stats

MDPI

Article

Model Selection with Missing Data Embedded in Missing-at-Random Data

Keiji Takai [1,*] and Kenichi Hayashi [2]

[1] Faculty of Commerce, Kansai University, Yamatecho 3-3-35, Osaka 564-8680, Japan
[2] Department of Mathematics, Keio University, Hiyoshi 3-14-1, Kohokuku, Yokohama 223-0061, Japan
* Correspondence: takai@kansai-u.ac.jp

Abstract: When models are built with missing data, an information criterion is needed to select the best model among the various candidates. Using a conventional information criterion for missing data may lead to the selection of the wrong model when data are not missing at random. Conventional information criteria implicitly assume that any subset of missing-at-random data is also missing at random, and thus the maximum likelihood estimator is assumed to be consistent; that is, it is assumed that the estimator will converge to the true value. However, this assumption may not be practical. In this paper, we develop an information criterion that works even for not-missing-at-random data, so long as the largest missing data set is missing at random. Simulations are performed to show the superiority of the proposed information criterion over conventional criteria.

Keywords: information criteria; missing at random; missing data; not missing at random

Citation: Takai, K.; Hayashi, K. Model Selection with Missing Data Embedded in Missing-at-Random Data. *Stats* **2023**, *6*, 495–505. https://doi.org/10.3390/stats6020031

Academic Editors: Hari Mohan Srivastava and Wei Zhu

Received: 8 March 2023
Revised: 31 March 2023
Accepted: 7 April 2023
Published: 11 April 2023

1. Introduction

In data analysis, all variables are not equally important in the final model; some variables are more important than others. Thus, what matters in practice is model selection. For model selection, several variants of the Kullback-Leibler (KL) information criteria have been developed. In the case of complete data, the criteria include Akaike's information criterion (AIC; [1]) and Takeuchi's information criterion (TIC; [2]). These all measure the "distance" from the true distribution to the model distribution in the sense of Kullback-Leibler information and find the model that minimizes the information.

Applying these information criteria to real data is not straightforward, as nearly all real data have missing values. In practical cases, missing values are an obstacle to applying statistical methods since the methods implicitly assume the availability of complete data. To handle such missing values with information criteria, modifications have been proposed by authors such as [3–7]. The first modification of the information criterion is attributed to [3]. The paper [4] derived the information criterion for missing data in a straight way, and the paper [5] used the symmetric KL divergence for the derivation. The paper [6] used the Hermite expansion with the Markov Chanin Monte Carlo method to approximate the H-function of the EM algorithm, and hence it needed much computation in application. The paper [7] also used the EM algorithm to develop the information criterion for the settings with missing covariates.

However, despite modification, these information criteria still face a major problem in their practical application. The common problem is their implicit or explicit assumption that any set of data selected by them produces consistent parameters in the model. This assumption is simply not realistic. In reality, these criteria can exclude a variable that causes missingness, resulting in inconsistent parameters, as the missing data are not missing at random (NMAR; [8]). For the parameters to be consistent, the missing-data mechanism requires modeling [9]. However, this is practically difficult since the data necessary for modeling are missing.

To overcome the problems described above, we propose a new information criterion for missing data. The primary advantage of our information criterion is that it can handle situations where a subset of the data is NMAR so long as the largest set of missing data is missing at random (MAR). In computing our criterion, we compute the parameters with the largest data set and extract the parameters necessary for the model under consideration. These two steps allow us to obtain an estimator that is consistent with the true value for any possible model. Since the existing criteria have no such feature, they are unable to produce a proper value for the information criterion when the data are NMAR.

The remainder of the paper is organized as follows. In the next section, we develop the asymptotic theory for a subset of MAR data. In Section 3, we derive our information criterion for a subset of MAR data. In Section 4, the results of two simulations are presented to demonstrate the effectiveness of the proposed information criterion in the practical situations. We conclude the paper with suggestions for future research.

2. Inference of Maximum Likelihood Estimator with Missing Data

We will represent the variables of the entire data set as a vector $\boldsymbol{x}$ and partition it to $(\boldsymbol{y}, \boldsymbol{z})$, where $\boldsymbol{y}$ is the variable of interest and $\boldsymbol{z}$ is not. The model density functions with respect to these variables are written as $f(\cdot)$ in common; however, they can be distinguished by the context. If the values of $\boldsymbol{x}$ are observed completely, we would denote this as X_{c} in matrix form. However, since $\boldsymbol{x}$ contains missing values, X_{c} consists of an observed part, X_{o}, and a missing part, X_{m}. For Y_{c}, Y_{o} and Y_{m} are defined in the same way. We denote the parameters of the distributions of $\boldsymbol{x}$ and $\boldsymbol{y}$ as $\boldsymbol{\theta}$ and $\boldsymbol{\alpha}$, respectively. The relationship between $\boldsymbol{\theta}$ and $\boldsymbol{\alpha}$ is assumed to be $\boldsymbol{\alpha} = S\boldsymbol{\theta}$, where S is a matrix to extract the necessary part of the parameter $\boldsymbol{\theta}$, and $\boldsymbol{\theta} = (\boldsymbol{\alpha}', \boldsymbol{\beta}')'$. The rest of $\boldsymbol{\theta}$ is written as $\boldsymbol{\beta}$. Note that although we could theoretically develop $\boldsymbol{\alpha} = g(\boldsymbol{\theta})$ for a nonlinear function, in this paper we confine our focus to the linear function that can be expressed as $\boldsymbol{\alpha} = S\boldsymbol{\theta}$ for simplicity.

In real data analysis, we often encounter a situation in which a subset of data is embedded in MAR data. With the MAR mechanism, the missingness of data is caused by observed values for variables included in the analysis; on the other hand, for NMAR data (i.e., data that are not MAR), the missingness is caused by variables not included in the analysis and/or by missing values of the included variables [8]. It follows that using more variables in the analysis makes it more likely that the missing data will be MAR (or close to MAR). One reason for this is that variables added into the analysis are likely to include those variables causing missingness. Another reason is that more variables are likely to include some variables related to missing values, recovering information lost by the missingness. These phenomena have been partly confirmed in simulations by [10–12]. Thus, MAR is a more plausible missing-data mechanism for data with a larger set of variables.

An example is given in which a subset of the missing data is NMAR but the entire data is MAR [9]. An investigation is conducted with an aim to estimate the correlation between the self-esteem and the sexual behavior for the teenagers. However, the question of the sexual behavior is so sensitive that the question is only asked for over 15. The resulting data are missing at random (MAR), since the question on the sexual behavior is missing for those less than 15 years of age. As described before, our interest is the relationship between the self-esteem and the sexual behavior for teenagers. If the variables for analysis are limited to the self-esteem and the sexual behavior, the data are NMAR since the data set excludes "age" which causes missingness, and thus the correlation estimate is biased. On the other hand, if the age is used as an additional variable, the data become MAR, and the likelihood estimation using the data on these three variables creates the consistent maximum likelihood estimator of the correlation coefficient between the self-esteem and the sexual behavior.

For missing data, an exact distribution such as the t-distribution for complete data is rarely obtained other than in special cases such as monotonic missing data [13,14]. Hence, asymptotic theory is preferable for deriving the distribution of an estimator with missing

data. In asymptotic theory, the most important properties of an estimator are consistency to the true value and asymptotic normality. Of these two properties, the former is more important in practice. That an estimator has consistency means that the estimated parameter converges to the true value of the assumed model and that the estimated parameter is close enough to the true value in a large sample. For an analyst, this is a highly desirable property. In fact, an estimator without such consistency is essentially useless to the analyst as a means of approximating the parameter that he/she wishes to estimate. However, even if an estimator lacks consistency, asymptotic normality still might hold [15]. Notably, it has been shown that a maximum likelihood (ML) estimator under MAR is a good estimator, since it will have both consistency and asymptotic normality [16].

We can now provide more detail regarding the estimation methods when a subset of MAR data is used to produce the ML estimator. It has been shown that the ML estimator based on the maximum likelihood method has consistency [16]. Expressing this in the form of an equation, we have

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \frac{1}{n} \log f(X_{\mathrm{o}}|\boldsymbol{\theta}) \xrightarrow{\text{a.s.}} \boldsymbol{\theta}_0 \stackrel{\text{def}}{=} \arg\max_{\boldsymbol{\theta}} E[\log f(\boldsymbol{x}|\boldsymbol{\theta})], \tag{1}$$

where n is the sample size, and the arrow indicates that the left-side term converges almost surely to the right-side term as n increases to infinity. What matters here is that the ML estimator is based on missing data X_{o}, and it is a function of the data and the corresponding missing-data indicator. This ML estimator converges to the value which maximizes $E[\log f(\boldsymbol{x}|\boldsymbol{\theta})]$ (where the expectation is taken with respect to $\boldsymbol{x}$), the same convergence point as the maximizer of $n^{-1} \log f(X_{\mathrm{c}}|\boldsymbol{\theta})$, which, in actuality, is not available under the presence of missing data. The asymptotic distribution of $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ is normal, with mean zero and variance $V(\boldsymbol{\theta}) = (E[-\nabla^2 \log f(X_{\mathrm{o}}|\boldsymbol{\theta})])^{-1}$, where ∇ is the first-derivative operator with respect to the parameter $\boldsymbol{\theta}$. On the other hand, the ML estimator

$$\tilde{\boldsymbol{\alpha}} = \arg\max_{\boldsymbol{\alpha}} \frac{1}{n} \log f(Y_{\mathrm{o}}|\boldsymbol{\alpha}) \tag{2}$$

would generally not be guaranteed to converge to the true value, $\boldsymbol{\alpha}_0$ in $\boldsymbol{\theta}_0$, since the set of data corresponding to $\boldsymbol{\alpha}$ might be NMAR. Hence, to estimate $\boldsymbol{\alpha}$, we construct an estimator by extracting the part corresponding to $\boldsymbol{\alpha}$ from the ML estimator $\hat{\boldsymbol{\theta}}$, and set $\hat{\boldsymbol{\alpha}} = S\hat{\boldsymbol{\theta}}$.

3. Information Criterion for Missing Data

In this section, we propose an information criterion (IC) for missing data that are MAR as a whole but might be NMAR as a subset. In selecting a model among multiple candidate models, the standard IC underlies the discrepancy from the true distribution $g(\boldsymbol{y})$ to the model $f(\boldsymbol{y}|\boldsymbol{\theta})$. This is measured using the KL information, which is defined as

$$E_{g(\boldsymbol{y})}\left[\log \frac{f(\boldsymbol{y}|\boldsymbol{\alpha})}{g(\boldsymbol{y})}\right] = E_{g(\boldsymbol{y})}[\log f(\boldsymbol{y}|\boldsymbol{\alpha})] - E_{g(\boldsymbol{y})}[\log g(\boldsymbol{y})]$$

Since the second term is constant, the first term determines the KL information. The first term works as a measure of a model's goodness. When we estimate the parameter $\boldsymbol{\alpha}$ using the ML method and obtain its estimate $\tilde{\boldsymbol{\alpha}}$, a naive measure of goodness of fit is $E_{g(\boldsymbol{y})}[\log f(\boldsymbol{y}|\tilde{\boldsymbol{\alpha}})]$. The problem is that the true distribution g is unknown. If the complete data are available and are a random sample, we can use $n^{-1} \sum_{j=1}^{n} \log f(\boldsymbol{y}_j|\tilde{\boldsymbol{\alpha}})$ as an estimated measure to the true distribution, where $\boldsymbol{y}_j$ $(j = 1, \ldots, n)$ are the data used to estimate $\tilde{\boldsymbol{\alpha}}$. However, the problem with using thisempirical likelihood is that the estimates depend on the data, and hence $n^{-1} \sum_{j=1}^{n} \log f(\boldsymbol{y}_j|\tilde{\boldsymbol{\alpha}})$ is anasymptotically biased estimator of $E_{g(\boldsymbol{y})}[\log f(\boldsymbol{y}|\tilde{\boldsymbol{\alpha}})]$. The likelihood with the corrected bias is known as the IC. Some ICs have a special name, such as AIC [1], TIC [2] and GIC [17]. In the correction and derivation of the IC, asymptotic properties play a central role. Of the various desirable properties of an ML estimator, the most important are asymptotic normality and the consistency of the

ML estimator with respect to the true value. See also Konishi and Kitagawa [17] for the technical details.

For missing data, an IC can be computed in a similar spirit. However, considerable caution must be exercised in constructing the criterion, as different subsets of the missing data may have different types of missing-data mechanisms. Even if a subset of the missing data, call it Y_o, is MAR, which would mean that a naive ML estimator based on $\log f(Y_o|\boldsymbol{\alpha})$ would have consistency, a smaller subset of Y_o might be NMAR because the variables causing missingness are excluded from the model under consideration, in which case the ML would not have consistency. This means that a straightforward extension of the conventional IC could not be effectively applied to such missing-data cases.

To derive the proposed IC for cases with missing data, we will assume that the data as a whole, represented as X_o, are MAR. As shown in [16], the ML estimator based on X_o has consistency and asymptotic normality. However, a subset of MAR data X_o, call it Y_o, might be NMAR, and thus the ML estimator based on Y_o would be asymptotically biased. To ensure that the ML estimator is consistent (i.e., converges to the true value), it is necessary to perform ML estimation with the entire data set, even when we are interested in models for a subset of X_o, Y_o. Let the ML estimator based on the entire data set X_o be $\hat{\boldsymbol{\theta}}$, and let the part of $\boldsymbol{\theta}$ associated with the model of interest be $\boldsymbol{\alpha}$. In this normality case, there is a selection matrix S such that $S\boldsymbol{\theta} = \boldsymbol{\alpha}$. The ML estimates of $\boldsymbol{\alpha}$ can be written as $\hat{\boldsymbol{\alpha}} = S\hat{\boldsymbol{\theta}}$. In this setting, we can evaluate the bias.

We will use $\log f(Y_o|\boldsymbol{\theta})$ to denote a log-likelihood function. With the regularity conditions and the MAR assumption, the ML estimator is proved to have consistency and asymptotic normality [16].

We can now evaluate the bias of $\log f(Y_o|\hat{\boldsymbol{\alpha}})$. The bias is given as

$$b(G) = E\Big[\log f(Y_o|\hat{\boldsymbol{\alpha}}) - nE_{g(\boldsymbol{y})}[\log f(\boldsymbol{y}|\hat{\boldsymbol{\alpha}})]\Big].$$

Decomposing the bias, we have

$$\begin{aligned} b(G) &= E[\log f(Y_o|\hat{\boldsymbol{\alpha}}) - \log f(Y_o|\boldsymbol{\alpha}_0)] \\ &\quad + E\Big[\log f(Y_o|\boldsymbol{\alpha}_0) - nE_{g(\boldsymbol{y})}[\log f(\boldsymbol{y}|\boldsymbol{\alpha}_0)]\Big] \\ &\quad + E\Big[nE_{g(\boldsymbol{y})}[\log f(\boldsymbol{y}|\boldsymbol{\alpha}_0)] - nE_{g(\boldsymbol{y})}[\log f(\boldsymbol{y}|\hat{\boldsymbol{\alpha}})]\Big]. \end{aligned}$$

Let the terms be, in order, D_1, D_2, and D_3.

In this paper, we have assumed that X_o is MAR and that $\hat{\boldsymbol{\theta}}$ is asymptotically normally distributed. First, we evaluate D_1. ∇ and $\nabla_{\boldsymbol{\alpha}}$ indicate the first derivative operator with respect to $\boldsymbol{\theta}$, and to $\boldsymbol{\alpha}$, respectively. Assume that the ML estimator $\hat{\boldsymbol{\theta}}$ takes the form

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = V_{X_o}(\boldsymbol{\theta}_0)\sqrt{n}\frac{1}{n}\nabla \log f(X_o|\boldsymbol{\theta}_0) + o_p(1), \tag{3}$$

where $\boldsymbol{\theta}_0$ is the true value of the parameter, and $V_{X_o}(\boldsymbol{\theta}_0)$ is the variance defined as $V_{X_o}(\boldsymbol{\theta}_0) = E[\nabla \log f(X_o|\boldsymbol{\theta}_0)\nabla \log f(X_o|\boldsymbol{\theta}_0)']^{-1}$. In addition, assume that $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ is asymptotically distributed as normal with mean zero and variance $V_{X_o}(\boldsymbol{\theta}_0)$. By $\sqrt{n}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0) = S\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$, $-\nabla^2_{\boldsymbol{\alpha}} \log f(Y_o|\boldsymbol{\alpha}_0) \approx V_{Y_o}(\boldsymbol{\theta}_0)^{-1}/n$, we obtain

$$\begin{aligned} &\log f(Y_o|\hat{\boldsymbol{\alpha}}) - \log f(Y_o|\boldsymbol{\alpha}_0) \\ &= \nabla_{\boldsymbol{\alpha}} \log f(Y_o|\boldsymbol{\alpha}_0)'(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0) + \frac{1}{2}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0)'\nabla^2_{\boldsymbol{\alpha}} \log f(Y_o|\boldsymbol{\alpha}_0)(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0) + o_p(1) \\ &= \left(\sqrt{n}\frac{1}{n}\nabla \log f(Y_o|\boldsymbol{\alpha}_0)\right)' S'SV_{X_o}(\boldsymbol{\theta}_0)\sqrt{n}\frac{1}{n}\nabla \log f(X_o|\boldsymbol{\theta}_0) \\ &\quad - \frac{1}{2}\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)'S'(V_{Y_o}(\boldsymbol{\theta}_0))^{-1}S\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + o_p(1). \end{aligned} \tag{4}$$

The expected first term in Equation (4) can be written as

$$\mathrm{tr}\Big\{SV_{\mathrm{X_o}}(\boldsymbol{\theta}_0)E\Big[\sqrt{n}\frac{1}{n}\nabla\log f(Y_\mathrm{o}|\boldsymbol{\alpha}_0)\sqrt{n}\frac{1}{n}\nabla\log f(Z_\mathrm{o}|Y_\mathrm{o};\boldsymbol{\beta}_0)'\Big]S'+SV_{\mathrm{X_o}}(\boldsymbol{\theta}_0)S'(V_{Y_\mathrm{o}}(\boldsymbol{\theta}_0))^{-1}\Big\},$$

where $\mathrm{tr}\{\}$ is the trace of the matrix in $\{\}$. Because $E_{g(Z_\mathrm{o}|Y_\mathrm{o})}[\nabla\log f(Z_\mathrm{o}|Y_\mathrm{o};\boldsymbol{\beta}_0)|Y_\mathrm{o}]=\mathbf{0}$, the first term in this trace is zero. Therefore, the expected first term is $\mathrm{tr}\big\{SV_{\mathrm{X_o}}(\boldsymbol{\theta}_0)S'(V_{Y_\mathrm{o}}(\boldsymbol{\theta}_0))^{-1}\big\}$. Next, we calculate the expected second term in Equation (4) as

$$\begin{aligned}&-\frac{1}{2}\mathrm{tr}\Big\{E\big[\sqrt{n}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\sqrt{n}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)'\big]S'(V_{Y_\mathrm{o}}(\boldsymbol{\theta}_0))^{-1}S\Big\}\\&=-\frac{1}{2}\mathrm{tr}\Big\{SV_{\mathrm{X_o}}(\boldsymbol{\theta}_0)S'(V_{Y_\mathrm{o}}(\boldsymbol{\theta}_0))^{-1}\Big\}.\end{aligned}$$

In summary, $D_1=\frac{1}{2}\mathrm{tr}\Big\{SV_{\mathrm{X_o}}(\boldsymbol{\theta}_0)S'(V_{Y_\mathrm{o}}(\boldsymbol{\theta}_0))^{-1}\Big\}+o_p(1)$.

Second, we evaluate D_2.

$$\begin{aligned}D_2&=E[\log f(Y_\mathrm{o}|\boldsymbol{\alpha}_0)]-nE_{g(\boldsymbol{y})}[\log f(\boldsymbol{y}|\boldsymbol{\alpha}_0)]\\&=E_{g(\boldsymbol{y})}[\log f(Y_\mathrm{c}|\boldsymbol{\alpha}_0)-\log f(Y_\mathrm{m}|Y_\mathrm{o};\boldsymbol{\alpha}_0)]-nE_{g(\boldsymbol{y})}[\log f(\boldsymbol{y}|\boldsymbol{\alpha}_0)]\\&=-E[E[\log f(Y_\mathrm{m}|Y_\mathrm{o};\boldsymbol{\alpha}_0)|Y_\mathrm{o}]]\\&=-E[H(\boldsymbol{\alpha}_0|\boldsymbol{\alpha}_0)],\end{aligned}$$

where $H(\boldsymbol{\alpha}_0|\boldsymbol{\alpha}_0)$ is the H-function of the EM algorithm based on $\log f(Y_\mathrm{o}|\boldsymbol{\alpha})$ evaluated at $\boldsymbol{\alpha}=\boldsymbol{\alpha}_0$.

Finally, we evaluate D_3. Define $\boldsymbol{\eta}(\boldsymbol{\alpha})=E_{g(\boldsymbol{y})}[\log f(\boldsymbol{y}|\boldsymbol{\alpha})]$. Since $\nabla\boldsymbol{\eta}(\boldsymbol{\alpha}_0)=\mathbf{0}$,

$$\begin{aligned}\boldsymbol{\eta}(\hat{\boldsymbol{\alpha}})-\boldsymbol{\eta}(\boldsymbol{\alpha}_0)&=-\frac{1}{2}(\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}_0)'(V_{Y_\mathrm{c}}(\boldsymbol{\theta}_0))^{-1}(\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}_0)+o_p(1)\\&=-\frac{1}{2}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)'S'(V_{Y_\mathrm{c}}(\boldsymbol{\theta}_0))^{-1}S(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)+o_p(1),\end{aligned}$$

where $V_{Y_\mathrm{c}}(\boldsymbol{\theta}_0)=\big(-E[\nabla^2_{\boldsymbol{\alpha}}\log f(Y_\mathrm{c}|\boldsymbol{\alpha}_0)]\big)^{-1}$. Taking the expectation of both sides, we have

$$\begin{aligned}D_3&=\frac{1}{2}\mathrm{tr}\Big\{SE\big[\sqrt{n}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\sqrt{n}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)'\big]S'(V_{Y_\mathrm{c}}(\boldsymbol{\theta}_0))^{-1}\Big\}\\&=\frac{1}{2}\mathrm{tr}\Big\{SV_{\mathrm{X_o}}(\boldsymbol{\theta}_0)S'(V_{Y_\mathrm{c}}(\boldsymbol{\theta}_0))^{-1}\Big\}.\end{aligned}$$

Therefore, ignoring the term $o_p(1)$, we obtain an unbiased estimator for $E[\log f(\boldsymbol{y}|\hat{\boldsymbol{\alpha}})]$ as

$$\log f(Y_\mathrm{o}|\hat{\boldsymbol{\alpha}})-E[H(\boldsymbol{\alpha}_0|\boldsymbol{\alpha}_0)]-\frac{1}{2}\mathrm{tr}\Big[SV_{\mathrm{X_o}}(\boldsymbol{\theta}_0)S'\Big\{(V_{Y_\mathrm{o}}(\boldsymbol{\theta}_0))^{-1}+(V_{Y_\mathrm{c}}(\boldsymbol{\theta}_0))^{-1}\Big\}\Big].\tag{5}$$

The bias cannot be computed since unknown quantities such as $\boldsymbol{\alpha}$ are involved.

We can construct an estimated version of the IC given in Equation (5). By replacing the unknown parameter with the consistent ML estimator and eliminating the expectation from the second term in Equation (5), we can substitute the first two terms with

$$\log f(Y_\mathrm{o}|\hat{\boldsymbol{\alpha}})-H(\hat{\boldsymbol{\alpha}}|\hat{\boldsymbol{\alpha}})=Q_Y(\hat{\boldsymbol{\alpha}}|\hat{\boldsymbol{\alpha}}).$$

This is a natural estimator for the first two terms in Equation (5). The penalty term is estimated by replacing $\boldsymbol{\theta}_0$ with its ML estimator, which can be consistently estimated under the assumption that the data as a whole are MAR. As a result, the following Equation (6) is obtained. As is conventional in deriving the AIC, we double the bias, obtaining the corrected bias as

$$IC = -2Q_Y(\hat{\boldsymbol{\alpha}}|\hat{\boldsymbol{\alpha}}) + \mathrm{tr}\left[S\hat{V}_{X_o}(\hat{\boldsymbol{\theta}})S'\left\{\hat{V}_{Y_o}(\hat{\boldsymbol{\theta}})^{-1} + \hat{V}_{Y_c}(\hat{\boldsymbol{\theta}})^{-1}\right\}\right], \tag{6}$$

where $Q_Y(\hat{\boldsymbol{\alpha}}|\hat{\boldsymbol{\alpha}})$ is the Q-function of the EM algorithm based on $\log f(Y_c|\boldsymbol{\alpha})$ evaluated at $\boldsymbol{\alpha} = \hat{\boldsymbol{\alpha}}$ and $\hat{V}_\bullet$ is the sample variance of $V_\bullet$ for each suffix $\bullet$. Under the regularity condition assumed in [16], the estimators of the variance-covariance matrices in (6) converges to their population version (i.e., $\hat{V}_{Y_o}(\hat{\boldsymbol{\theta}})$ converges to $V_{Y_o}(\boldsymbol{\theta}_0)$ in probability as $n \to \infty$).

This IC is an extension of the AIC for missing data. It becomes the original AIC [1] when $X_o = X_c = Y_o = Y_c$ and $\boldsymbol{\theta} = \boldsymbol{\alpha}$ without missing data. Similar to the original AIC, our IC shares the starting point $E[\log f(\boldsymbol{y}|\hat{\boldsymbol{\theta}})]$ with [3,4]. However, the first and second terms of our IC differ from those of [3]. Our IC shares the first term with that of [4] and that of [6], but not the second term. This difference arises from the way in which the bias is approximated. The former [4] approximated the bias under the setting that all the subsets are MAR. The latter [6] approximated the bias by using the Hermite expansion and the Markov Chain Monte Carlo methods.

4. Simulations

To confirm the validity of our proposed IC, two simulations were performed.

4.1. Selection of the Parameter Structure

We will confirm through simulations that our IC selects the correct model when there is missing data for variables (y_1, y_2) in $\boldsymbol{x} = (y_1, y_2, z)'$. In so doing, we compare our IC with a slightly modified AIC, since the original AIC proposed by [1] cannot deal with missing data:

$$\mathrm{AIC} = -2\log f(Y_o|\hat{\boldsymbol{\alpha}}) + 2p,$$

where p is the dimension of vector $\hat{\boldsymbol{\alpha}}$. This has been used as a competitor to the information criterion proposed by [4]. As an additional competitor, we use $\mathrm{AIC}_{\mathrm{cd}}$, which Cavanaugh and Shumway [4] developed for missing data:

$$\mathrm{AIC}_{\mathrm{cd}} = -2Q(\hat{\boldsymbol{\alpha}}|\hat{\boldsymbol{\alpha}}) + 2\mathrm{tr}\left(\hat{V}_{Y_c}(\hat{\boldsymbol{\alpha}})^{-1}\left\{\frac{1}{2}\nabla^2_{\boldsymbol{\alpha}}\log f(Y_o|\hat{\boldsymbol{\alpha}})\right\}\right).$$

Four candidate models were considered, all of which are based on a trivariate normal distribution with variables $\boldsymbol{x} = (y_1, y_2, z)$. The respective models restrict the means of (y_1, y_2), and/or the variances of (y_1, y_2), or impose no restrictions at all. Model 1 restricts the two expectations to be equal and the two variances to be equal. Model 2 restricts that only the expectations be equal. Model 3 restricts that only the two variances be equal. Model 4 has no restrictions. In the simulations, one of the four models is the true data-generating model, and a model is selected according to each of the three information criteria: AIC, $\mathrm{AIC}_{\mathrm{cd}}$, and our proposed IC.

With respect to model selection, it should be noted that all the criteria assume that the true model must be included as a special case of any of the candidate models. Model 1, for example, is a special case of the other three candidate models. On the other hand, Model 2 is not a special case of Model 3, since Model 2 has no restrictions on the variances. Theoretically, models 2, 3, and 4 cannot be compared using an information criterion. For each model, the models that can or cannot be theoretically compared are shown in Table 1. A number in parentheses is that of a model that cannot be theoretically compared. Nevertheless, we have included in the simulation these theoretically un-comparable models in the set of candidate models for comparison because such a comparison is often conducted in practice, as pointed out in [18].

Table 1. True models and the corresponding candidate models. A number in parentheses indicates the number of a model which is theoretically inappropriate to use for comparison but compared in practice.

True Model	Candidate Models
1	1, 2, 3, 4
2	1, 2, (3), 4
3	1, (2), 3, 4
4	(1), (2), (3), 4

The simulation was conducted using the following procedure. First, a complete data set of a given size was generated from a trivariate normal distribution $N(\boldsymbol{\mu}, \Sigma)$ for variables $(y_1, y_2, z)'$, where $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3)'$ and $\Sigma = [\sigma_{ab}]_{a,b=1,2,3}$. The mean and variance for the data generation are shown in Table 2, where the covariances are all set to 0.7. The model that is assumed to be true is varied.

Table 2. True values of parameters. The covariances are common.

True Model	μ_1	μ_2	μ_3	σ_{11}	σ_{22}	σ_{33}	σ_{ab} $(a \neq b)$
1	0	0	0	1	1	1	0.7
2	0	0	0	1	2	1	0.7
3	0.5	−0.5	0	1	1	1	0.7
4	0.5	−0.5	0	1	2	1	0.7

Second, denoting the missing indicators for (y_1, y_2) as (r_1, r_2), we generated missing values according to the following two missing-data mechanisms. The first mechanism is, for $i = 1, 2$,

$$P(r_i = 1|z) = \begin{cases} \psi_i & (z \geq 0), \\ 1 & (z < 0). \end{cases} \tag{7}$$

This is a case where the missing-data mechanism is non-smooth. We set ψ_1 at 0.3, and ψ_2 at 0.7. The second of the two missing-data mechanisms used the binomial distribution, with the observing probability given as, for $i = 1, 2$,

$$P(r_i = 1|z) = 1 - \frac{2\omega_i}{\exp(z) + \exp(-z)}. \tag{8}$$

This is a smooth missing-data mechanism. We set ω_1 at 0.3, and ω_2 at 0.7. With the first missing-data mechanism, the values for (y_1, y_2) beyond the threshold $z = 0$ are always missing, while those below $z = 0$ are always observed. In the second mechanism, the missingness of (y_1, y_2) is stochastically determined based on the inverse of the hyperbolic cosine. Notice that the missing data for (y_1, y_2, z) are MAR for both mechanisms and that those for (y_1, y_2) are NMAR since z, which is not included in the models, causes the missingness of y_1 and y_2.

Third, for the missing data, the model that gives the minimum IC value is selected for each of the three ICs. The procedure is repeated 1000 times, and the number of times that each of the four models is selected is recorded.

The results for the first missing-data mechanism are summarized in Table 3. As can be seen in the table, our IC is generally superior to its two competitors. In the cases where Models 1 to 3 are true, the proposed IC outperforms the other criteria for any of the sample sizes. In the case where Model 4 is true, the proposed IC underperforms the other criteria when the sample size is small. However, as the sample size increases from 200 to 800, the proposed IC gradually selects the correct model as often as the other criteria. In addition

to the above findings, the AIC is found to be more robust against the violation of the missing-data mechanism than AIC_{cd}.

The results for the second missing-data mechanism are given in Table 4. Here, again, our IC is shown to be generally superior to its competitors. While all the information criteria are capable of selecting the true model, as the sample size increases, our IC is more likely to select the true model. For any true model, our IC outperforms the other information criteria except when Model 4 is true. Even in that case, our IC is equivalent to the others. In sum, our IC nearly consistently outperforms its major competitors and selects the correct model in all of the cases.

Table 3. Performance comparison under the missing-data mechanism (7).

		True Model															
		1				**2**				**3**				**4**			
n	criterion	Selected				Selected				Selected				Selected			
		1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
100	Proposed	754	117	106	23	67	630	87	216	0	0	855	145	0	0	131	869
	AIC_{cd}	506	163	243	88	6	538	3	453	0	0	749	251	0	0	8	992
	AIC	597	134	211	58	10	578	4	408	0	0	808	192	0	0	13	987
200	Proposed	743	99	143	15	12	652	38	298	0	0	892	108	0	0	46	954
	AIC_{cd}	441	142	295	122	0	397	0	603	0	0	724	276	0	0	0	1000
	AIC	559	111	255	75	0	432	0	568	0	0	811	189	0	0	0	1000
400	Proposed	672	104	197	27	1	638	15	346	0	0	881	119	0	0	17	983
	AIC_{cd}	298	118	399	185	0	205	0	795	0	0	685	315	0	0	0	1000
	AIC	429	93	345	133	0	247	0	753	0	0	752	248	0	0	0	1000
800	Proposed	652	83	234	31	0	592	5	403	0	0	893	107	0	0	6	994
	AIC_{cd}	190	58	466	286	0	40	0	960	0	0	646	354	0	0	0	1000
	AIC	270	58	464	208	0	51	0	949	0	0	727	273	0	0	0	1000

Table 4. Performance comparison under the missing-data mechanism (8).

		True Model															
		1				**2**				**3**				**4**			
n	criterion	Selected				Selected				Selected				Selected			
		1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
100	Proposed	829	88	74	9	134	743	9	114	0	0	914	86	0	1	121	878
	AIC_{cd}	484	200	232	84	73	638	24	265	0	0	716	284	0	0	98	902
	AIC	670	165	138	27	154	661	17	168	0	0	808	192	0	0	170	830
200	Proposed	820	99	74	7	28	852	3	117	0	0	895	105	0	0	28	972
	AIC_{cd}	465	210	212	113	8	710	3	279	0	0	675	325	0	0	12	988
	AIC	662	166	134	38	20	800	4	176	0	0	799	201	0	0	24	976
400	Proposed	795	124	70	11	2	882	0	116	0	0	867	133	0	0	1	999
	AIC_{cd}	418	251	193	138	0	710	1	289	0	0	611	389	0	0	1	999
	AIC	598	239	120	43	0	813	1	186	0	0	717	283	0	0	1	999
800	Proposed	765	158	60	17	0	885	0	115	0	0	827	173	0	0	0	1000
	AIC_{cd}	375	303	163	159	0	718	0	282	0	0	536	464	0	0	0	1000
	AIC	549	279	106	66	0	824	0	176	0	0	656	344	0	0	0	1000

4.2. Selection of Linear Regression Models

In this simulation, we demonstrate how our IC works for linear regression models. Since the derivation given above does not consider the covariate variables, our IC does not directly apply to a regression model. However, it is in the practical analysis of a regression model that we need information criteria to select from the various candidate models. Thus, we naively applied our IC in a simulation and monitored its performance.

In this simulation, we compared the performance of our proposed IC with that of the AIC and the AIC_{cd} criteria. Four regression models were used in the comparison. From here on, for simplicity, we will use more common notation. Let y be the dependent variable and $\boldsymbol{x} = (x_1, x_2)'$ be independent variables. The regression model is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon,$$

where ϵ has mean zero and a finite variance conditional on (x_1, x_2). The four models considered are Model 5 with $\beta_1 = \beta_2 = 0$, Model 6 with β_1 nonzero and $\beta_2 = 0$, which is the true model, Model 7 with $\beta_1 = 0$ and β_2 nonzero, and Model 8 with both β_1 and β_2 nonzero. For example, the data-generating model of Model 6 with $(\beta_0, \beta_1, \beta_2) = (1, \frac{1}{\sqrt{2}}, 0)$ is that the joint distribution of (y, x_1, x_2) is a trivariate normal with mean $[0, 0, 0]'$ and variance

$$\begin{bmatrix} 1 & \frac{1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{2}} & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

These four models do not necessarily have a nested/nesting relationship. Model 5 is a special case of Models 6 to 8. Models 6 and 7 are a special case of Model 8, but they do not nest with one another. Models 6 and 7 are not in a nesting relationship. Thus, theoretically speaking, these four models cannot be compared with the AIC-type information criteria. However, in practice, the AIC criterion has been commonly used for the comparison of nonnested/nonnesting models [18]. Thus, we compare the four models according to this convention.

The missingness of y and x_1 depends only on x_2. To represent the missing-data mechanism in a more formal way, we introduce missing-data indicators (r_y, r_{x_1}) for (y, x_1), each of which takes a value of 1 when the corresponding variable is observed and 0 when it is not. The missing-data mechanisms that were used are

$$\begin{aligned} P(r_y = 1|x_2) &= 1 - \frac{2\omega_y}{\exp(x_2) + \exp(-x_2)}, \\ P(r_{x_1} = 1|x_2) &= 1 - \frac{2\omega_{x_1}}{\exp(x_2) + \exp(-x_2)}, \end{aligned}$$

with setting ω_y at 0.7, and ω_{x_1} at 0.3.

The simulation was conducted according to the following procedure. First, a sample without missing values is created by drawing random sample of size n from the joint normal distribution of (y, x_1, x_2) which is specified above. Second, the missing values in the sample are created through either of the missing-data mechanisms for (y, x_1) as introduced above. Finally, the proposed IC and the competitor ICs are computed with such a sample for Models 1 to 4, and each is used to select the model corresponding to the minimum value. We repeated this procedure 1000 times, counting the number of times that each model is selected. The counts are used as an indicator of how well each of the ICs work in selecting the best model.

The results of the simulation are shown in Table 5. Regardless of which model is the actual true model, our proposed IC appears to work well. In the case where Model 5 is true, our proposed IC outperforms the other two competitors for any sample size. Here, AIC_{cd} works much better than AIC, but falls behind our IC. In the case where either Models 6 or 7 is true, AIC_{cd} shows the best performance for any sample size. The second best is our IC. AIC selects the wrong model, Model 8, more than half the time. Interestingly, all three IC's select neither Model 5 that is in a nesting relationship, nor Model 7(6) that is not in a nesting relationship when Model 6(7) is true. In the case in which Model 8 is true, AIC performs best. However, while for a sample size of 200, AIC selects the true model more often than our proposed IC, for sample sizes of more than 200, our proposed IC is comparable. AIC_{cd} always selects the wrong models. In summary, our proposed IC performed best or, at worst, second best throughout the simulation. Even in the second-best case, the performance of our proposed IC is very close to that of its higher-performing competitor.

Note that even for different values of the parameters in the missing-data mechanisms, we observed almost the same trend.

Table 5. Performance comparison with the regression model.

		True Model															
		5				**6**				**7**				**8**			
		Selected				Selected				Selected				Selected			
n	Criterion	5	6	7	8	5	6	7	8	5	6	7	8	5	6	7	8
200	Proposed	912	17	36	35	0	983	0	17	0	0	969	31	5	21	115	859
	AIC_{cd}	703	153	144	0	0	1000	0	0	0	0	1000	0	0	271	729	0
	AIC	157	205	213	425	0	439	0	561	0	0	442	558	0	6	19	975
400	Proposed	915	18	31	36	0	982	0	18	0	0	980	20	0	0	3	997
	AIC_{cd}	706	138	156	0	0	1000	0	0	0	0	1000	0	0	188	812	0
	AIC	160	223	201	416	0	430	0	570	0	0	452	548	0	0	1	999
800	Proposed	890	30	30	50	0	987	0	13	0	0	976	24	0	0	0	1000
	AIC_{cd}	699	162	139	0	0	1000	0	0	0	0	1000	0	0	112	888	0
	AIC	178	217	180	425	0	420	0	580	0	0	401	599	0	0	0	1000
1600	Proposed	887	39	31	43	0	976	0	24	0	0	971	29	0	0	0	1000
	AIC_{cd}	705	141	154	0	0	1000	0	0	0	0	1000	0	0	28	972	0
	AIC	180	191	177	452	0	425	0	575	0	0	445	555	0	0	0	1000

5. Conclusions

This paper considered the practical situation in which a set of missing data may not be MAR and developed a new information criterion to handle such a situation. This contrasts with previous ICs, which implicitly consider only the unrealistic case in which the data are MAR as a whole and subsets of the data are MAR as well. Our proposed IC uses the largest data set to estimate model parameters and circumvents the problem that the conventional ICs ignore. Using numerical simulations, it was shown that our new IC works better than, or at worst, equivalently to, its competitors.

The study of missing-data information criterion is far from complete. The present paper requires refinement. Although we applied our IC to regression, this was done without a rigorous basis, as noted in the body of the paper. A solid foundation for the comparison of regression models needs to be developed. Furthermore, application of the proposed IC can be extended to cases in which the parameters of the model are not a linear function of the parameters for the entire data. The approach used to develop our IC should not be limited to regression models; rather, it can be applied to other conventional statistical models such as the generalized linear regression and the explanatory factor model. Further refinements of our IC is also possible. Currently our IC uses $H(\hat{\boldsymbol{\alpha}}|\hat{\boldsymbol{\alpha}})$ as the estimator of the H-function $H(\hat{\boldsymbol{\alpha}}|\hat{\boldsymbol{\alpha}})$. That is, implementing the MCMC-based estimator of the H-function presented in [6] to our IC can improve the performance of model selection.

Further, the framework of our proposed IC might be applied to the data taken under the two-phase design. Under the two-phase design, an efficient estimation of the regression parameter has been extensively developed [19,20]. In the two-phase design, some part of the first-phase variables are observed for all subjects and the second-phase variables are observed only for some of the subjects who are selected based on the first-phase variables to avoid high cost. Since the missingness of the two-phase variables are caused by the values of the first-phase variables, the entire data are MAR. However, the variables of interest vary from analysis to analysis, and in some cases, the subset of the data on the model become NMAR. Hence, the variable selection becomes necessary. In such a case, our proposed IC might contribute to select variables. It is necessary to develop our IC to deal with such a two-phase design.

Author Contributions: The authors carried out this work and drafted the manuscript collaboratively. In particular, K.T. focused on construction of the estimator of the proposed criteria. K.H. conducted all simulations. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by JSPS KAKENHI (Grants-in-Aid for Scientific Research) Grant Numbers 18K11205 (Takai) and 15K15950 (Hayashi).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors are grateful to the associate editor and anonymous reviewers for their comments and suggestions that served to materially improve the manuscript.

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

References

1. Akaike, H. A new look at the statistical model identification. *IEEE Trans. Autom. Control* **1974**, *19*, 716–723. [CrossRef]
2. Takeuchi, K. Distribution of informational statistics and a criterion for model fitting. *Suri-Kagaku* **1976**, *153*, 12–18. (In Japanese)
3. Shimodaira, H. A new criterion for selecting models from partially observed data. In *Selecting Models from Data: AI and Statistics IV*; Cheeseman, P., Oldford, R.W., Eds.; Springer: New York, NY, USA, 1994; pp. 21–30.
4. Cavanaugh, J.E.; Shumway, R.H. An Akaike information criterion for model selection in the presence of incomplete data. *J. Stat. Plan. Inference* **1998**, *67*, 45–65. [CrossRef]
5. Seghouane, A.K.; Bekara, M.; Fleury, G. A criterion for model selection in the presence of incomplete data based on Kullback's symmetric divergence. *Signal Process.* **2005**, *85*, 1405–1417. [CrossRef]
6. Ibrahim, J.G.; Zhu, H.; Tang, N. Model Selection Criteria for Missing-Data Problems Using the EM Algorithm. *J. Am. Stat. Assoc.* **2008**, *103*, 1648–1658. [CrossRef] [PubMed]
7. Consentino, G.; Claeskens, F. Variables selection with incomplete covariate data. *Biometrics* **2008**, *64*, 1062–1069.
8. Little, R.J.A.; Rubin, D.B. *Statistical Analysis with Missing Data*; Wiley: Hoboken, NJ, USA, 2002.
9. Enders, C.K. *Applied Missing Data Analysis (Methodology in the Social Sciences)*; Guilford Press: New York, NY, USA, 2010.
10. Schafer, J.L.; Graham, J.W. Missing data: Our view of the state of the art. *Psychol. Methods* **2002**, *7*, 147–177. [CrossRef] [PubMed]
11. Graham, J.W. Adding missing-data-relevant variables to FIML-based structural equation models. *Struct. Equ. Model. Multidiscip. J.* **2003**, *10*, 80–100. [CrossRef]
12. Collins, L.M.; Schafer, J.L.; Kam, C.M. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychol. Methods* **2001**, *6*, 330–351. [CrossRef] [PubMed]
13. Chang, W.Y.; Richards, D.S.P. Finite-sample inference with monotone incomplete multivariate normal data, I. *J. Multivar. Anal.* **2009**, *100*, 1883–1899. [CrossRef]
14. Chang, W.Y.; Richards, D.S.P. Finite-sample inference with monotone incomplete multivariate normal data, II. *J. Multivar. Anal.* **2010**, *101*, 603–620. [CrossRef]
15. White, H. Maximum likelihood estimation of misspecified models. *Econometrics* **1982**, *50*, 1–25. [CrossRef]
16. Takai, K.; Kano, Y. Asymptotic Inference with Incomplete Data. *Commun. Stat. Theory Methods* **2013**, *42*, 3174–3190. [CrossRef]
17. Konishi, S.; Kitagawa, G. *Information Criteria and Statistical Modeling*; Springer: New York, NY, USA, 2007.
18. Akaike, H. Prediction and Entropy. In *A Celebration of Statistics*; Atkinson, A.C., Fienberg, S.E., Eds.; Springer: New York, NY, USA, 1985; pp. 1–24.
19. Tao, R.; Zeng, L.; Lin, D.-Y. Optimal designs of two-phase studies. *J. Am. Stat. Assoc.* **2020**, *115*, 1946–1959. [CrossRef] [PubMed]
20. Yang, C.; Diao, L.; Cook, R.J. Adaptive response-dependent two-phase designs: Some results on robustness and efficiency. *Stat. Med.* **2022**, *41*, 4403–4425. [CrossRef] [PubMed]