \(X\), thereby fixing \(n\) finite but sufficiently large, one need not worry about the \(n\) and \(\epsilon(n)\) terms in the result. Intuitively, if one wants to have more and more precise approximations, the sample size \(N\) just has to be greater than or equal to \(n\) in order for the result to hold, which the Theorem suggests.

Theorem 1 shows that the average-loglikelihood function (ignoring the missingness mechanism) generated by partial data will approximate that which would have been generated by complete data as sample size increases and the extent of missingness decreases. This implies that inferences drawn from the likelihood function generated by partial data such as posterior modes, uncertainty estimates, confidence intervals, likelihood ratios, and test statistics will be close to their analogous values from the complete dataset. The posterior modes are obtained by taking the mode of the parameters given the distribution generated by the average-loglikelihood function, and the uncertainty estimates are obtained by measuring the distance between the parameters in this space and the mode given these distributions. Since these distributions tend to be identical in large samples with small amounts of missingness, so will the modes and corresponding uncertainties, and hence the confidence intervals. Indeed, because these distributions will become approximately identical, any feature or quantity that can be obtained from the partial-data loglikelihood function will approximate its true value. One practical implication of this is that inferences from partial datasets will tend to be valid in sufficiently large samples with sufficiently small amounts of missing data. The next section proves the interesting result that if the data are Missing at Random, then this approximation depends only on the sample size and not the extent of missingness. Prior to stating and proving Theorem 2, Missing at Random will be defined. It is similar to the definition of “Everywhere Missing at Random” given by Seaman et al. (2013).
Definition 2.1 (Missing at Random). The data are Missing at Random if

\[ P_k(M|X_1) = P_k(M|X_2) \forall M, X_1, X_2 : O(X_1, M) = O(X_2, M) \] almost surely

This is technically not the same definition that (Little and Rubin 2002) use to prove the consistency of the posterior mode (Seaman et al. 2013), but this does not make a difference when one considers asymptotic inference. The definition they use is “Realized Missing at Random”, which is analogous to the “Everywhere Missing at Random” definition, except it holds only for the realized \( M \)'s. However, it does not make much of a difference for large sample sizes, because as more and more observations are drawn, the space of realized \( M \)'s will contain all possible \( M \)'s of non zero probability. As a result, the application of the “Realized” definition will result in its application to almost every \( M \) that can be observed, which makes it akin to the “(Almost) Everywhere” definition. Thus, while this definition is not technically the same as that used to prove previous results about consistency of the posterior mode, it does not pose an issue here because this article is concerned with very large samples. With this definition in hand, this article proceeds to Theorem 2.

The key intuition behind Theorem 2 is that for any missingness pattern, the complete data can be partitioned into “classes” that each contain partial observations that will be identical under the specified pattern. For example, if one considers the class of missingness vectors where only the first entry is observed, then \((1, 0)\) and \((1, 1)\) will belong to the same class generated by that missingness vector. The expectation operator can be taken over these different classes to show that the error induced by the missingness mechanism tends to zero as the sample size increases if the data are Missing at Random.

**Theorem 2.** If the data are Missing at Random, then

\[ \lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} f(O(X(F_i), M_i)|\theta) = \lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} f(X(F_i)|\theta) + 2 \sum_{i=1}^{N} \epsilon_i(n) \]

in probability on \( K_X \times K_{\theta} \) for any parameter \( \psi \in K_{\psi} \) that describes the missingness mechanism from which the \( \{M_i\}_{i \in \mathbb{N}} \) are generated, and where \( \{\phi_n\}_{n=1}^{\infty} \) is a sequence of simple functions such that \( \lim_{n \to \infty} \phi_n = X \) uniformly over \( K_{\theta} \).

Theorem 2 shows that if the data are Missing at Random, then the convergence of the average-loglikelihood function (ignoring the missingness mechanism) generated by partial data to that which would have been generated by complete data depends only on the sample size and not the extent of missingness. In other words, the effect of the missing data mechanism becomes arbitrarily small if the data are missing at random. This has three key implications for when data are Missing at Random. First, the validity of inferences drawn from partial data depends only on sample size and not the extent of missingness. This means that one could have an incredibly large amount of missing data, and their inferences drawn from the average-loglikelihood would still tend to be valid and consistent for sufficiently large datasets (obviously the more missing data one has, the larger the dataset that one might need). This is because the posterior modes, uncertainty estimates, confidence intervals, likelihood ratios, test statistics, and any other quantities that can be obtained from the likelihood function will get arbitrarily close to their true values no matter the missingness mechanism, so long as the data are Missing
at Random. Practically speaking, this means that even if one has 90% of their data missing, they can still do valid inference on that partial dataset. Any quantity they estimate from the partial loglikelihood will approximate its complete-data counterpart. So far, only the posterior mode has been shown to be consistently estimated in previous literature (Little and Rubin 2002; Little, Rubin, and Zangeneh 2017; Seaman et al. 2013; Takai and Kano 2013; Nielsen 1997). This is what makes this result noteworthy.

Second, the missing data mechanism will asymptotically have no effect on parameter estimation and hypothesis testing if the data are Missing at Random. Parameter estimation and hypothesis testing are done using precisely this likelihood function. The former is done analyzing the extremums of the likelihood function, $\frac{1}{N} \sum_{i=1}^{N} f(O(X(F_i), M_i)|\theta)$ on a set of parameters. Since the missingness mechanism has asymptotically no effect on the function if the data are Missing at Random, it follows that the missingness mechanism has asymptotically no effect on parameter estimation. The latter is usually done by analyzing the set of likelihood ratios per the Wilks Test:

$$-2 \ln \left( \frac{\sup_{\theta \in \Theta} \frac{1}{N} \sum_{i=1}^{N} f(O(X(F_i), M_i)|\theta)}{\sup_{\theta \in \Theta} \frac{1}{N} \sum_{i=1}^{N} f(O(X(F_i), M_i)|\theta)} \right)$$

where $\theta, \theta_0 \subseteq K_\theta$. Thus, hypothesis testing is also done directly on the likelihood function. Since the missing data mechanism has asymptotically no effect on this function if the data are Missing at Random, it follows also that it has asymptotically no effect on hypothesis testing either.

Finally, when doing inference on large samples with missing data, Maximum Likelihood is preferable to Multiple Imputation, which some studies have recently hinted at (Shin, Davison, and Long 2017; von Hippel 2016). This is due to a few reasons. First, the average-loglikelihood function (ignoring the missingness mechanism) generated by large partial datasets is going to mirror that which would have been generated by the complete dataset. Multiple Imputation works by first obtaining the average-loglikelihood function generated by partial data, and then imputing multiple values for the missing entries based on this function. Multiple Imputation thus involves an extra, and perhaps harmful step. The average-loglikelihood function is already going to yield parameter estimates, uncertainty estimates, confidence intervals, etc. that are close to those that would have been obtained from complete data, so there is little need to impute missing values to “re-estimate” these quantities. Second, since these “re-estimates” are obtained from datasets with imputed values, one is at increased risk of obtaining biased estimates for the parameters and their corresponding uncertainties. By imputing values, one is introducing additional noise in their analysis of the parameters, which reduces the quality of the estimates. Indeed, the risk of Multiple Imputation leading to bias with respect to certain parameter estimates has been observed in other settings (Madley-Dowd et al. 2019; Mishra and Khare 2014; Jochen et al. 2013; Wu and Jia 2013; Black, Harel, and Betsy McCoach 2011; Knol et al. 2010; Leite and Beretvas 2010; Schafer 1997). One justification for the imputation of values is to account for estimation uncertainty (Honaker, King, and Blackwell 2011), but there is no need to do this, because Theorems 1 and 2 show that the uncertainty estimate that one obtains from a large enough partial dataset will already be close to the true uncertainty. Multiple imputation is superfluous in large samples and brings inferences farther from their true values. In the next section, this article concretely shows that
because of this resulting bias in Multiple Imputation, Maximum Likelihood inference will tend to outperform Multiple Imputation in large samples with respect to obtaining parameter estimates and corresponding uncertainties that are close to the truth.

3. Inference with missing data: simulation example

Following previous studies, this article now shows by way of simulation how sample size and the extent of missingness can affect the validity of inference from partial data (Madley-Dowd et al. 2019; Mishra and Khare 2014; Jochen et al. 2013; Wu and Jia 2013; Black, Harel, and Betsy McCoach 2011; Knol et al. 2010; Leite and Beretvas 2010; Schafer 1997). This simulation is situated in the theoretical context of a study by Lupu (2013) evaluating the impact of International Human Rights Law on public support for human rights abuses. It shows (1) that the maximum distance between the average-loglikelihood function (ignoring the missingness mechanism) generated by incomplete data and that which would have been generated by complete data decreases with respect to sample size and missingness, and (2) that for data that is Missing at Random, this approximation depends only on sample size. It further shows this convergence of the average-loglikelihood functions results in parameter estimates and corresponding uncertainties that are close to their true values (what they would have been without missingness), and much closer than those that are obtained with Multiple Imputation. Altogether, these results show that the theorems proven above do hold in practice.

The data used in this part of the article are purely simulated. However, to make better sense of the effects of MAR and Missing not at Random (MNAR) data on statistical inference, the simulated data are treated as if they came from the empirical survey study done by Lupu (2013). It is easier to make sense of MAR and MNAR data in this concrete way, because whether data end up being MAR or MNAR depends on the substance of what is being measured. For example, survey data on depression can be MAR if male participants are less likely to answer questions than female participants because of an aversion to showing vulnerability. In this example, the data are MAR because of some substantive reason about the differences in how males and females relate to mental health. It is therefore helpful to situate the simulated data in some empirical context like the Lupu (2013) study so that the MAR and MNAR mechanisms are better understood. This is precisely the approach taken in other studies on missing data: Mishra and Khare (2014) simulated data from a diabetic clinical trial, Jochen et al. (2013) simulated data from an existing dataset on parent-child relationships, Wu and Jia (2013) simulated data from an existing dataset on dietary habits, and Black, Harel, and Betsy McCoach (2011) simulated data from an existing dataset on socioeconomic status and student achievement (other studies that do this include Jakobsen et al. (2017), Sung and Geyer (2007), Tang and Qin (2012)). Situating data in some empirical example does not alter the simulated nature of the data in any way, it is merely a means of framing it. Thus, the overall simulation results do not depend on the empirical example used.

3.1. Theoretical context of the simulation

The Lupu and Wallace (2019) study analyzes whether international law affects public support for human rights abuses. Through a survey experiment across 3 countries,
Lupu (2013) estimates the average treatment effect of whether an individual would support their government if told that their government has violated international human rights law. The study found that in India, an individual would tend to support their government less upon hearing this information, whereas in Israel, an individual would tend to support their government more.

For simplicity, the dataset from this study as it is used in this article contains only on the India subsample. It includes information about the following for each individual: whether they were told that their country has violated international human rights law, gender, age, ideology, education level, whether they participated in political activism, their level of political interest, and their level of approval for their government.

### 3.2. Methodology

The main objects of interest in this article are the average-loglikelihood functions (ignoring the missingness mechanism) generated by partial and complete datasets. To study these objects, this article uses simulated normally distributed data. The main advantage of focusing on normally distributed data is that it allows for the convenient use of the Norm software (Schafer 2013) in R to evaluate likelihoods of partial observations. This is useful because obtaining likelihoods from partial data can be computationally complex (Little and Rubin 2002; Sung and Geyer 2007; Honaker, King, and Blackwell 2011). Although this article focuses on likelihood functions that are generated by normally distributed data, the Theorems above hold for any kind of data with a continuous probability distribution so long as the likelihood function corresponds to that distribution.

To simulate normally distributed data from the International Human Rights Law dataset, both the mean vector and variance-covariance matrix of the original dataset are calculated. These parameters are then used to simulate 500 observations according to a multivariate normal model, \( \mathcal{N}(\theta) = \mathcal{N}(\mu, \Sigma) \), that are treated as a perfectly representative sample of the population, from which simulated samples of the population can be drawn. Simulated samples from size \( N = 30 \) to \( N = 110 \) are taken in increments of 1. For each sample, a distribution of \( (\mu, \Sigma) \), \( K \), is obtained by bootstrapping: taking 100 “subsamples” of size \( N \) of each \( N \)-sized sample, and calculating \( (\mu, \Sigma) \) for each subsample. For each simulated sample, Missing at Random (MAR) data and Missing not at Random (MNAR) data are introduced in varying degrees. To introduce the former, it removes observations under the assumption that people with less interest in politics are less likely to answer the question about whether they approve of their government. To introduce the latter, it removes observations under the assumption that people who generally approve of the government tend not to indicate this on their surveys. In each case, missingness is introduced according to a Bernoulli distribution. The article looks at what happens when 20%, 30%, and 40% of a specific variable (the subject’s approval of their government) are MAR/MNAR. Then, the average-loglikelihood functions generated by the complete and incomplete datasets are approximated, using the Norm package to evaluate the average-loglikelihoods of partial data. These functions are evaluated over the space of parameters generated by bootstrapping, \( K \), and the supremum (which
is a maximum given that all of this is done on a compact set) of the differences between the two functions over \( K \) is evaluated.

The sample sizes considered for this part of the empirical analysis, namely the analysis of the average-loglikelihood functions, are relatively small; however, this is done intentionally. The average-loglikelihood functions considered here can yield values that are too small for \( R \) to return as non zero numbers. One could work with the loglikelihoods instead, but this can also be difficult because one would have to take the log of the difference of two average-loglikelihood functions, and the average-loglikelihood functions will still replace some small values with zero, leading to a difference of zero and thus producing \(-\infty\). The reason one would have to take the log of the difference of the functions rather than the difference of log of the two functions is that as sample size increases, the average-loglikelihoods tend to get really small; and the smaller they are, the greater the effect of a slight shift in the input (this is because the derivative of the log is \( \frac{1}{x} \)). Thus, differences between the two functions will tend to be exaggerated as the sample size increases, making the effects of sample size and the missingness mechanism harder to detect. To ensure that effects are correctly and precisely measured, the sample sizes are kept relatively small. This does not prove to be a big issue, since the effects of sample size and the extent of missingness are still clear with these small sample sizes.

Afterward, this article illustrates the effect that this convergence of the average-loglikelihood functions has on the Maximum Likelihood Estimates (MLEs) and their corresponding uncertainties. Using the likelihood functions generated by partial data, this article obtains the MLEs and corresponding uncertainties of the parameter, \( (\mu, \Sigma) \), and compares them to the estimates and corresponding uncertainties of these parameters that are obtained through Multiple Imputation by way of the Amelia software in R (Honaker, King, and Blackwell 2011). This is done by leveraging Amelia’s Bootstrap to obtain a distribution of estimates for the parameters \( (\mu, \Sigma) \) whose size is equal to the number of imputations, which is set to \( m = 200 \). This comparison is done over samples ranging from size \( N = 200 \) to \( N = 5000 \) under different Missing at Random mechanisms. The Missing at Random mechanisms are simulated in the same way as above for \( \psi = 0.1, 0.3, \) and \( 0.7 \).

### 3.3. Results

Figure 1 illustrates how the supremum of the differences between the average-loglikelihood functions (ignoring the missingness mechanism) generated by partial and complete data changes with respect to sample size and the extent of Missing at Random data. Figure 2 is analogous for Missing not at Random data. Figure 3 shows how the parameter estimates generated by the likelihood function approximate the true parameter as sample size increases under the Missing at Random assumption, and compares this approximation to that obtained by Multiple Imputation. Figure 4 shows this same phenomenon with respect to the uncertainty of the parameter estimates.

### 3.4. Discussion

Figures 1 and 2 show that as the sample size increases, the maximum distance between the average-loglikelihood functions (ignoring the missingness mechanism) generated by
partial and complete data tends to decrease. They also show that as this maximum distance tends to decreases as the extent of missingness decreases; however, this effect appears to be more pronounced in Missing not at Random data than in Missing at Random data. This is expected, since Theorem 2 proves that for Missing at Random data, only the sample size matters for convergence. It may also be interesting to note that the value of the supremum of the differences between average-loglikelihood
functions generated by partial and complete data tends to be the same given similar sample sizes and degrees of missingness. Overall, these results verify that as the sample size increases and the extent of missingness decreases, an average-loglikelihood function generated by incomplete data will approximate that which would have been generated by complete data.

Figures 3 and 4 show further that the inferences drawn from partial datasets under the Missing at Random assumption will tend to approximate the truth as sample size increases. The difference between the MLE generated by Maximum Likelihood on partial datasets and the true MLE decreases as sample size increases. The same is true for the MLE uncertainty. Moreover, in both cases these differences are smaller than the differences that would result from using Multiple Imputation instead of Maximum Likelihood under different sample size and missingness specifications. This suggests that not only is Likelihood good at approximating the true parameters and corresponding uncertainties, it also does a better job than Multiple Imputation

4. Conclusion

Missing data are often an inevitable and difficult part of scientific research. Researchers have tried to deal with this problem by doing inference on partial datasets. However, it is not always clear when inferences done on partial datasets will be asymptotically valid. This article fills in this knowledge gap by offering a new asymptotic theory about inference with missing data. It proves that as sample size increases and the extent of missingness decreases, the average-loglikelihood function (ignoring the missingness mechanism) generated by a partial dataset will approximate that generated by a complete dataset. Moreover, it shows that if the data are Missing at Random, this approximation depends only on sample size. This result adds to previous research by proving that all possible inferences that one can draw from a loglikelihood function generated by partial data will approximate their true values (their complete-data analogues), whereas previous research has only proved this to be true for the posterior modes (Little and Rubin 2002; Little, Rubin, and Zangeneh 2017; Seaman et al. 2013; Takai and Kano 2013; Nielsen 1997). This result has important practical implications: (1) the inferences obtained from partial datasets will approximate those that would have been obtained from complete ones for sufficiently large samples and small amounts of
missing data, (2) when the data are Missing at Random, only the sample size matters for the validity of these inferences, not the proportion of missing data, (3) if the data are Missing at Random, then the missingness mechanism will have no asymptotic effect on parameter estimation and hypothesis testing, and (4) Maximum Likelihood inference is preferable to Multiple Imputation in large samples.

Acknowledgments

This project has not been funded.

References

Black, A. C., O. Harel, and D. Betsy McCoach. 2011. Missing data techniques for multilevel data: Implications of model misspecification. *Journal of Applied Statistics* 38 (9):1845–65. doi:10.1080/02664763.2010.529882.

Bobb, J. F., D. O. Scharfstein, M. J. Daniels, F. S. Collins, and S. Kelada. 2011. Multiple imputation of missing phenotype data for QTL mapping. *Statistical Applications in Genetics and Molecular Biology* 10 (1):1–27. doi:10.2202/1544-6115.1676.

Folland, G. B. 2007. *Real analysis: Modern techniques and their applications*. 2nd ed. New York: Wiley.

George, S. V., G. Ambika, and R. Misra. 2015. Effect of data gaps on correlation dimension computed from light curves of variable stars. *Astrophysics and Space Science* 360 (1):5. doi:10.1007/s10509-015-2516-z.

Honaker, J., G. King, and M. Blackwell. 2011. Amelia II: A program for missing data. *Journal of Statistical Software* 45 (7):1–47. doi:10.18637/jss.v045.i07.

Jakobsen, J. C., C. Gluud, J. Weterslev, and P. Winkel. 2017. When and how should multiple imputation be used for handling missing data in randomised clinical trials — a practical guide with flowcharts. *BMC Medical Research Methodology* 17 (1):162. doi:10.1186/s12874-017-0442-1.

Jochen, H., H. Max, B. Tamara, and L. Wilfried. 2013. Multiple imputation of missing data: A simulation study on a binary response. *Open Journal of Statistics* 2013: 370–8.

King, G., J. Honaker, A. Joseph, and K. Scheve. 2001. Analyzing incomplete political science data: An alternative algorithm for multiple imputation. *American Political Science Review* 95 (1): 49–69. doi:10.1017/S0003055401000235.

Knol, M. J., K. J. M. Janssen, A. R. T. Donders, A. C. G. Egberts, E. R. Heerdink, D. E. Grobbee, K. G. M. Moons, and M. I. Geerlings. 2010. Unpredictable bias when using the missing indicator method or complete case analysis for missing confounder values: An empirical example. *Journal of Clinical Epidemiology* 63 (7):728–36. doi:10.1016/j.jclinepi.2009.08.028.

Leite, W., and S. Beretvas. 2010. The performance of multiple imputation for Likert-type items with missing data. *Journal of Modern Applied Statistical Methods* 9 (1):64–74. doi:10.22237/jmasm/1272686820.

Little, R. J. A., and D. B. Rubin. 2002. *Statistical analysis with missing data*, 2nd ed., Wiley-Interscience.

Little, R. J., D. B. Rubin, and S. Z. Zangeneh. 2017. Conditions for ignoring the missing-data mechanism in likelihood inferences for parameter subsets. *Journal of the American Statistical Association* 112 (517):314–20. doi:10.1080/01621459.2015.1136826.

Lupu, Y. 2013. Best evidence: The role of information in domestic judicial enforcement of international human rights agreements. *International Organization* 67 (3):469–503. doi:10.1017/S002081831300012X.

Lupu, Y., and G. P. R. Wallace. 2019. Violence, nonviolence, and the effects of international human rights law. *American Journal of Political Science* 63 (2):411–26. https://onlinelibrary.wiley.com/doi/pdf/10.1111/ajps.12416.
Madley-Dowd, P., R. Hughes, K. Tilling, and J. Heron. 2019. The proportion of missing data should not be used to guide decisions on multiple imputation. *Journal of Clinical Epidemiology* 110:63–73. doi:10.1016/j.jclinepi.2019.02.016.

Mishra, S., and D. Khare. 2014. On comparative performance of multiple imputation methods for moderate to large proportions of missing data in clinical trials: A simulation study. *Journal of Medical Statistics and Informatics* 2 (1):9. doi:10.7243/2053-7662-2-9.

Murray, J. S. 2018. Multiple imputation: A review of practical and theoretical findings. *Statistical Science* 33 (2):142–59. doi:10.1214/18-STS644.

Nielsen, S. F. 1997. Inference and missing data: Asymptotic results. *Scandinavian Journal of Statistics* 24 (2):261–74. doi:10.1111/1467-9469.00062.

Nielsen, S. F. 2007. Proper and improper multiple imputation. *International Statistical Review* 71 (3):593–607. https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1751-5823.2003.tb00214.x.

Nielsen, S. F. V., and M. Mørup. 2014. Non-negative tensor factorization with missing data for the modeling of gene expressions in the human brain. In 2014 IEEE International Workshop on Machine Learning for Signal Processing (MLSP), pages 1–6. ISSN: 2378-928X.

Schafer, J. L. 1997. *Analysis of incomplete multivariate data*. New York: Chapman and Hall/CRC.

Schafer, P. t R. b A. A. N. O. b J. L. 2013. norm: Analysis of multivariate normal datasets with missing values. Accessed on 22 February 2022. https://CRAN.R-project.org/package=norm

Seaman, S., J. Galati, D. Jackson, and J. Carlin. 2013. What is meant by “Missing at Random”? *Statistical Science* 28 (2):257–68. doi:10.1214/13-STS415.

Shin, T., M. L. Davison, and J. D. Long. 2017. Maximum likelihood versus multiple imputation for missing data in small longitudinal samples with nonnormality. *Psychological Methods* 22 (3):426–49. doi:10.1037/met0000094.

Sung, Y. J., and C. J. Geyer. 2007. Monte Carlo likelihood inference for missing data models. *The Annals of Statistics* 35 (3):990–1011. doi:10.1214/009053606000001389.

Takai, K., and Y. Kano. 2013. Asymptotic Inference with Incomplete Data. *Communications in Statistics - Theory and Methods* 42 (17):3174–90. doi:10.1080/03610926.2011.621577.

Tang, C. Y., and Y. Qin. 2012. An efficient empirical likelihood approach for estimating equations with missing data. *Biometrika* 99 (4):1001–7. doi:10.1093/biomet/ass045.

von Hippel, P. T. 2016. New confidence intervals and bias comparisons show that maximum likelihood can beat multiple imputation in small samples. *Structural Equation Modeling: A Multidisciplinary Journal* 23 (3):422–37. doi:10.1080/10705511.2015.1047931.

Wu, W., and F. Jia. 2013. A new procedure to test mediation with missing data through nonparametric bootstrapping and multiple imputation. *Multivariate Behavioral Research* 48 (5):663–91. doi:10.1080/00273171.2013.816235.

**Appendix**

**Proof of Lemma 1**

**Lemma 1.** \( \lim_{k \to \infty} \mathbb{E}_{M \in \{0,1\}^m}[f(O(X,M)|\theta)|X,\psi_k] = f(X|\theta) \) almost uniformly on \( K_X \times K_\theta \).

**Proof.** Let \( \theta \in K_\theta, X \in K_X \).

\[
\lim_{k \to \infty} \mathbb{E}_{M \in \{0,1\}^m}[f(O(X,M)|\theta)|X,\psi_k] = \lim_{k \to \infty} \int_{\{0,1\}^m} f(O(X,M)|\theta)dP_k
\]

Observe that \( f(O(X,M)|\theta) \) is bounded and Lipschitz over \( \{0,1\}^m \). Observe also that \( \forall M \in \{0,1\}^m, \limsup_{k \to \infty} P_k(M|X,\psi_k) \leq 1_{\{0\}^m}(M) \) by construction of \( \{P_k\}_{k=1}^\infty \). Therefore, by the Portmanteau Theorem,
\[
\lim_{k \to \infty} \int_{(0,1)^m} f(O(X,M)|\theta) dP_k \\
= \int_{(0,1)^m} f(O(X,M)|\theta) d1_{\{1\}^m} \\
= f(X|\theta) \\
\Rightarrow \lim_{k \to \infty} E_{M \in \{0,1\}^m} [f(O(X,M)|\theta)|X,\psi_k] = f(X|\theta) \text{ for a given } X, \theta
\]  

(A1)

Since \(K_X \times K_\theta\) is compact, it is of finite measure. Therefore, by Egorov’s Theorem, this convergence happens almost uniformly on \(K_X \times K_\theta\):

\[
E_{M \in \{0,1\}^m} [f(O(X,M)|\theta)|X,\psi_k] \text{ converges almost uniformly on } K_X \times K_\theta \text{ to } f(X|\theta)
\]

\(\square\)

**Proof of Lemma 2**

**Lemma 2.** There exists a sequence of simple functions \(\{\phi_n\}_{n=1}^\infty\) of the form \(\sum_{j=1}^n z_j 1_{E_j}\) such that \(\lim_{n \to \infty} \phi_n = X\) uniformly. Moreover,

1. \(f(O(X,M)|\theta) = f(O(\phi_n,M)|\theta) + \varepsilon(n)\) uniformly over \(K_X \times K_\theta\) for all \(M \in \{0,1\}^m\), and
2. \(f(X|\theta) = f(\phi_n|\theta) + \varepsilon(n)\) uniformly over \(K_X \times K_\theta\).

**Proof.** Let \(M \in \{0,1\}^m\). \(f(O(X,M)|\theta)\) is continuous on \(K_X \times K_\theta\), which is a compact set. So, \(f(O(X,M)|\theta)\) is uniformly continuous on \(K_X \times K_\theta\). This means that \(\forall \varepsilon > 0, \exists \delta(M,\varepsilon) > 0 : \forall X_1, X_2 \in K_X : ||X_1 - X_2|| < \delta(M,\varepsilon), \frac{|f(O(X_1,M)|\theta) - f(O(X_2,M)|\theta)|}{|} < \varepsilon\).

By Theorem 2.10 in Folland (2007), \(\exists \{\phi_n\}_{n=1}^\infty\), a sequence of simple functions of the form \(\sum_{j=1}^n z_j 1_{E_j}\) such that \(\lim_{n \to \infty} \phi_n = X\) uniformly. This implies that \(\forall \varepsilon > 0, \exists N \in \mathbb{N} : \forall n \geq N, \sup_{F \in X^{-1}(K_X)} |X(F) - \phi_n(F)| < \varepsilon\). Here \(X^{-1}(K_X)\) can be thought of as the space of outcomes that, when input as arguments into \(X\), return values that span \(K_X\).

So, if \(\varepsilon = \delta(M,\varepsilon), \exists N_1 \in \mathbb{N} : \forall n \geq N_1, \sup_{F \in X^{-1}(K_X)} |X(F) - \phi_n(F)| < \delta(M,\varepsilon)\) \(\Rightarrow |f(O(X,M)|\theta) - f(O(\phi_n,M)|\theta)| < \varepsilon\) \(\Rightarrow f(O(X,M)|\theta) = f(O(\phi_n,M)|\theta) + \varepsilon(n)\) \(\text{ uniformly over } K_X \times K_\theta\) for the given \(M \in \{0,1\}^m\). Since \(M\) can only adopt finitely many values in \(\{0,1\}^m, \forall \varepsilon > 0 \exists \delta = \min_{M \in \{0,1\}^m} \{\delta(M,\varepsilon)\}\), and hence a corresponding \(N_1\) analogous to \(N_1\) in Equations (A2) and (A3) such that:

\[
f(O(X,M)|\theta) = f(O(\phi_n,M)|\theta) + \varepsilon(n)
\]

(A4)

uniformly over \(K_X \times K_\theta\) for all \(M \in \{0,1\}^m\). Letting \(M = \{1\}^m\) shows that:

\[
f(X|\theta) = f(\phi_n|\theta) + \varepsilon(n)
\]

(A5)

uniformly over \(K_X \times K_\theta\).  

\(\square\)
Proof of Theorem 1

Theorem 1.

\[
\lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} f(O(X(F_i), M_i)|\theta) = \lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} f(X(F_i)|\theta) + \frac{1}{N} \sum_{i=1}^{N} \varepsilon_i(k) + \frac{2}{N} \sum_{i=1}^{N} \varepsilon_i(n)
\]

in probability on \( K_X \times K_\theta \), where \( \{\psi_k\}_{k \in \mathbb{N}} \) is a sequence of parameters generating the \( \{M_i\}_{i \in \mathbb{N}} \), and \( \{\phi_n\}_{n=1}^\infty \) is a sequence of simple functions, so \( \lim_{n \to \infty} \phi_n = X \) uniformly over \( K_\theta \).

Proof. Begin by noting that, uniformly over \( K_\theta \),

\[
\sum_{i=1}^{N} f(O(X(F_i), M_i)|\theta)
\]

= \( \sum_{i=1}^{N} \left[ f(O(\phi_n(F_i), M_i)|\theta) + \varepsilon_i(n) \right] \) by Lemma 2.

Now, note that given a sequence of simple functions, \( \{\phi_n\}_{n=1}^\infty \) of the form \( \sum_{j=1}^{n} z_j 1_{E_j} \), such that \( \lim_{n \to \infty} \phi_n = X \) uniformly,

\[
f(O(\phi_n, M)|\theta) = \sum_{j=1}^{n} f(O(z_j, M)|\theta) 1_{E_j}
\]

(A6)

for all \( M \in \{0, 1\}^m \), \( \theta \in K_\theta \).

Thus, uniformly over \( K_\theta \),

\[
\sum_{i=1}^{N} \left[ f(O(\phi_n(F_i), M_i)|\theta) + \varepsilon_i(n) \right] = \sum_{i=1}^{N} \sum_{j=1}^{n} \left[ f(O(z_j, M_i)|\theta) 1_{E_j(F_i)} + \varepsilon_i(n) 1_{E_j(F_i)} \right] \] by Equation 6

Now, note that the \( \{E_j\}_{j=1}^{n} \) are mutually exclusive and collectively exhaustive of \( X^{-1}(K_X) \) by properties of simple functions, so \( \forall i \in \{1, ..., N\}, \exists j(i) \in \{1, ..., n\} : F_i \in E_{j(i)} \). This implies that \( \forall j \in \{1, ..., n\}, \exists I(j) \subseteq \{1, ..., N\} : \forall i \in I(j), F_i \in E_j \), and \( I(j_1) \cap I(j_2) \neq \emptyset \) if and only if \( j_1 = j_2 \). This is a formal way of saying that every outcome in \( \{F_i\}_{i=1}^{N} \) is associated with some \( z_j \) in the construction of \( \phi_n \), and vice versa. Thus, uniformly over \( K_\theta \),

\[
\sum_{i=1}^{N} \sum_{j=1}^{n} \left[ f(O(z_j, M_i)|\theta) 1_{E_j(F_i)} + \varepsilon_i(n) 1_{E_j(F_i)} \right] = \sum_{j=1}^{n} \sum_{I \subseteq \{j\}} \left[ f(O(z_j, M_i)|\theta) + \varepsilon_i(n) \right]
\]

By the Strong Law of Large Numbers and Lemma 1, for any \( j \in \{1, ..., n\} \),

\[
\frac{1}{|I(j)|} \sum_{I \subseteq \{j\}} f(O(z_j, M_i)|\theta) = E_{M \in \{0, 1\}^n} \left[ f(O(\phi_n, M)|\theta) z_j, \psi_k \right] + \varepsilon_j(|I(j)|, \theta)
\]

is the largest among \( |I(j)| \) almost uniformly over \( K_X \times K_\theta \).

Further, note that \( \forall j \in \{1, ..., n\}, \varepsilon > 0, \exists N_\varepsilon(j) \in \mathbb{N} : \forall N \geq N_\varepsilon(j), |\varepsilon_j(|I(j)|)| < \varepsilon \), because \( P(F_i) > 0 \forall i \in \{1, ..., N\} \), so \( \forall \varepsilon > 0, \exists N_\varepsilon = \max_{j \in \{1, ..., n\}} N_\varepsilon(j) : \forall N \geq N_\varepsilon, |\varepsilon_j(|I(j)|)| < \varepsilon \). This is a formal way of saying that for any \( j \in \{1, ..., n\} \), the corresponding set of outcomes \( I(j) \) that, when input as arguments into the random variable \( X \), would produce \( z_j \), depends on \( N \). This is because each outcome has non zero probability, so increasing \( N \) (the sample size), and then drawing random samples will lead to a larger \( |I(j)| \) for all \( j \). This means that the error term, \( \varepsilon_j(|I(j)|) \) becomes arbitrarily small as the sample size, \( N \), increases. Formally,
If the data are Missing at Random, then

\[
\frac{1}{|I(j)|} \sum_{i \in I(j)} f(O(z_i, M_i)|\theta) = f(z_j|\theta) + e_j(k) + e_j(N)
\]

almost uniformly on \( K_X \times K_\theta \).

Thus, the following convergence happens in probability on \( K_X \times K_\theta \):

\[
\begin{align*}
&= \sum_{j=1}^{n} \sum_{i \in I(j)} \left[ f(O(z_i, M_i)|\theta) + e_i(n) \right] \\
&= \sum_{j=1}^{n} |I(j)| \left[ f(z_j|\theta) + e_j(k) + e_j(N) \right] + \sum_{i=1}^{N} e_i(n) \text{ by Equation 7} \\
&= \sum_{j=1}^{n} \left[ |I(j)| f(z_j|\theta) + |I(j)| e_j(k) + |I(j)| e_j(N) \right] + \sum_{i=1}^{N} e_i(n) \\
&= \sum_{i=1}^{N} \sum_{j=1}^{n} f(z_j|\theta) 1_{E_i}(F_i) + \sum_{i=1}^{N} \sum_{j=1}^{n} e_j(k) 1_{E_i}(F_i) + \sum_{i=1}^{N} \sum_{j=1}^{n} e_j(N) 1_{E_i}(F_i) + \sum_{i=1}^{N} e_i(n) \\
&= \sum_{i=1}^{N} f(\psi_n(F_i)|\theta) + \sum_{i=1}^{N} e_i(k) + \sum_{i=1}^{N} e_i(N) + \sum_{i=1}^{N} e_i(n) \text{ by Lemma 2} \\
&= \sum_{i=1}^{N} f(X(F_i)|\theta) + \sum_{i=1}^{N} e_i(k) + \sum_{i=1}^{N} e_i(N) + 2\sum_{i=1}^{N} e_i(n)
\end{align*}
\]

This implies that:

\[
\lim_{k \to \infty, N \to \infty} \frac{1}{N} \sum_{i=1}^{N} f(O(X(F_i), M_i)|\theta)
\]

\[
= \frac{1}{N} \left\{ \sum_{i=1}^{N} f(X(F_i)|\theta) + \sum_{i=1}^{N} e_i(k) + \sum_{i=1}^{N} e_i(N) + 2\sum_{i=1}^{N} e_i(n) \right\}
\]

\[
= \frac{1}{N} \sum_{i=1}^{N} f(X(F_i)|\theta) + \frac{1}{N} \sum_{i=1}^{N} e_i(k) + \frac{1}{N} \sum_{i=1}^{N} e_i(N) + \frac{2}{N} \sum_{i=1}^{N} e_i(n)
\]

in probability on \( K_X \times K_\theta \), where \( \lim_{k \to \infty} \sup_{n \in \mathbb{N}} e_i(k) = \lim_{N \to \infty} \sup_{n \in \mathbb{N}} e_i(N) = 0 \). Thus,

\[
= \lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} f(X(F_i)|\theta) + \frac{1}{N} \sum_{i=1}^{N} e_i(k) + \frac{2}{N} \sum_{i=1}^{N} e_i(n)
\]

in probability on \( K_X \times K_\theta \), where \( \{\psi_k\}_{k \in \mathbb{N}} \) is a sequence of parameters generating the \( \{M_i\}_{i \in \mathbb{N}} \), and \( \{\phi_n\}_{n=1}^{\infty} \) is a sequence of simple functions such that \( \lim_{n \to \infty} \phi_n = X \) uniformly over \( K_\theta \).

\[\square\]

**Proof of Theorem 2**

**Theorem 2.** If the data are Missing at Random, then

\[
\lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} f(O(X(F_i), M_i)|\theta) = \lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} f(X(F_i)|\theta) + \frac{2}{N} \sum_{i=1}^{N} e_i(n)
\]
Proof. Let $\phi_n$ be a simple function approximation of $X$ as discussed in Theorem 1. It is sufficient to show that:

$$\mathbb{E}_{K_X}[v(k)] = 0$$

since as $N$ approaches infinity, $\frac{1}{N}\sum_{i=1}^{N} v_i(k) \approx \mathbb{E}_{K_X}[v(k)]$ by the Strong Law of Large Numbers. Note that

$$\mathbb{E}_{K_X}[v(k)] = \mathbb{E}_{K_X}\left[\mathbb{E}_{M \in \{0,1\}^n}[f(O(\phi_n,M)|\theta)] - f(\phi_n|\theta)\right]$$

Thus, it is sufficient to show that

$$\mathbb{E}_{K_X}\left[\mathbb{E}_{M \in \{0,1\}^n}[f(O(\phi_n,M)|\theta)]\right] = \mathbb{E}_{K_X}[f(\phi_n|\theta)]$$

Observe that

$$\mathbb{E}_{K_X}\left[\mathbb{E}_{M \in \{0,1\}^n}[f(O(\phi_n,M)|\theta)]\right] = \sum_{j=1}^{n} \sum_{M \in \{0,1\}^n} f(O(z_j,M)P_k(M|O(z_j,M),\psi_k)P(E_j)$$

$$= \sum_{M \in \{0,1\}^n} \sum_{j=1}^{n} f(O(z_j,M)P_k(M|O(z_j,M),\psi_k)P(E_j)$$

Because the data are Missing at Random, note that $\forall M \in \{0,1\}^n$, there is a set of subsets of $\{1,...,n\}$, denoted $C(M) = \{C(M)_1,\ldots,C(M)_{n(M)}\}$ that are mutually exclusive and collectively exhaustive of $\{1,...,n\}$ such that $O(z_j,M) = O(z_j,M)$ if and only if $\exists q \in \{1,...,n(M)\}$ such that $z_j \in C(M)_q$. This is a formal way of saying that for any given missingness vector, one can partition the space of $\{E_1,...,E_n\}$ into clusters of $E_j$'s that would have the same observed values under the missingness vector given. For any $M \in \{0,1\}^n$, $q \in \{1,...,n(M)\}$, let the “representative” of $C(M)_q$ be $z(M,q) = O(z_j,M)$, where $j \in C(M)_q$. Thus,

$$\sum_{M \in \{0,1\}^n} \sum_{j=1}^{n} f(O(z_j,M)P_k(M|O(z_j,M),\psi_k)P(E_j)$$

$$= \sum_{M \in \{0,1\}^n} \sum_{q=1}^{n(M)} \left[ \sum_{j \in C(M)_q} \left[ \ldots \left[ \ldots \left[ \sum_{j \in C(M)_q} \ldots \right] \right] \right] \right]$$

$$= \sum_{M \in \{0,1\}^n} \sum_{q=1}^{n(M)} \left[ \sum_{j \in C(M)_q} \ldots \left[ \sum_{j \in C(M)_q} \ldots \right] \right]$$

since $f(O(X,M))$ is obtained by “integrating out” the missing values by taking a conditional expectation.
by the fact that the data are Missing at Random

\[
\sum_{M \in \{0,1\}^n} \sum_{q=1}^{n(M)} \left[ \sum_{j \in C(M)_q} \mathbb{E}_{K_x} \left[ f(\phi_n | \theta) | O(\phi_n, M) = z(M, q) \right] P_k(M | O(z_j, M), \psi_k) P(E_j) \right] \\
= \sum_{M \in \{0,1\}^n} \sum_{q=1}^{n(M)} \left[ \sum_{j \in C(M)_q} \mathbb{E}_{K_x} \left[ f(\phi_n | \theta) | j \in C(M)_q, M \right] P_k(M | O(z_j, M), \psi_k) P(E_j) \right] \\
= \sum_{M \in \{0,1\}^n} \sum_{q=1}^{n(M)} \mathbb{E}_{K_x} \left[ \sum_{j \in C(M)_q} f(\phi_n | \theta) | j \in C(M)_q, M \right] P_k(M | O(z_j, M), \psi_k) \sum_{j \in C(M)_q} P(E_j) \\
= \mathbb{E}_{K_x} \left[ f(\phi_n | \theta) \right] 
\]

\[\square\]