Published in final edited form as:
J Am Stat Assoc. 2008 December 1; 103(484): 1648–1658. doi:10.1198/016214508000001057.

Model Selection Criteria for Missing-Data Problems Using the EM Algorithm

Joseph G. Ibrahim,
Joseph G. Ibrahim is Alumni Distinguished Professor (E-mail: ibrahim@bios.unc.edu), Department of Biostatistics, University of North Carolina, Chapel Hill.

Hongtu Zhu, and
Hongtu Zhu is Associate Professor (E-mail: hzhu@bios.unc.edu), Department of Biostatistics, University of North Carolina, Chapel Hill.

Niansheng Tang
Niansheng Tang is Professor, Department of Statistics, Yunnan University, Kunming (E-mail: nstang@ynu.edu.cn)

Abstract

We consider novel methods for the computation of model selection criteria in missing-data problems based on the output of the EM algorithm. The methodology is very general and can be applied to numerous situations involving incomplete data within an EM framework, from covariates missing at random in arbitrary regression models to nonignorably missing longitudinal responses and/or covariates. Toward this goal, we develop a class of information criteria for missing-data problems, called $\mathrm{IC}_{H,Q}$, which yields the Akaike information criterion and the Bayesian information criterion as special cases. The computation of $\mathrm{IC}_{H,Q}$ requires an analytic approximation to a complicated function, called the *H*-function, along with output from the EM algorithm used in obtaining maximum likelihood estimates. The approximation to the *H*-function leads to a large class of information criteria, called $\mathrm{IC}_{\tilde{H}(k),Q}$. Theoretical properties of $\mathrm{IC}_{\tilde{H}(k),Q}$, including consistency, are investigated in detail. To eliminate the analytic approximation to the *H*-function, a computationally simpler approximation to $\mathrm{IC}_{H,Q}$, called IC_Q, is proposed, the computation of which depends solely on the *Q*-function of the EM algorithm. Advantages and disadvantages of $\mathrm{IC}_{\tilde{H}(k),Q}$ and IC_Q are discussed and examined in detail in the context of missing-data problems. Extensive simulations are given to demonstrate the methodology and examine the small-sample and large-sample performance of $\mathrm{IC}_{\tilde{H}(k),Q}$ and IC_Q in missing-data problems. An AIDS data set also is presented to illustrate the proposed methodology.

Keywords

EM algorithm; *H*-function; Kullback; Leibler divergence; Missing data; *Q*-function

1. INTRODUCTION

Missing data have long been a problem in various settings, including surveys, clinical trials, and longitudinal studies. Responses and/or covariates may be missing, and methods for handling the missing data often depend on the mechanism that generated the missing values. Unless the data are missing completely at random (MCAR), a complete-case analysis can be both inefficient and biased; therefore, distributional and modeling assumptions often are made in missing-data problems, and the resulting estimates and tests may be sensitive to these assumptions. For this reason, sensitivity analyses are commonly done to check the robustness of the parameters of interest and their standard errors under different modeling schemes (see, e.g., Rubin 1977; Little 1993, 1994, 1995; Copas and Li 1997; van Steen, Molenberghs, and Thijs 2001; Verbeke, Molenberghs, Thijs, Lasaffre, and Kenward 2001; Jansen, Molenberghs, Aerts, Thjis, and van Steen 2003; Troxel, Ma, and Heitjan 2004). Although these analyses demonstrate the effect of assumptions on estimates and tests, they do not indicate which modeling strategy is best, nor do they specifically address model selection for a given class of models.

Model selection criteria typically depend on the likelihood function based on the observed data, and any sensible model selection criterion must depend on this quantity in some way. In missing-data problems, it is very challenging to obtain a suitable and accurate approximation of the observed data likelihood, which involves intractable multiple integration, and/or directly maximize the observed data likelihood and compute the Akaike information criterion (AIC) and/or the Bayesian information criterion (BIC), for example, as well other model selection criteria. The EM algorithm maximizes the Q-function (formally defined in Sec. 2.1) at each iteration, avoiding direct maximization of the observed data likelihood, which typically is a more difficult function to maximize. A natural and important question is whether we can use the key components of the EM algorithm, such as the Q-function, to develop an easily computable model selection criterion.

In this article we consider a class of information-based model selection criteria, called $\mathrm{IC}_{H,Q}$, for missing-data problems. The class of model selection criteria includes AIC and BIC as special cases, as well other model selection criteria that have been proposed in the literature, mainly for settings not involving missing-data. The essential novel feature of the proposed model selection criteria is that they essentially depend only on output from the EM algorithm for their computation. Our development is based on the fact that the observed data log-likelihood in missing data problems can be written as a difference between two functions, the Q-function of the EM algorithm and another quantity called the H-function. The Q-function and the H-function are formally defined in Section 2.1. The Q-function can be computed solely from the EM output, but the H-function cannot; however, we show that after the H-function is analytically approximated, it then can be computed as part of the EM output, resulting in model selection criteria, $\mathrm{IC}_{\tilde{H}(k),Q}$, that depend solely on the EM output. We give a theoretical justification for $\mathrm{IC}_{\tilde{H}(k),Q}$ and derive the asymptotic properties of $\mathrm{IC}_{\tilde{H}(k),Q}$. We also consider another class of model selection criteria, IC_{Q}, which use only the Q-function in their construction and thus omit the H-function in their construction. We also show that compared with $\mathrm{IC}_{\tilde{H}(k),Q}$, IC_{Q} is an inferior approximation to $\mathrm{IC}_{H,Q}$, but it may be adequate when the fraction of missing information is small.

The rest of the article is organized as follows. In Section 2 we introduce $\mathrm{IC}_{H,Q}$, $\mathrm{IC}_{\tilde{H}(k),Q}$, and IC_{Q}. We present three theorems characterizing consistency and asymptotic properties of $\mathrm{IC}_{\tilde{H}(k),Q}$ as general model selection criteria. In Section 3 we present two extensive simulation studies, one involving missing-at-random (MAR) covariates in linear models and one involving MAR covariates in generalized linear models (GLMs). These simulations compare the finite-sample performance of $\mathrm{IC}_{\tilde{H}(k),Q}$ and IC_{Q} and examine how these criteria can be used to determine the best-fitting model from a candidate set of proposed models. In Section 3.3 we analyze a data set from a study of the relationship between acquired immune deficiency syndrome (AIDS) and the use of condoms that includes not missing-at-random (NMAR) (i.e., nonignorable) covariates as well as responses. We conclude with a discussion in Section 4.

2. EM–BASED MODEL SELECTION CRITERIA

2.1 EM Algorithm

For simplicity, we only consider an independent-type incomplete-data (ITID) model throughout the article, even though most of the development here is valid for a large class of statistical models involving missing data. Assume the observed data $D_{obs} = (z_{1,obs}, \ldots, z_{n,obs})$, the missing data $D_{mis} = (z_{1,mis}, \ldots, z_{n,mis})$, and the complete data $D_{com} = (z_{1,com}, \ldots, z_{n,com})$, in which $z_{i,com} = (z_{i,mis}, z_{i,obs})$ for $i = 1,\ldots, n$. The ITID model assumes that $z_{i,com}$ and $z_{j,com}$ are independent for $i \neq j$. Moreover, the dimensions of $z_{i,mis}$ and $z_{i,obs}$ may vary across i; for instance, in GLMs with missing covariates, some observations may have missing covariates and others may not. This kind of model structure is very general and subsumes most commonly used models, such as GLMs with missing responses and/or covariates and random-effects models (Zhu, Lee, Wei, and Zhou 2001; Ibrahim, Chen, and Lipstiz 1999, 2001).

Suppose that we want to compare a general model for the complete data, $g(D_{com}; \theta)$, with the true model for the complete data, $f(D_{com})$. The model for the complete data is the product of a model for the observed data, $g(D_{obs}; \theta)$, and a model for the missing data given the observed data, $g(D_{mis}|D_{obs}; \theta)$. Correspondingly, $f(D_{com}) = f(D_{obs})f(D_{mis}|D_{obs})$, where $f(D_{mis}|D_{obs})$ and $f(D_{obs})$ are the true model for the missing data given the observed data and that for the observed data. Specifically, for the ITID model, we have

$$\begin{aligned} g(D_{obs};\theta) &= \prod_{i=1}^{n} g(z_{i,obs};\theta), \\ f(D_{obs}) &= \prod_{i=1}^{n} f(z_{i,obs}), \end{aligned} \tag{1}$$

$$\begin{aligned} g(D_{com};\theta) &= \prod_{i=1}^{n} g(z_{i,com};\theta), \quad \text{and} \\ f(D_{com}) &= \prod_{i=1}^{n} f(z_{i,com}), \end{aligned} \tag{2}$$

where $f(z_{i,obs})$ and $g(z_{i,obs}; \theta)$ denote the true and postulated models for $z_{i,obs}$, and $f(z_{i,com})$ and $g(z_{i,com}; \theta)$ denote the true and postulated models for $z_{i,com}$.

The EM algorithm (Dempster, Laird, and Rubin 1977) has been a popular technique for obtaining maximum likelihood (ML) estimates in missing-data problems (Little and Rubin 2002; Meng and van Dyk 1997; Ibrahim 1990; Ibrahim and Lipsitz 1996). The EM algorithm consists of two key steps as follows. At the sth step of the EM algorithm, given $\theta^{(s)}$, the E-step involves evaluating the Q-function given by

$$Q(\theta|\theta^{(s)}) = E[\log g(D_{com};\theta)|D_{obs};\theta^{(s)}], \tag{3}$$

where $E[\cdot|D_{obs}; \theta^{(s)}]$ denotes the conditional expectation with respect to $g(D_{mis}|D_{obs}; \theta^{(s)})$. Recall that the Q-function can be written as

$$Q(\theta|\theta^{(s)}) = \log g(D_{obs};\theta) + H(\theta|\theta^{(s)}), \tag{4}$$

where

$$H(\theta|\theta^{(s)}) = E[\log g(D_{mis}|D_{obs};\theta)|D_{obs};\theta^{(s)}] \tag{5}$$

is called the *H*-function. The M-step is to maximize $Q(\theta|\theta^{(s)})$ to compute $\theta^{(s+1)}$. At EM convergence, we can obtain three byproducts: $\hat{\theta}$, $Q(\hat{\theta}|\hat{\theta})$, and samples drawn from $g(D_{mis}|D_{obs};\hat{\theta})$. We use these three quantities in constructing our proposed model selection criteria in the subsequent sections.

2.2 Development of IC$_{H,Q}$

Our main interest is to develop a class of model selection criteria for missing-data problems based on the observed data likelihood $g(D_{obs};\theta)$. However, some missing-data problems have very complicated observed data likelihood functions, for which $g(D_{obs};\theta)$ has no closed form, so that its direct evaluation is not computationally feasible or computationally accurate. Because

$$\log g(D_{obs};\theta)=Q(\theta|\theta^{(s)})-H(\theta|\theta^{(s)}), \tag{6}$$

this suggests that we may compute $g(D_{obs};\theta)$ from the EM output—namely, from the *Q*-function $Q(\hat{\theta}|\hat{\theta})$ and the *H*-function $H(\hat{\theta}|\hat{\theta})$ at EM convergence. Thus we consider the class of model selection criteria given by

$$\begin{aligned}\mathrm{IC}_{H,Q}&=-2\log g(D_{obs};\widehat{\theta})+\widehat{c}_n(\widehat{\theta})\\&=-2Q(\widehat{\theta|\theta})+2H(\widehat{\theta|\theta})+\widehat{c}_n(\widehat{\theta}),\end{aligned} \tag{7}$$

where $\hat{c}_n(\hat{\theta})$ is a penalty term that is a function of the data and the fitted model. Different forms of the model penalty $\hat{c}_n(\hat{\theta})$ lead to different criteria; for instance, when $\hat{c}_n(\hat{\theta})=2d$ in (7), where *d* denotes the dimension of θ, we obtain the AIC of Akaike (1973), given by AIC $=-2\log g(D_{obs};\hat{\theta})+2d$. When $\hat{c}_n(\hat{\theta})=\log(n)d$, then (7) reduces to the BIC of Schwarz (1978). We note that the penalty term $\hat{c}_n(\hat{\theta})$ is neither *Q*-function–based nor specific to missing-data problems; rather, it is a general penalty term chosen by the user, mimicking the penalty terms for general model selection information criteria as discussed in the literature (Macquarrie and Tsai 1998;Konishi and Kitagawa 2008).

There is a subtle computational problem with (7) in that although the *Q*-function is a direct byproduct of the EM output, the *H*-function is not a direct byproduct of the EM output. Specifically, the density $g(D_{mis}|D_{obs};\theta)$ in the *H*-function does not have a closed form for many missing-data problems and typically is quite complicated, and thus the integrand of the *H*-function itself does not have a closed form. Thus $g(D_{mis}|D_{obs};\theta)$ first needs an analytic approximation to allow computation of the *H*-function through the EM output. Once $g(D_{mis}|D_{obs};\theta)$ is analytically approximated (i.e., the integrand of the *H*-function is analytically approximated), the *H*-function can be computed by Monte Carlo integration using samples from $g(D_{mis}|D_{obs};\hat{\theta})$ at EM convergence. Samples from this density are obtained by carrying out Markov chain Monte Carlo (MCMC) methods and are direct byproducts of the Monte Carlo EM algorithm (MCEM), as discussed by Ibrahim, Lipsitz, and Chen (1999). Using these samples, we then can obtain an EM-based estimator of the approximation to the *H*-function, which we discuss in detail in the next section. We note that when $\hat{c}_n(\hat{\theta})=2d$, an EM-based approximation to the AIC is obtained by replacing the *H*-function by its estimator.

2.3 Approximation of $g(D_{mis}|D_{obs},\hat{\theta})$ in IC$_{H,Q}$

We propose a simple but useful method for approximating the *H*-function. In general, given the MCMC samples from $g(D_{mis}|D_{obs};\hat{\theta})$ (Ibrahim, Lipsitz, and Chen 1999), we can get a

Monte Carlo approximation of the integral $\int w(D_{mis})g(D_{mis}|D_{obs};\hat{\theta})\,dD_{mis}$ only if $w(D_{mis})$ has an analytic closed form. Although $w(D_{mis}) = \log g(D_{mis}|D_{obs};\hat{\theta})$ for $H(\hat{\theta}|\hat{\theta})$, $g(D_{mis}|D_{obs};\hat{\theta})$ does not have a closed form for most missing-data problems.

We propose using a truncated Hermite expansion as an approximation of each $g(z_{i,mis}|z_{i,obs};\hat{\theta})$, leading to

$$\tilde{g}\,(z_{i,mis};\widehat{\mu}_i,\widehat{\sum}_i,\psi_i,k)=P_i^2(t;\psi_i,k)\varphi(z_{i,mis};\widehat{\mu}_i,\widehat{\sum}_i), \tag{8}$$

where $t=R_i^{-1}(z_{i,mis}-\widehat{\mu}_i)$, $\widehat{\sum}_i=R_iR_i^T$, and $\varphi(z_{i,mis};\hat{\mu}_i,\hat{\Sigma}_i)$ is a multivariate normal density with mean $\hat{\mu}_i$ and covariance matrix $\hat{\Sigma}_i$. In addition, $\hat{\mu}_i = \mu_i(\hat{\theta})$ and $\hat{\Sigma}_i = \Sigma_i(\hat{\theta})$ are the conditional mean and covariance matrix of $z_{i,mis}$ given $z_{i,obs}$ at $\hat{\theta}$. Here $P_i(t;\psi_i,k)$ is a multivariate polynomial of order k and ψ_i are the coefficients of $P_i(t;\psi_i,k)$. If $g(z_{i,mis}|z_{i,obs};\hat{\theta})$ belongs to a smooth class of functions, then $\tilde{g}(z_{i,mis};\hat{\mu}_i,\hat{\Sigma}_i,\psi_i,k)$ approximates $g(z_{i,mis}|z_{i,obs};\hat{\theta})$ well for even small k, say $k = 1$ and 2 (Gallant and Nychka 1987); for instance, if $z_{i,mis}$ is uni-variate and $k = 2$, then

$$P_i(t;\psi_i,2)=\frac{1+\psi_{i1}t+\psi_{i2}t^2}{\sqrt{1+\psi_{i1}^2+3\psi_{i2}^2+2\psi_{i2}}}.$$

If $k = 0$, then we obtain $P_i^2(t;\psi_i,0)=1$ and $\tilde{g}(z_{i,mis};\hat{\mu}_i,\hat{\Sigma}_i,\psi_i,k) = \varphi(z_{i,mis};\hat{\mu}_i,\hat{\Sigma}_i)$. It has been shown both numerically and theoretically that the truncated Hermite expansion can provide an accurate approximation to $g(z_{i,mis}|z_{i,obs};\hat{\theta})$ as $k \to \infty$ (Fenton and Gallant 1996). Moreover, in the truncated Hermite expansion, the multivariate normal density can be replaced by another density, such as a multivariate t, Poisson, or gamma density (Cameron and Johansson 1997; Kim 2007).

We can use $\tilde{g}(z_{i,mis};\hat{\mu}_i,\hat{\Sigma}_i,\psi_i,k)$ to produce a Monte Carlo estimate of $H(\hat{\theta}|\hat{\theta})$. The detailed steps are summarized as follows. In step 1 we draw a set of random samples, $\{z_{i,mis}^{(s)}:s=1,\ldots,S_0\}$, from $g(z_{i,mis}|z_{i,obs};\hat{\theta})$ using MCMC sampling, where S_0 is a prefixed number. In step 2 we use the sample mean and covariance matrix of $\{z_{i,mis}^{(s)}:s=1,\ldots,S_0\}$ to approximate $\hat{\mu}_i$ and $\hat{\Sigma}_i$. In step 3, because $\{z_{i,mis}^{(s)}:s=1,\ldots,S_0\}$ are observations from $g(z_{i,mis}|z_{i,obs};\hat{\theta})$, we can then obtain estimators (e.g., ML estimators) of ψ_i, denoted by $\hat{\psi}_i(k)$, for given k and $i = 1, \ldots, n$. Because S_0 can be arbitrarily large, we can assume that $\hat{\mu}_i$ and that $\hat{\Sigma}_i$ are exact and that $\hat{\psi}_i(k)$ is the minimizer of the Kullback–Leibler divergence between $\tilde{g}(z_{i,mis};\hat{\mu}_i,\hat{\Sigma}_i,\psi_i,k)$ and $g(z_{i,mis}|z_{i,obs};\hat{\theta})$, that is,

$$\widehat{\psi}_i(k)=\arg\min_{\psi_i}\left\{\int\log\frac{g(z_{i,mis}|z_{i,obs};\widehat{\theta})}{\tilde{g}(z_{i,mis};\widehat{\mu}_i,\widehat{\sum}_i,\psi_i,k)}\times g(z_{i,mis}|z_{i,obs};\widehat{\theta})dz_{i,mis}\right\}.$$

In step 4 we calculate

$$\begin{aligned}\tilde{H}(k|\widehat{\theta}) &= S_0^{-1}\sum_{s=1}^{S_0}\sum_{i=1}^{n}\log\tilde{g}(z_{i,mis}^{(s)};\widehat{\mu}_i,\widehat{\Sigma}_i,\widehat{\psi}_i(k),k)\\ &= E[\log\widehat{g}(D_{mis}|D_{obs};k,\widehat{\theta})|z_{i,obs};\widehat{\theta}]+o(1),\end{aligned} \tag{9}$$

where $\tilde{g}(D_{mis}|D_{obs};k,\widehat{\theta})=\prod_{i=1}^{n}\tilde{g}(z_{i,mis};\widehat{\mu}_i,\widehat{\sum}_i,\widehat{\psi}_i(k),k)$ and $o(1)$ converges to 0 as $S_0 \to \infty$. In general, the computational burden in steps 1, 2, and 4 is minimal, whereas computing $\hat{\psi}_i(k)$ for each i can be computationally cumbersome when k is relatively large. If we set k at 0, then we can avoid the maximization in step 3.

Based on $\tilde{H}(k|\hat{\theta})$, we can obtain an approximation of $\mathrm{IC}_{H,Q}$ as

$$\begin{aligned}&\mathrm{IC}_{H,Q} = -2Q(\widehat{\theta}|\widehat{\theta})+2\tilde{H}(k|\widehat{\theta})+2H(\widehat{\theta}|\widehat{\theta})\\ &-2\tilde{H}(k|\widehat{\theta})+c_n(\widehat{\theta})\\ &\approx \mathrm{IC}_{\tilde{H}(k),Q} = -2Q(\widehat{\theta}|\widehat{\theta})+2\tilde{H}(k|\widehat{\theta})+c_n(\widehat{\theta}).\end{aligned} \tag{10}$$

Moreover, because $\tilde{H}(k|\hat{\theta}) \le H(\hat{\theta}|\hat{\theta})$ according to Jensen's inequality, $\mathrm{IC}_{\tilde{H}(k),Q} \le \mathrm{IC}_{H,Q}$. Although $\tilde{H}(k|\hat{\theta})$ converges to $H(\hat{\theta}|\hat{\theta})$ as $k \to \infty$, choosing a large k is computationally inefficient. Moreover, we observe that $\tilde{H}(k|\hat{\theta})$ based on a small k, say 0 or 1, also can produce reasonable results, as shown in Section 3. Thus this Hermite approximation for $g(z_{i,mis}|z_{i,obs};\hat{\theta})$ is quite attractive, because model choice is quite robust with respect to the choice of k.

2.4 General Theoretical Development for $\mathrm{IC}_{\tilde{H}(k),Q}$

Here we present a formal theoretical development for $\mathrm{IC}_{\tilde{H}(k),Q}$, which was defined in the previous section. We define

$$\begin{aligned}&\tilde{g}_{(k)}(D_{obs};\theta_1,\theta_2)=\exp\{E[\log g(D_{com};\theta_1)\\ &-\log\tilde{g}(D_{mis}|D_{obs};k,\theta_1)|D_{obs};\theta_2]\Big\}\end{aligned} \tag{11}$$

as an approximation to $g(D_{obs};\theta_1)$, where $E[\cdot|D_{obs};\theta_2]$ denotes the conditional expectation taken with respect to $g(D_{mis}|D_{obs};\theta_2)$. As $k \to \neq$, it can be shown that under some conditions, $\tilde{g}(D_{mis}|D_{obs};k,\theta_1)$ converges to $g(D_{mis}|D_{obs};\theta_1)$, and thus $\tilde{g}_{(k)}(D_{obs};\theta_1,\theta_2)$ converges to $g(D_{obs};\theta_1)$. To develop a general class of model selection criteria, we consider the Kullback–Leibler divergence between $\tilde{g}_{(k)}(D_{obs};\theta_1,\theta_2)$ and $f(D_{obs})$, defined by

$$\begin{aligned}K(\theta_1,\theta_2) &= -\int\log(\tau(D_{obs};\theta_1,\theta_2))f(D_{obs})dD_{obs}\\ &= \int f(D_{obs})\log f(D_{obs})dD_{obs}\\ &-\int f(D_{obs})\log\tilde{g}_{(k)}(D_{obs};\theta_1,\theta_2)dD_{obs},\end{aligned} \tag{12}$$

where $\tau(D_{obs};\theta_1,\theta_2)=\tilde{g}_{(k)}(D_{obs};\theta_1,\theta_2)/f(D_{obs})$. The quantity $K(\theta,\theta)$ is an overall measure of the goodness of fit of $\tilde{g}_{(k)}(D_{obs};\theta,\theta)$ relative to $f(D_{obs})$. Because the first term in (12) is independent of any fitted model and can be ignored, our goal of selecting a model can be accomplished using the second term of (12).

If $g(D_{obs};\theta)$ is specified correctly, then $\hat{\theta}$ is asymptotically efficient, and the likelihood ratio statistic is a most sensitive criterion for detecting deviations of the model parameters from their true values. But even though $g(D_{obs};\theta)$ is "always" mis-specified, White (1994) established consistency and asymptotic normality of $\hat{\theta}$ under some conditions. Thus it is desirable to evaluate $K(\hat{\theta},\theta^{\star})$. A simple estimator of $K(\hat{\theta},\theta^{\star})$ is given by substituting for the distribution of D_{obs}, denoted by F_{obs}, the empirical distribution function $\hat{F}_{obs}$. Thus, except for a constant, $K(\hat{\theta},\theta^{\star})$ can be approximated by

$$\tilde{K}_{(k)}(\widehat{\theta},\theta^{\star})=-\log\tilde{g}_{(k)}(D_{obs};\widehat{\theta},\theta^{\star}).$$

We obtain the following theorems, whose detailed proofs are given in Appendix A. The following conditions are needed to facilitate development of our methods, although they may not be the weakest possible conditions. Even though $g(D_{obs};\theta)$ may be misspecified, the ML estimator, $\hat{\theta}$, converges to the θ_{n*} that minimizes $E\{\sum_{i=1}^{n}\ell(z_{i,obs};\theta)\}=\sum_{i=1}^{n}\int\ell(z_{i,obs};\theta)f(z_{i,obs})dz_{i,obs}$, where $\ell(z_{i,obs};\theta)=\log g(z_{i,obs};\theta)$ (see, e.g., White 1994). For simplicity, we further assume that $\theta_{n*}=\theta_{*}$ for all n and $E\{\partial_{\theta}\ell(z_{i,obs};\theta_{*})\}=0$ for all i. The conditions are as follows:

(C1) θ_{*} is unique and an interior point of Θ, where Θ is a compact set in R^{p}.

(C2) $\hat{\theta}\rightarrow\theta_{*}$ in probability as $n\rightarrow\infty$.

(C3) For all i, $\ell(z_{i,obs};\theta)$ is three times continuously differentiable on θ, and $|\partial_{j}\ell(z_{i,obs};\theta)|^{2}$ and $|\partial_{j}\partial_{j'}\partial_{l}\ell(z_{i,obs};\theta)|$ are dominated by $B_{i}(z_{i,obs})$ for all $j, j', l = 1, \ldots, d$, where $\partial_{j}=\partial/\partial\theta_{j}$. The same smoothness condition also holds for $h_{(k)}(z_{i,obs};\theta)=E[\log\tilde{g}(z_{i,mis}|z_{i,obs};k,\theta)|z_{i,obs};\theta]$.

(C4) For each $\varepsilon>0$, there exists a finite C such that

$$\sup_{n\geq 1}n^{-1}\sum_{i=1}^{n}E[B_{i}(z_{i,obs})\mathbf{1}\{B_{i}(z_{i,obs})>C\}]<\varepsilon$$

for all n, where $\mathbf{1}\{B_{i}(z_{i,obs})>C\}$ is the indicator function of $B_{i}(z_{i,obs})>C$.

(C5)

$$\lim_{n\rightarrow\infty}n^{-1}E\left\{-\sum_{i=1}^{n}\partial_{\theta}^{2}\ell(z_{i,obs};\theta_{*})\right\}=A(\theta_{*})$$

and

$$\lim_{n\rightarrow\infty}n^{-1}\sum_{i=1}^{n}E\left\{\partial_{\theta}\ell(z_{i,obs};\theta_{*})\partial_{\theta}\tilde{K}_{(k)}(\theta,\theta^{\star})^{T}\right\}|_{\theta=\theta_{*}}=B(\theta_{*}|\theta^{\star}),$$

where $A(\theta_{*})$ is positive definite.

Condition (C1) defines the uniqueness of the "true" parameter value. Condition (C2) is the consistency of $\hat{\theta}$. Condition (C3) is a smoothness condition on $\ell(z_{i,obs};\theta)$ and $h_{(k)}(z_{i,obs};\theta)$.

Condition (C4) is a standard Lindeberg condition, and (C5) can be easily proved by the law of large numbers.

Theorem 1—For ITID models, if conditions (C1), (C2), and (C3) hold true, then

$$\begin{aligned}&n^{-1}|\tilde{K}_{(k)}(\widehat{\theta},\theta^{\star})-E[\tilde{K}_{(k)}(\widehat{\theta},\theta^{\star})]|\\&+n^{-1}|E[\tilde{K}_{(k)}(\widehat{\theta},\theta^{\star})]-E[\tilde{K}_{(k)}(\theta_{*},\theta^{\star})]|\to 0\end{aligned}\tag{13}$$

in probability, where $E[\tilde{K}_{(k)}(\cdot,\cdot)]$ denotes the expectation with respect to the observed data, $E[\tilde{K}_{(k)}(\hat{\theta},\theta^{\star})]$ denotes $E[\tilde{K}_{(k)}(\theta,\theta^{\star})]$ evaluated at $\theta=\hat{\theta}$, and θ_* is the pseudo true value of θ based on $g(D_{obs};\theta)$.

Theorem 1 indicates that $n^{-1}\tilde{K}_{(k)}(\hat{\theta},\theta^{\star})$ is a consistent estimator of $n^{-1}E[\tilde{K}_{(k)}(\theta_*,\theta^{\star})]$. Now consider the situation in which we want to compare values of $\tilde{K}_{(k)}(\hat{\theta},\theta^{\star})$ under different models for $g(D_{com};\theta)$. Although $n^{-1}\tilde{K}_{(k)}(\hat{\theta},\theta^{\star})$ is a consistent estimator of $n^{-1}E[\tilde{K}_{(k)}(\hat{\theta},\theta^{\star})]$, it is an overestimate of $n^{-1}E[\tilde{K}_{(k)}(\hat{\theta},\theta^{\star})]$, because the same data are used to estimate θ and to approximate F_{obs}. Following Akaike (1973) and Konishi and Kitagawa (2008), we calculate the bias of $n^{-1}\tilde{K}_{(k)}(\hat{\theta},\theta^{\star})$ in estimating $n^{-1}E[\tilde{K}_{(k)}(\hat{\theta},\theta^{\star})]$ as

$$\begin{aligned}&b(\theta^{\star})=E_{D_{obs}}\left\{\tilde{K}_{(k)}(\widehat{\theta},\theta^{\star})-E[\tilde{K}_{(k)}(\widehat{\theta},\theta^{\star})]\right\}\\&=b_1(\theta^{\star})+o(1),\end{aligned}\tag{14}$$

where $E_{D_{obs}}$ denotes the expectation taken with respect to the observed data. Although it may be difficult to calculate the explicit form of $b(\theta^{\star})$, we can derive an asymptotic bias expression, denoted $b_1(\theta^{\star})$.

Theorem 2—For ITID models, if conditions (C1)–(C5) are true, then the asymptotic bias of $\tilde{K}_{(k)}(\hat{\theta},\theta^{\star})$ in estimating $E[\tilde{K}_{(k)}(\hat{\theta},\theta^{\star})]$ is given by

$$\begin{aligned}&E_{D_{obs}}\{\tilde{K}_{(k)}(\widehat{\theta},\theta^{\star})-E[\tilde{K}_{(k)}(\widehat{\theta},\theta^{\star})]\}\\&=b(\theta^{\star})\\&=\mathrm{tr}\{A(\theta_*)^{-1}B(\theta_*|\theta^{\star})\}+o_p(1),\end{aligned}\tag{15}$$

where $A(\theta)$ and $B(\theta|\theta^{\star})$ are defined in condition (C5).

Theorem 2 provides a theoretical basis for using $-2\tilde{K}_{(k)}(\hat{\theta},\theta^{\star})+b(\theta^{\star})$ as a model selection criterion, and this quantity is precisely a bias-corrected estimate of $-2E_{D_{obs}}[\tilde{K}_{(k)}(\hat{\theta},\theta^{\star})]$. In particular, if $\theta^{\star}=\theta_*$ and $g(D_{obs};\theta)$ is specified correctly, then $A(\theta_*)-B(\theta_*|\theta_*)$ converges to a zero matrix and $b(\theta_*)\approx 2d$ as $k\to\infty$. But because $\theta^{\star}$ is unknown, we replace θ_* and $\theta^{\star}$ by $\hat{\theta}$. In particular, under the correct specification of $g(D_{obs};\theta)$, $b(\hat{\theta})$ should be close to $2d$ for large k. This leads to an approximation to the AIC as $\mathrm{AIC}_{\tilde{H}(k),Q}=-2\tilde{K}_{(k)}(\hat{\theta},\hat{\theta})+2d$.

We now establish sufficient conditions to ensure consistency of $\mathrm{IC}_{\tilde{H}(k),Q}$. Following Nishii (1988), we consider two parametric models for the complete data, with densities given by

$$\mathcal{M}_t=\left\{g_{(t)}(D_{com};\theta_{(t)})=g_{(t)}(D_{obs};\theta_{(t)})g_{(t)}(D_{mis}|D_{obs};\theta_{(t)}):\theta_{(t)}\in\Theta_{(t)}\subset R^{d_t}\right\} \tag{16}$$

for $t = 1, 2$. For each $\mathfrak{M}_t$, the ML estimator $\hat{\theta}_{(t)}$ converges in probability to the pseudo true value, denoted by $\theta_{*(t)}$. To select a better model, we first calculate

$$\begin{aligned}&\mathrm{dIC}_{\tilde{H}(k),Q21}\\&=\mathrm{IC}_{\tilde{H}(k),Q2}-\mathrm{IC}_{\tilde{H}(k),Q1}\\&=2Q(\widehat{\theta}_{(1)}|\widehat{\theta}_{(1)})-2\tilde{H}(k|\widehat{\theta}_{(1)})-\widehat{c}_n(\widehat{\theta}_{(1)})\\&-2Q(\widehat{\theta}_{(2)}|\widehat{\theta}_{(2)})+2\tilde{H}(k|\widehat{\theta}_{(2)})+\widehat{c}_n(\widehat{\theta}_{(2)}).\end{aligned} \tag{17}$$

We choose $\mathfrak{M}_2$ if $\mathrm{dIC}_{\tilde{H}(k),Q21} < 0$ and $\mathfrak{M}_1$ otherwise. Define

$$\begin{aligned}&\delta_{21,k}=E[Q(\theta_{*(1)}|\theta_{*(1)})-\tilde{H}(k|\theta_{*(1)})]\\&-E[Q(\theta_{*(2)}|\theta_{*(2)})-\tilde{H}(k|\theta_{*(2)})]\end{aligned}$$

and $\delta_{c21} = \hat{c}_n(\hat{\theta}_{(2)}) - \hat{c}_n(\hat{\theta}_{(1)})$. Moreover, without loss of generality, we assume that $d_2 > d_1$ and $\hat{c}_n(\hat{\theta}_{(2)}) > \hat{c}_n(\hat{\theta}_{(1)})$; for instance, if $\hat{c}_n(\hat{\theta}_{(2)}) = d_2 \log(n)$, then $\delta_{c21} = (d_2 - d_1) \log(n)$.

Theorem 3—Suppose that $\mathfrak{M}_1$ and $\mathfrak{M}_2$ are ITID models and satisfy conditions (C1)–(C5). We then have the following results:

- **a.** If $\liminf_n n^{-1}\delta_{21,k} > 0$ and $\delta_{c21} = o_p(n)$, then $\mathrm{dIC}_{\tilde{H}(k),Q21} > 0$ in probability.
- **b.** Assume that

$$\begin{aligned}&\limsup_n n^{-1/2}\{E[Q(\theta_{*(2)}|\widehat{\theta}_{(2)})-\tilde{H}(k|\widehat{\theta}_{(2)})]\\&-E[Q(\theta_{*(1)}|\widehat{\theta}_{(1)})-\tilde{H}(k|\widehat{\theta}_{(1)})]\}<\infty,\end{aligned}$$

  , $n^{-1/2}\{\tilde{H}(k|\hat{\theta}_{(t)}) - E[\tilde{H}(k|\hat{\theta}_{(t)})]\} = O_p(1)$, and $n^{-1/2} \times \{Q(\hat{\theta}_{(t)}|\hat{\theta}_{(t)}) - E[Q(\theta_{*(t)}|\ \hat{\theta}_{(t)})]\} = O_p(1)$ for $t = 1, 2$. Then $\mathrm{dIC}_{\tilde{H}(k),Q21} \le 0$ in probability as $n^{-1/2}\delta_{c21} \to \infty$.
- **c.** Assume that $Q(\theta_{*(2)}|\hat{\theta}_{(2)}) - Q(\theta_{*(1)}|\hat{\theta}_{(1)}) = O_p(1)$ and $\tilde{H}(k|\hat{\theta}_{(2)}) - \tilde{H}(k|\hat{\theta}_{(1)}) = O_p(1)$. Then $\mathrm{dIC}_{\tilde{H}(k),Q21} \le 0$ in probability as $\delta_{c21} \to \infty$.

Theorem 3 has some important implications. Theorem 3a indicates that $\mathrm{IC}_{\tilde{H}(k),Q}$ chooses$\mathfrak{M}_2$ as $\liminf_n n^{-1}\delta_{21,k} > 0$ and $\delta_{c21} = o_p(n)$. Generally, the most commonly used $\hat{c}_n(\hat{\theta})$, such as $2d$, $d \log(n)$, and $d \log \log(n)$ ($d > 0$), all satisfy the condition $\delta_{c21} = o_p(n)$ (Nishii 1988). The condition $\liminf_n n^{-1}\ \delta_{21,k} > 0$ ensures that $\mathrm{IC}_{\tilde{H}(k),Q}$ chooses a model with large $E[Q(\theta_*|\theta_*) - \tilde{H}(k|\theta_*)]$. If $\mathfrak{M}_1$ and $\mathfrak{M}_2$ have the same average $n^{-1}E[Q(\theta_*|\theta_*) - \tilde{H}(k|\theta_*)]$ (i.e., $\liminf_n n^{-1} \times \delta_{21,k} = 0$), then Theorem 3b and c indicate that $\mathrm{IC}_{\tilde{H}(k),Q}$ picks out the “simpler” $\mathfrak{M}_1$ when δ_{c21} increases to ∞ at a certain rate [e.g., $\log(n)$]. But $\hat{c}_n = 2d$ does not satisfy this condition. Thus, because $\mathrm{IC}_{\tilde{H}(k),Q}$ with $\hat{c}_n = 2d$ is the EM-based estimate of the AIC, it tends to overfit the data in this scenario.

2.5 Using $\mathrm{IC}_{\tilde{H}(k),Q}$ in the Presence of Nonignorable Missing Data

Although our model selection criteria $\mathrm{IC}_{\tilde{H}(k),Q}$ are quite general and can be used with MAR or NMAR covariate and/or response data, here we offer some caution and advice on using these criteria with NMAR data. First, it is often argued that in missing-data problems, there is little information in the data regarding the form of the missing-data mechanism, and the parametric assumption of the missing-data mechanism itself is not "testable" from the data. Thus nonignorable modeling should be viewed as a sensitivity analysis involving a more complicated model. In this sense, it is dangerous to use any model selection criterion to directly compare MAR and NMAR models. Formally, we give the following guidelines on using $\mathrm{IC}_{\tilde{H}(k),Q}$:

1. $\mathrm{IC}_{\tilde{H}(k),Q}$ should be used to choose among a family of MAR models and/or choose among a family of NMAR models. They should not be used to choose among an aggregate set of MAR and NMAR models nor should they be used to judge the fit of MAR models versus NMAR models.

2. Once the best MAR model and best NMAR model are found using step 1, further sensitivity analyses can be done on those two models to examine changes in estimates of the main regression coefficients of interest in the sampling model. These sensitivity analyses can be carried out by examining estimates of the regression coefficients of the sampling model under several different parametric forms of the missing-data mechanism.

2.6 IC_Q

Because the analytic approximation to the integrand of the H-function and its computation may be cumbersome for large k, it also might be desirable to obtain a model selection criterion that does not involve the H-function and whose components depend only on quantities obtained directly from the EM output. Toward this goal, we can obtain such a criterion by dropping $H(\hat{\theta}|\hat{\theta})$ from (7), leading to the criterion

$$\mathrm{IC}_Q = -2Q(\widehat{\theta}|\widehat{\theta}) + \widehat{c}_n(\widehat{\theta}). \tag{18}$$

Thus IC_Q can be viewed as a crude approximation to $\mathrm{IC}_{H,Q}$ in which $H(\hat{\theta}|\hat{\theta})$ is omitted. When $\hat{c}_n(\hat{\theta}) = 2d$ in (18), this leads to the criterion

$$\mathrm{AIC}_Q = -2Q(\widehat{\theta}|\widehat{\theta}) + 2d. \tag{19}$$

There are clear advantages and disadvantages to using IC_Q instead of $\mathrm{IC}_{\tilde{H}(k),Q}$. One advantage of using IC_Q is that it is computationally easier than $\mathrm{IC}_{\tilde{H}(k),Q}$, not requiring an approximation to the integrand of the H-function. But one clear disadvantage of IC_Q is that as a result of omitting the H-function, a model selection criterion based on the Q-function alone can overstate the amount of information in the missing data compared with the observed data log-likelihood function. Omitting the H-function can lead to a criterion with poor model selection properties in some cases, especially when the missing-data fraction is high. In general, we recommend using $\mathrm{IC}_{\tilde{H}(k),Q}$ over IC_Q.

3. SIMULATION STUDIES

In this section we report on several simulation studies used to investigate the finite-sample performance of $\mathrm{IC}_{\tilde{H}(k),Q}$ and IC_Q in linear models and GLMs with MAR covariates. More specifically, we demonstrate how $\text{IC}_{\tilde{H}(k),Q}$ and IC_Q can be used as model selection criteria for choosing the best-fitting model. In the simulation for the linear model with MAR and normally distributed covariates (Sec. 3.1), $\text{IC}_{H,Q}$ has an analytic closed form, and thus $g(z_{i,mis}|z_{i,obs}; \hat{\theta})$ has a closed form, and thus neither approximation or MCMC sampling is needed in this case. Therefore, we can assess the performance of the approximation in this setting by comparing $\{\text{IC}_{\tilde{H}(k),Q}:k = 0, 1\}$ to $\text{IC}_{H,Q}$, which is analytically equivalent to AIC in this case when $\hat{c}_n(\hat{\theta}) = 2d$. We also compare IC_Q with $\text{IC}_{H,Q}$.

But for the GLM with MAR covariates neither $\text{IC}_{H,Q}$ nor $g(z_{i,mis}|z_{i,obs}; \hat{\theta})$ has a closed form, and thus both the Hermite approximation and MCMC sampling are needed to compute $\text{IC}_{H,Q}$. In this setting, we do not attempt to compute AIC or BIC directly using Laplace approximations or numerical integration techniques, because these methods are not easy and quite cumbersome to implement, and, more importantly, the resulting approximations are very difficult to assess in terms of accuracy. Thus for GLMs, we only compute $\{\text{IC}_{\tilde{H}(k),Q}:k = 0, 1\}$ and IC_Q through the MCEM algorithm under several values of $\hat{c}_n(\hat{\theta})$.

3.1 Missing-at-Random Covariates in Linear Models

We generated simulated data sets from a linear regression model with one MAR covariate. This simulation study had three goals: (i) to demonstrate how $\text{IC}_{\tilde{H}(k),Q}$ for different k can be used as a tool for selecting a model from a candidate set of proposed models and evaluate and compare them with $\text{IC}_{H,Q}$, (ii) to compare IC_Q with $\text{IC}_{H,Q}$, and (iii) to compare the performance of IC_Q with $\text{IC}_{\tilde{H}(k),Q}$. To save space, we focus on $c_n(\hat{\theta}) = 2d$ throughout, although several additional simulation results are available for other values of $\hat{c}_n(\hat{\theta})$ including $\hat{c}_n(\hat{\theta}) = d\log(n)$.

Consider the true model $y_i|x_i \sim \text{N}(\beta_0 + \beta_1 x_i, \sigma^2)$, where $x_i \sim \text{N}(\mu, \tau^2)$ for $i = 1, \dots, n$. We generated the data set $\{(x_{i,} y_{i)}: i = 1, \dots, n\}$ as follows. First, we generated n independent random variables x_i from a $\text{N}(\mu, \tau^2)$ distribution; and then generated independent responses y_i from a $\text{N}(\beta_0 + \beta_1 x_i, \sigma^2)$ distribution. We then generated n independent standard normally distributed variables z_i that are independent of y_i and x_i. The true parameter values were taken to be $\beta_0 = .8$, $\beta_1 = .8$, $\sigma^2 = .8$, $\mu = .8$, $\tau^2 = .8$, and $n = 100, 300, 500$.

Furthermore, we assume that the response y_i and the additional covariate z_i are completely observed for $i = 1, \dots, n$, but the covariate x_i can be missing for some cases. We note that because z_i is fully observed for all cases, we need not specify a covariate distribution for z_i in the modeling strategy, but a covariate distribution for x_i must be specified, because x_i is missing for some cases. The missing-data mechanism for the x_i 's is defined as follows. We let $r_i = 1$ if x_i is missing and $r_i = 0$ if x_i is observed. Then the following logistic regression model is considered for the missing-data mechanism:

$$p(r_i=1|y_i, z_i)=\frac{\exp(\varphi_0+\varphi_1 y_i)}{1+\exp(\varphi_0+\varphi_1 y_i)}, \tag{20}$$

implying MAR covariates. To investigate the effect of the missingness fraction on the performance of the model selection criteria, we consider the following sets of true parameter values for φ_0 and φ_1: (I) $\varphi_0 = -4.0$, $\varphi_1 = 1.0$ giving an average missingness fraction for x_i roughly equal to 11%, and (II) $\varphi_0 = -3.5$, $\varphi_1 = 1.5$, giving an average missingness fraction of 29%.

We considered five candidate models:

- Model M1 (true model): $y_i|x_i \sim \text{N}(\beta_0 + \beta_1 x_i, \sigma^2)$, $x_i \sim \text{N}(\mu, \tau^2)$
- Model M2: $y_i|x_i \sim \text{N}(\beta_0, \sigma^2)$, $x_i \sim \text{N}(\mu, \tau^2)$

- Model M3: $y_i|x_{i,}\ z_i \sim \mathrm{N}(\beta_0 + \beta_1 x_i + \beta_2 z_{i,}\ \sigma^2)$, $x_i \sim \mathrm{N}(\mu, \tau^2)$
- Model M4: $y_i|x_{i,}\ z_i \sim \mathrm{N}(\beta_0 + \beta_1 x_i + \beta_2 x_i\, z_i, \sigma^2)$, $x_i \sim \mathrm{N}(\mu, \tau^2)$
- Model M5: $y_i|x_{i,}\ z_i \sim \mathrm{N}(\beta_0 + \beta_1 x_i + \beta_2 z_i + \beta_3 x_i\, z_{i,}\ \sigma^2)$, $x_i \sim \mathrm{N}(\mu, \tau^2)$.

We generated $R = 500$ simulated data sets from M1 and then calculated $\{\mathrm{IC}_{\tilde{H}\,(k),Q}$: $k = 0, 1\}$ and IC_Q for $\hat{c}_n(\hat{\theta}) = 2d$, and AIC ≡ $\mathrm{IC}_{H,Q}$ when $\hat{c}_n(\hat{\theta}) = 2d$ (Table 1).

Table 1 shows the number of times out of $R = 500$ simulations that each rank was achieved for M1, the true model for all model selection criteria. The columns in Table 1 correspond to the rankings of AIC_Q [$\mathrm{AIC}_Q \equiv \mathrm{IC}_Q$ when $\hat{c}_n(\hat{\theta}) = 2d$] under different settings, and the rows of Table 1 corresponds to the proposed criteria for different choices of $\hat{c}_n(\hat{\theta})$ and k. With $n = 100$ and case (I), M1 got ranked as number one 332 = 331 + 1 times by AIC_Q, 360 = 331 + 25 + 4 times by AIC [$\mathrm{IC}_{H,Q}$ with $\hat{c}_n(\hat{\theta}) = 2d$], 357 times by $\mathrm{IC}_{\tilde{H}\,(0),Q}$ with $\hat{c}_n(\hat{\theta}) = 2d$, and 355 times by $\mathrm{IC}_{\tilde{H}\,(1),Q}$ with $\hat{c}_n(\hat{\theta}) = 2d$. With $n = 100$ and case (II), M1 got ranked as number one 306 times by AIC_Q, 364 times by $\mathrm{IC}_{H,Q}$ with $\hat{c}_n(\hat{\theta}) = 2d$, 361 times by $\mathrm{IC}_{\tilde{H}\,(0),Q}$ with $\hat{c}_n(\hat{\theta}) = 2d$, and 358 times by $\mathrm{IC}_{\tilde{H}\,(1),Q}$ with $\hat{c}_n(\hat{\theta}) = 2d$. These results imply that AIC_Q performs reasonably well in all scenarios, but $\mathrm{IC}_{H,Q}$ outperforms AIC_Q, particularly for large missingness fractions. The $\{\mathrm{IC}_{\tilde{H}\,(k),Q:}\ k = 0, 1\}$ for $\hat{c}_n(\hat{\theta}) = 2d$ perform as well as $\mathrm{IC}_{H,Q}$ even for large missingness fractions, which is an attractive result demonstrating the suitability of the approximation. Moreover, increasing k does not seem to improve the performance of $\mathrm{IC}_{\tilde{H}\,(k),Q,}$ demonstrating its high degree of robustness. The $\{\mathrm{IC}_{\tilde{H}\,(k),Q}$: $k = 0, 1\}$ for $c_n(\hat{\theta}) = 2d$ outperform AIC_Q, particularly for large missingness fractions. Finally, we note that AIC yields very similar results to $\{\mathrm{IC}_{\tilde{H}\,(k),Q:} k = 0, 1\}$ for $\hat{c}_n(\hat{\theta}) = 2d$.

3.2 Missing-at-Random Covariates in Generalized Linear Models

In this section we consider a logistic regression model with one continuous covariate. Our primary aim is to evaluate $\{\mathrm{IC}_{\tilde{H}\,(k),Q:} k = 0, 1\}$ and IC_Q and compare them with each other. In this simulation study, covariates $x_1, \ldots, x_n$ are iid and generated from a N(.5, 1.0) distribution, and responses $y_1, \ldots, y_n$ are generated independently from a Bernoulli distribution with success probability $p(y_i=1|\beta_0,\beta_1,x_i)=\frac{\exp(\beta_0+\beta_1 x_i)}{1+\exp(\beta_0+\beta_1 x_i)}(i=1,\ldots,n)$. We also assume that $y_1, \ldots, y_n$ are completely observed, whereas $x_1, \ldots, x_n$ are MAR for some cases.

The missing data for the x_i were generated according to the missing-data mechanism in (20), and the z_i 's were generated exactly as described in Section 3.1. The true parameter values were taken to be $\beta_0 = \beta_1 = .8$ and $n = 100$, 300, and 500. To investigate the effect of the missingness fraction on our model selection criteria, we again considered two sets of true values for φ_0 and φ_1: (I) $\varphi_0 = -1.2$ and $\varphi_1 = -.8$, giving a missingness fraction of about 15%, and (II) $\varphi_0 = -.5$ and $\varphi_1 = -.8$, giving a missingness fraction of about 26%.

As in Section 3.1, we considered five candidate models:

- Model M_1 (true model): $\mathrm{logit}(p_i) = \beta_0 + \beta_1 x_{i,}\ x_i \sim \mathrm{N}(\mu, \tau^2)$
- Model M_2: $\mathrm{logit}(p_i) = \beta_0$, $x_i \sim \mathrm{N}(\mu, \tau^2)$
- Model M_3: $\mathrm{logit}(p_i) = \beta_0 + \beta_1 x_i + \beta_2 z_{i,}\ x_i \sim \mathrm{N}(\mu, \tau^2)$
- Model M_4: $\mathrm{logit}(p_i) = \beta_0 + \beta_1 x_i + \beta_2 z_i\, x_{i,}\ x_i \sim \mathrm{N}(\mu, \tau^2)$
- Model M_5: $\mathrm{logit}(p_i) = \beta_0 + \beta_1 x_i + \beta_2 z_i\, x_i + \beta_3 z_i$, $x_i \sim \mathrm{N}(\mu, \tau^2)$.

We simulated 500 data sets and then calculated $\{\mathrm{IC}_{\tilde{H}\,(k),Q:} k = 0, 1\}$ and IC_Q for $\hat{c}_n(\hat{\theta}) = 2d$ for each simulated data set. Table 2 shows the number of times out of $R = 500$ simulations that each rank was achieved for M1, the true model for all model selection criteria. Again, the columns in Table 2 correspond to the rankings of AIC_Q, and the rows correspond to several settings of the proposed criteria. The results are very similar to those reported in Section 3.1. For instance, with n = 100 and case (I), M1 was ranked as number one 302 times by AIC$_Q$, 319 times by IC $_{\tilde{H}\ (0),Q}$ with $\hat{c}_n(\hat{\theta}) = 2d$, and 317 times by IC $_{\tilde{H}\ (1),Q}$ with $\hat{c}_n(\hat{\theta}) = 2d$. These results imply that AIC$_Q$ performs reasonably well in all scenarios, and that increasing the missing-data fraction does not have a strong effect on AIC$_Q$ for accurately selecting the true model M1. The {IC $_{\tilde{H}\ (k),Q}$:k = 0, 1} for $\hat{c}_n(\hat{\theta}) = 2d$ perform reasonably well even for large missingness fractions. Moreover, increasing k does not seem to improve the performance of IC $_{\tilde{H}\ (k),Q}$. Again, the {IC $_{\tilde{H}\ (k),Q}$:k = 0, 1} for $\hat{c}_n(\hat{\theta}) = 2d$ outperform AIC$_Q$, particularly for large missingness fractions.

3.3 AIDS Data

We considered a data set from a study of the relationship between AIDS and the use of condoms (Morisky et al. 1998; Lee and Tang 2006). This complex data set requires sophisticated structural equations modeling in the presence of NMAR covariate and response data. An intriguing question is whether there is any model selection criterion for selecting the best-fitting model from a candidate set of structural equation models whose observed data likelihood functions involve high-dimensional integrals. Directly computing AIC and BIC, for example, using Laplace methods or high-dimensional numerical integration techniques is simply too hard and computationally cumbersome in this scenario, and moreover, the accuracy of such approximations is impossible to assess in this high-dimensional setting. Thus this example greatly motivates the need for EM-based criteria, such as IC$_{\tilde{H}\ (k),Q}$ and IC$_Q$.

For simplicity, we used only the data obtained from female sex workers in Philippine cities (Lee and Tang 2006). These data are related to knowledge of AIDS and attitude toward AIDS, beliefs, self-efficiency of condom use, and other variables. Nine variables in the original data set (items 33, 32, 31, 43, 72, 74, 27h, 27e, and 27i on the questionnaire) were taken as manifest variables in $\mathbf{y}_i = (y_{i1}, \ldots, y_{i9})^{T,}$ a continuous item x_{i1} (item 37) and an ordered categorical item x_{i2} (item 21, treated as continuous) were taken as covariates. The definitions of these nine items are given in Appendix B. In this data set, the variables $y_{i1}, y_{i2}, y_{i3}, y_{i7}, y_{i8}$, and y_{i9} were measured on a 5-point scale and thus were treated as continuous; variables y_{i4}, y_{i5}, and y_{i6} were continuous. There are n = 1,116 observations in this data set, and the manifest variables and covariates are missing at least once for 361 of them (32%). The missingness patterns for the manifest variables are shown in table 4 of Lee and Tang (2006). In this data set, the covariate x_{i2} is completely observed.

Following Lee and Tang (2006), the manifest variables (y_{i1}, y_{i2}, y_{i3}) are related to a latent variable, $\eta_{i,}$ that can be interpreted as the "threat of AIDS," whereas the manifest variables (y_{i4}, y_{i5}, y_{i6}) and (y_{i7}, y_{i8}, y_{i9}) are related to the latent variables ξ_{i1} and ξ_{i2}, which can be interpreted as "aggressiveness of the sex worker" and "worry of contracting AIDS." Specifically, to identify the relationship between the manifest variables $\mathbf{y}_i$ and the latent variables $\boldsymbol{\omega}_i = (\eta_{i,}\ \xi_{i1}, \xi_{i2})^{T,}$ we consider the following measurement equation:

$$\mathbf{y}_i=\mu+\mathbf{\Lambda}\omega_i+\varepsilon_i, \quad i=1,\ldots,n,$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_9)^T$ is a vector of intercepts, (ξ_{i1}, ξ_{i2}) is independent of the measurement error vector $\boldsymbol{\varepsilon}_{i,}$ $(\xi_{i1}, \xi_{i2}) \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Phi})$, and $\boldsymbol{\varepsilon}_i \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Psi})$, in which $\boldsymbol{\Psi} = \mathrm{diag}(\psi_1, \ldots, \psi_9)$ and $\boldsymbol{\Phi} = (\varphi_{ij)}$ is a 2 × 2 covariance matrix. We also assume the following structure for $\mathbf{\Lambda}$:

$$\mathbf{\Lambda}^T=\begin{pmatrix} 1.0^* & \lambda_{21} & \lambda_{31} & 0^* & 0^* & 0^* & 0^* & 0^* & 0^* \\ 0^* & 0^* & 0^* & 1.0^* & \lambda_{52} & \lambda_{62} & 0^* & 0^* & 0^* \\ 0^* & 0^* & 0^* & 0^* & 0^* & 0^* & 1.0^* & \lambda_{83} & \lambda_{93} \end{pmatrix},$$

where 0^* and 1.0^* are regarded as fixed values to identify the scale of the latent factor. We let $r_{yij} = 1$ if y_{ij} is missing and $r_{yij} = 0$ if y_{ij} is observed and $r_{xi1} = 1$ if x_{i1} is missing and $r_{xi1} = 0$ if x_{i1} is observed. Based on the missingness patterns, we assume that both the missing-data mechanisms of the manifest variables and the covariates are NMAR. In particular, we consider the following missing-data mechanisms for y_{ij} and x_{i1}:

$$M_y: \quad \text{logit}\{\Pr(r_{yij}=1|\tau)\}=\tau^T\mathbf{y}_{io}^*$$

and

$$M_x: \quad \text{logit}\{\Pr(r_{xi1}=1|\phi)\}=\phi_0+\phi_1 y_{i1}+\cdots+\phi_9 y_{i9},$$

where $\boldsymbol{\tau}$ is a vector of logistic regression coefficients, $\mathbf{y}_{io}^*=(1,\mathbf{y}_{io}^T)^T$ in which $\mathbf{y}_{io}$ is a vector corresponding to the observed data of $\mathbf{y}_{i,}$ and $\boldsymbol{\phi} = (\boldsymbol{\phi}_0, \boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_9)^T$. Because x_{i1} may be missing, we need to specify its distribution. For simplicity, we assume that $x_{i1} \sim N(0, \psi_{x)}$.

To study the relationship between η and (x_1, x_2, ξ_1, ξ_2), we consider four nonlinear structural equations models,

$$\begin{aligned}
&M_0:\ \eta_i=b_1x_{i1}+b_2x_{i2}+\gamma_1\xi_{i1}+\gamma_2\xi_{i2}+\gamma_3\xi_{i1}\xi_{i2}+\delta_i;\\
&M_1:\ \eta_i=b_1x_{i1}+b_2x_{i2}+\gamma_1\xi_{i1}+\gamma_2\xi_{i2}\\
&+\gamma_3\xi_{i1}\xi_{i2}+\gamma_4\xi_{i1}^2+\delta_i;\\
&M_2:\ \eta_i=b_1x_{i1}+b_2x_{i2}+\gamma_1\xi_{i1}+\gamma_2\xi_{i2}\\
&+\gamma_3\xi_{i1}\xi_{i2}+\gamma_4\xi_{i2}^2+\delta_i;
\end{aligned}$$

and

$$\begin{aligned}
&M_3:\ \eta_i=b_1x_{i1}+b_2x_{i2}+\gamma_1\xi_{i1}+\gamma_2\xi_{i2}\\
&+\gamma_3\xi_{i1}\xi_{i2}+\gamma_4\xi_{i1}^2+\gamma_5\xi_{i2}^2+\delta_i,
\end{aligned}$$

where $\delta_i \sim N(0, \psi_\delta)$. Clearly, all four models include the linear effect of “aggressiveness,” ξ_{i1}, and “worry,” ξ_{i2} and an interaction of “aggressiveness” and “worry.” The models M_1 and M_2, respectively, have the additional quadratic terms of “aggressiveness” and “worry.” Because M_3 includes all the possible terms of ξ_{i1} and ξ_{i2}, it may be considered the “full model.”

We calculated values of $\{\text{IC}_{\tilde{H}(k),Q}: k = 0, 1\}$ and IC_Q with $\hat{c}_n(\hat{\theta}) = 2d$ and $d \log(n)$ for all four models (Table 3). The calculation of $\{\text{IC}_{\tilde{H}(k),Q}: k = 0, 1\}$ and IC_Q was straightforward, because it only required quantities from the output of the EM algorithm for obtaining parameter estimates. Model M_0 was selected as best by all model selection criteria. The ML estimates of the parameters were obtained through the MCECM algorithm and specific parameter estimates for model M_0 are presented in Table 4. The factor loading estimates are positive and quite large, which implies a strong positive association between the latent variables and their corresponding indicators, and the estimated nonlinear structural equation is given by $\hat{\eta_i} = -.0579x_{i1} + .0821x_{i2} - .2711\xi_{i1} + .2505\xi_{i2} + .1897\xi_{i1}\xi_{i2}$. We note the fact that comparatively large (positive) values of $(\eta_{i,}\ x_{i2})$ (or x_{i1}, ξ_{i1}) and ξ_{i2} indicate that an individual feels a high (or low) threat from AIDS and is more worried about contracting AIDS. The foregoing equation has the following interpretation:

1. $\hat{b}_1 = -.0579$ indicates that the longer sex workers are in their jobs, the less threat they feel from AIDS, and $\hat{b}_2 = .0821$ implies that the more they think that they know about AIDS, the more threat they feel from AIDS.
2. $\hat{\gamma}_1 = -.2711$ shows that the more aggressive the sex workers are, the less threat they feel from AIDS, and $\hat{\gamma}_2 = .2505$ shows that sex workers who are more worried about contracting AIDS feel more of a threat from AIDS.
3. $\hat{\gamma}_3 = .1897$ indicates that ξ_{i1} and ξ_{i2} have a positive interaction effect on "threat of AIDS."

It is easily seen from the foregoing analysis that introducing an interaction term in the nonlinear structural equation to interpret the relationship between η_i and ξ_{i1}, ξ_{i2} is very necessary, and we could get various different effects for different cases. The estimated correlation between "aggressiveness," ξ_{i1}, and "worry," ξ_{i2}, is $-.1819$, which indicates that they are negatively correlated.

4. DISCUSSION

We have proposed a general class of model selection criteria, IC $_{\tilde{H}(k),Q}$, for missing-data problems. The computation of IC $_{\tilde{H}(k),Q}$ can be obtained directly from the EM output. The theory of IC $_{\tilde{H}(k),Q}$ is quite general and can be applied to various types of missing-data models for which the EM algorithm is applicable. Moreover, IC $_{\tilde{H}(k),Q}$ can be directly applied to many other problems in which the ECM algorithm and the ECME algorithm can be applied (Liu and Rubin 1994; Meng and Rubin 1993). We have given theoretical underpinnings for these criteria and have shown that they are consistent. We note, however, that although consistency is a desirable and interesting property, it does not shed light on how to penalize the observed data likelihood for model parsimony in finite samples. Further research is needed to determine the best choice of penalty in missing-data problems. We have also demonstrated that the Hermite approximation to the integrand of the H -function, $\log(g(D_{mis}|D_{obs}; \hat{\theta}))$, is quite robust for model choice for several choices of k, leading to an attractive feature of the proposed approximation. Choices of $k = 0, 1$ worked as well as those of $k = 10$ and larger. This is a comforting feature, because it shows that model choice is not sensitive to the degree of the Hermite approximation to $g(D_{mis}|D_{obs}; \hat{\theta})$.

The penalty terms $\hat{c}_n(\hat{\theta})$ can have a profound effect on the finite-sample performance of IC $_{\tilde{H}(k),Q}$ and IC_Q. Compared with $\hat{c}_n(\hat{\theta}) = 2d$, the use of the penalty $d \log(n)$ for $\text{IC}_{\tilde{H}(k),Q}$ and IC_Q leads to a significant improvement in correctly determining the true model (not presented). According to Theorem 3, this is not surprising, because the $2d$ penalty tends to pick larger models. For instance, because the true model in Section 3.1 has one covariate, the $d \log(n)$ penalty will be expected to outperform the $2d$ penalty (not presented). Furthermore, combining different degrees of approximation in the truncated Hermite expansion and different penalty terms can lead to nonlinear behavior in $\text{IC}_{\tilde{H}(k),Q}$ and IC_Q.

The MCEM algorithm converged in a reasonable number of steps for the GLM simulation and the AIDS data set, and the Gibbs sampling followed the same steps as described by Ibrahim, Lipsitz, and Chen (1999). In the Gibbs steps of the MCEM algorithm, the Metropolis–Hastings algorithm (Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller 1953; Hastings 1970) was used to simulate observations from the complex, nonstandard conditional distributions. For the GLM and AIDS data examples, EM convergence was obtained in fewer than 50 iterations using an increasing Gibbs sample size of 2,000 within EM. Gibbs sample sizes of 5,000 and 10,000 also were used to check sensitivity to the choice of the Gibbs sample size, and the estimates were extremely robust to these choices; for example, the estimates based on Gibbs sample sizes of 2,000 and 10,000 matched to the third decimal place. In addition, values of the Gibbs sample size that changed with each EM iteration were considered. For example, at the beginning of EM, we started with 50 Gibbs samples and gradually increased the number of Gibbs samples as the EM iterations increased. The results obtained were quite similar to those obtained using a constant value of 2,000 Gibbs iterations throughout all of the EM iterations. The convergence criterion used for the EM algorithm was that the distance between the *k*th iteration and the (*k* + 1)st iteration for all of the parameters was less than 5×10^{-4}. The reason for choosing such a tolerance level is the Gibbs sample size used in each iteration. We also tried a tolerance level of 10^{-4} when the Gibbs sample size was 10,000, and EM convergence was obtained in a similar number of iterations. We further note that if the tolerance level were chosen too small, then it would be impossible to achieve convergence due to the Monte Carlo error induced by the Gibbs sampler. Finally, we note that slightly more computing time was required for the AIDS data set than the GLM simulation.

Acknowledgments

The authors wish to deeply thank the editor, the associate editor, and three referees for extremely helpful comments and suggestions that have substantially improved the article. Dr. Ibrahim's research was supported in part by National Institutes of Health grants GM 70335 and CA 74015. Dr. Zhu's research was supported in part by National Science Foundation grant SES-06-43663 and BCS-0826844 and NIH grant 1-UL1-RR025747-01. Dr. Tang's research was supported in part by NSFC (10561008) and NCET (NCET-07-0737).

References

Akaike, H. Information Theory and an Extension of the Maximum Likelihood Principle. In: Petrov, BN.; Czáki, F., editors. Second International Symposium on Information Theory. Budapest: Akademiai Kiadó; 1973. p. 267-281.

Andrews DWK. Generic Uniform Convergence. Econometric Theory 1992;8:241–257.

Cameron AC, Johansson P. Count Data Regression Using Series Expansions, With Applications. Journal of Applied Econometrics 1997;12:203–223.

Chen MH, Ibrahim JG, Shao QM. Propriety of the Posterior Distribution and Existence of the Maximum Likelihood Estimator for Regression Models With Covariates Missing at Random. Journal of the American Statistical Association 2004;99:421–438.

Copas JB, Li HG. Inference for Non-Random Samples. (with discussion). Journal of the Royal Statistical Society, Ser B 1997;59:55–96.

Dempster AP, Laird NM, Rubin DB. Maximum Likelihood for Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society, Ser B 1977;39:1–38.

Diggle PJ, Kenward MG. Informative Drop-Out in Longitudinal Data Analysis. Applied Statistics 1994;43:49–93.

Fenton VM, Gallant AR. Qualitative and Asymptotic Performance of SNP Density Estimators. Journal of Econometrics 1996;74:77–118.

Gallant AR, Douglas WN. Semi-Nonparametric Maximum Likelihood Estimation. Econometrica 1987;55:363–390.

Gallant AR, Nychka DW. Semi-Nonparametric Maximum Likelihood Estimation. Econometrica 1987;55:363–390.

Hastings WK. Monte Carlo Sampling Methods Using Markov Chains and Their Application. Biometrika 1970;57:97–109.

Huang L, Chen MH, Ibrahim JG. Bayesian Analysis for Generalized Linear Models With Nonignorable Missing Covariates. Biometrics 2005;61:729–737. [PubMed: 16135024]

Ibrahim JG. Incomplete Data in Generalized Linear Models. Journal of the American Statistical Association 1990;85:765–769.

Ibrahim JG, Lipsitz SR. Parameter Estimation From Incomplete Data in Binomial Regression When the Missing-Data Mechanism is Nonignorable. Biometrics 1996;52:1071–1078. [PubMed: 8805768]

Ibrahim JG, Chen M-H, Lipsitz SR. Monte Carlo EM for Missing Covariates in Parametric Regression Models. Biometrics 1999;55:591–596. [PubMed: 11318219]

Ibrahim JG, Chen M-H, Lipsitz SR. Missing Responses in Generalised Linear Mixed Models When the Missing Data Mechanism Is Nonignorable. Biometrika 2001;88:551–564.

Ibrahim JG, Lipsitz SR, Chen MH. Missing Covariates in Generalized Linear Models When the Missing-Data Mechanism Is Nonignorable. Journal of the Royal Statistical Society, Ser B 1999;61:173–190.

Jansen I, Molenberghs G, Aerts M, Thjis H, van Steen K. A Local Influence Approach to Binary Data From a Psychiatric Study. Biometrics 2003;59:410–419. [PubMed: 12926726]

Kim JI. Uniform Convergence Rate of the Seminonparametric Density Estimator and Testing for Similarity of Two Unknown Densities. Econometrics Journal 2007;10:1–34.

Konishi, S.; Kitagawa, G. Information Criteria and Statistical Modeling. New York: Springer; 2008.

Lee SY, Tang NS. Analysis of Nonlinear Structural Equation Models With Nonignorable Missing Covariates and Ordered Categorical Data. Statistica Sinica 2006;16:1117–1141.

Little RJA. Pattern-Mixture Models for Multivariate Incomplete Data. Journal of the American Statistical Association 1993;88:125–134.

Little RJA. A Class of Pattern-Mixture Models for Normal Incomplete Data. Biometrika 1994;81:471–483.

Little RJA. Modeling the Drop-Out Mechanism in Repeated-Measures Studies. Journal of the American Statistical Association 1995;90:1112–1121.

Little, RJA.; Rubin, DB. Statistical Analysis With Missing Data. Vol. 2. Hoboken, NJ: Wiley; 2002.

Liu CH, Rubin DB. The ECME Algorithm: A Simple Extension of EM and ECM With Fast Monotone Convergence. Biometrika 1994;81:633–648.

Macquarrie, ADR.; Tsai, CL. Regression and Time Series Model Selection. River Edge, NJ: World Scientific; 1998.

Meng XL, Rubin DB. Maximum Likelihood Estimation via the ECM Algorithm: A General Framework. Biometrika 1993;80:267–278.

Meng XL, van Dyk D. The EM Algorithm: An Old Folk Song Sung to a Fast New Tune. Journal of the Royal Statistical Society, Ser B 1997;59:511–540.

Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. Equations of State Calculations by Fast Computing Machine. Journal of Chemical Physics 1953;21:1087–1091.

Morisky DE, Tiglao TV, Sneed CD, Tempongko SB, Baltazar JC, Detels R, Stein JA. The Effects of Establishment Practices, Knowledge, and Attitudes on Condom Use Among Filipina Sex Workers. AIDS Care 1998;10:213–320. [PubMed: 9625904]

Nishii R. Maximum Likelihood Principle and Model Selection When the True Model Is Unspecified. Journal of Multivariate Analysis 1988;27:392–403.

Rubin DB. Formalizing Subjective Notions About the Effect of Non-respondents in Sample Surveys. Journal of the American Statistical Association 1977;72:538–543.

Schwarz G. Estimating the Dimension of a Model. The Annals of Statistics 1978;6:461–464.

Troxel AB, Ma G, Heitjan DF. An Index of Local Sensitivity to Nonignorability. Statistica Sinica 2004;14:1221–1237.

van Steen K, Molenberghs G, Thijs H. A Local Influence Approach to Sensitivity Analysis of Incomplete Longitudinal Ordinal Data. Statistical Modelling: An International Journal 2001;1:125–142.

Verbeke G, Molenberghs G, Thijs H, Lasaffre E, Kenward MG. Sensitivity Analysis for Non-Random Dropout: A Local Influence Approach. Biometrics 2001;57:43–50. [PubMed: 11252617]

White, H. Estimation, Inference, and Specification Analysis. New York: Cambridge University Press; 1994.

Zhu HT, Zhang HP. Asymptotics for Estimation and Testing Procedures Under Loss of Identifiability. Journal of Multivariate Analysis 2006;97:19–45.

Zhu HT, Lee SY, Wei BC, Zhou J. Case-Deletion Measures for Models With Incomplete Data. Biometrika 2001;88:727–737.

Appendix

APPENDIX A: PROOFS OF THEOREMS 1, 2, AND 3

Proof of Theorem 1

We need only show that $\sup_{(\theta_1,\ \theta_2)\in\Theta\times\Theta} n^{-1}|\tilde{K}_{(k)}(\theta_1, \theta_2) - E[\tilde{K}_{(k)}(\theta_1, \theta_2)]| \to 0$ in probability and $E[\tilde{K}_{(k)}(\theta_1, \theta_2)]$ is continuous in θ_1 and θ_2 uniformly over $\Theta \times \Theta$. Conditions (C3) and (C4) are sufficient for assumption W–LIP of Andrews (1992), which ensures the continuity of $E[\tilde{K}_{(k)}(\theta_1, \theta_2)]$ and the stochastic equicontinuity (SE) of $\tilde{K}_{(k)}(\theta_1, \theta_2)$. Furthermore, conditions (C3) and (C4) ensure pointwise convergence; that is, $n^{-}{}_{1}\{\tilde{K}_{(k)}(\theta_1, \theta_2) - E[\tilde{K}_{(k)}(\theta_1, \theta_2)]\}$ converges to 0 for each θ_1 and θ_2 in probability. Thus combining SE and the pointwise convergence yields Theorem 1.

Proof of Theorem 2

We prove Theorem 2 in three steps. We first show that

$$\sqrt{n}(\widehat{\theta} - \theta_*) = A(\theta_*)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\partial_\theta \ell(z_{i,obs};\theta_*) + o_p(1). \tag{A.1}$$

Conditions (C1)–(C5) are sufficient for establishing (A.1) (Zhu and Zhang 2006). The second step is to obtain the stochastic expansions for $\tilde{K}_{(k)}(\hat{\theta}, \theta^\star)$ and $E[\tilde{K}_{(k)}(\hat{\theta}, \theta^\star)]$ as follows:

$$\begin{aligned}
&\tilde{K}_{(k)}(\widehat{\theta}, \theta^\star) = \tilde{K}_{(k)}(\theta_*, \theta^\star) + [\partial_\theta \tilde{K}_{(k)}(\theta_*, \theta^\star)]^T \Delta\widehat{\theta} \\
&+ \Delta\widehat{\theta}^T \partial_\theta^2 \tilde{K}_{(k)}(\theta_*, \theta^\star)\Delta\widehat{\theta} + o_p(1), \\
&E[\tilde{K}_{(k)}(\widehat{\theta}, \theta^\star)] = E[\tilde{K}_{(k)}(\theta_*, \theta^\star)] + E[\partial_\theta \tilde{K}_{(k)}(\theta_*, \theta^\star)]^T \Delta\widehat{\theta} \\
&+ \Delta\widehat{\theta}^T E[\partial_\theta^2 \tilde{K}_{(k)}(\theta_*, \theta^\star)]\Delta\widehat{\theta} + o_p(1),
\end{aligned} \tag{A.2}$$

where $\Delta\hat{\theta} = \hat{\theta} - \theta_*$. Taking expectation yields

$$\begin{aligned}
&E_{D_{obs}}[\tilde{K}_{(k)}(\widehat{\theta}, \theta^\star)] \\
&= E_{D_{obs}}[\tilde{K}_{(k)}(\theta_*, \theta^\star)] + E_{D_{obs}}\{[\partial_\theta \tilde{K}_{(k)}(\theta_*, \theta^\star)]^T \Delta\widehat{\theta}\} \\
&+ E_{D_{obs}}[\Delta\widehat{\theta}^T \partial_\theta^2 \tilde{K}_{(k)}(\theta_*, \theta^\star)\Delta\widehat{\theta}] + o(1), \\
&E_{D_{obs}}\{E[\tilde{K}_{(k)}(\widehat{\theta}, \theta^\star)]\} \\
&= E_{D_{obs}}[\tilde{K}_{(k)}(\theta_*, \theta^\star)] + E[\partial_\theta \tilde{K}_{(k)}(\theta_*, \theta^\star)]^T E_{D_{obs}}[\Delta\widehat{\theta}] \\
&+ E_{D_{obs}}\{\Delta\widehat{\theta}^T E[\partial_\theta^2 \tilde{K}_{(k)}(\theta_*, \theta^\star)]\Delta\widehat{\theta}\} + o(1).
\end{aligned} \tag{A.3}$$

Following the same arguments as Konishi and Kitagawa (2008), we can get

$$\begin{aligned}
&E_{D_{obs}}\{\tilde{K}_{(k)}(\widehat{\theta}, \theta^\star) - E[\tilde{K}_{(k)}(\widehat{\theta}, \theta^\star)]\} \\
&= E_{D_{obs}}\{[\partial_\theta \tilde{K}_{(k)}(\theta_*, \theta^\star)]^T (\widehat{\theta} - \theta_*)\} + o(1) \\
&= \mathrm{tr}\{A(\theta_*)^{-1} B(\theta_* | \theta^\star)\} + o(1).
\end{aligned} \tag{A.4}$$

Proof of Theorem 3

Based on Theorem 1 and $\delta_{c21} = o_p(n)$, we have

$$n^{-1}\mathrm{dIC}_{\tilde{H}(k),Q21} = 2n^{-1}\delta_{Q21} + o_p(1),$$

which yields Theorem 3a.

Theorem 3b can be proved by noting that $n^{-1/2}\mathrm{dIC}_{\tilde{H}(k),Q21}$ can be written as the sum of

$$\begin{aligned}
& n^{-1/2}\delta_{c21}, \\
& -2n^{-1/2}\{E[Q(\theta_{*(2)}|\widehat{\theta}_{(2)}) - \tilde{H}(k|\widehat{\theta}_{(2)})] \\
& -E[Q(\theta_{*(1)}|\widehat{\theta}_{(1)}) - \tilde{H}(k|\widehat{\theta}_{(2)})]\}, \\
& -2n^{-1/2}\{Q(\widehat{\theta}_{(2)}|\widehat{\theta}_{(2)}) - E[Q(\theta_{*(2)}|\widehat{\theta}_{(2)})]\} \\
& +2n^{-1/2}\{Q(\widehat{\theta}_{(1)}|\widehat{\theta}_{(1)}) - E[Q(\theta_{*(1)}|\widehat{\theta}_{(1)})]\}, \quad \text{and} \\
& 2n^{-1/2}\{\tilde{H}(k|\widehat{\theta}_{(1)}) - E[\tilde{H}(k|\widehat{\theta}_{(1)})]\} \\
& -2n^{-1/2}\{\tilde{H}(k|\widehat{\theta}_{(2)}) - E[\tilde{H}(k|\widehat{\theta}_{(2)})]\}.
\end{aligned}$$

Note that for $t = 1, 2$, $Q(\hat{\theta}_{(t)}|\hat{\theta}_{(t)})$ can be written as

$$\begin{aligned}
& Q(\theta_{*(t)}|\widehat{\theta}_{(t)}) \\
& +.5(\widehat{\theta}_{(t)} - \theta_{*(t)})^T \partial_\theta^2 Q(\theta_{*(t)}|\widehat{\theta}_{(t)})(\widehat{\theta}_{(t)} - \theta_{*(t)})[1 + o_p(1)].
\end{aligned}$$

Because $\hat{\theta}_{(t)} - \theta_{*(t)} = O_p(n^{-1/2})$, $Q(\hat{\theta}_{(t)}|\hat{\theta}_{(t)}) = Q(\theta_{*(t)}|\hat{\theta}_{(t)}) + O_p(1)$. Thus $\mathrm{dIC}_{\tilde{H}(k),Q21}$ can be written as

$$2Q(\theta_{*(1)}|\widehat{\theta}_{(1)}) - 2Q(\theta_{*(2)}|\widehat{\theta}_{(2)}) + \delta_{c21} + O_p(1).$$

Theorem 3c can be proved by noting that $Q(\theta_{*(1)}|\hat{\theta}_{(1)}) - Q(\theta_{*(2)}|\hat{\theta}_{(2)} = O_p(1)$ and $\delta_{c21} \to \infty$

APPENDIX B: SELECTED ITEMS IN THE AIDS DATA

The number of the variables in the questionnaire is given in parentheses.

y_1 (item 33): How worried are you about getting AIDS? not at all worried 1/2/3/4/5 extremely worried.

y_2 (item 32): What are the chances that you yourself might get AIDS?

none 1/2/3/4/5 very great.

y_3 (item 31): How much of a threat do you think AIDS is to the health of people?

no threat at all 1/2/3/4/5 very great.

y_4 (item 43): How many times did you have vaginal sex in the last 7 days?

y_5 (item 72): How many "hand jobs" did you give in the last 7 days?

y_6 (item 74): How many "blow jobs" did you give in the last 7 days? How great is the risk of getting AIDS from the following activities.

y_7 (item 27h): Sexual intercourse with someone you don't know very well without using a condom.

y_8 (item 27e): Sexual intercourse with someone who has the AIDS virus using a condom?

y_9 (item 27i): Sexual intercourse with someone who injects drugs? The scale for y_7, y_8, and y_9 is: no risk 1/2/3/4/5 great risk.

x_1 (item 37): How long (in months) have you been working at a job where people pay to have sex with you?

x_2 (item 21): How much do you think you know about the disease called AIDS?

nothing 1/2/3/4/5 a great deal.

Table 1

Comparison of ranks of the true model M1 from various model selection criteria for MAR covariates in linear models

	(I)												**(II)**											
	***n* = 100**				***n* = 300**				***n* = 500**				***n* = 100**				***n* = 300**				***n* = 500**			
Rank	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
$\mathrm{IC}_{H,Q}$ with 2*d*																								
1	331	25	4	0	329	17	1	0	325	14	1	0	301	51	13	0	325	37	4	0	327	32	6	0
2	1	48	8	0	5	57	12	0	4	63	10	0	5	31	21	4	6	35	17	5	3	35	15	3
3	0	1	49	3	1	3	53	2	0	3	50	3	0	9	35	8	0	2	36	9	0	7	36	8
4	0	0	1	29	0	0	0	20	0	0	2	25	0	0	1	21	0	0	2	22	0	1	1	26
$\mathrm{IC}_{\tilde{H}(0),Q}$ with 2*d*																								
1	329	23	5	0	328	16	3	0	322	14	0	0	302	49	10	0	323	35	5	0	320	31	10	0
2	3	50	8	1	6	57	12	0	7	59	10	1	4	34	25	5	8	36	15	5	10	34	16	1
3	0	1	48	1	1	4	51	2	0	7	50	3	0	8	34	9	0	3	37	9	0	8	31	10
4	0	0	1	30	0	0	0	20	0	0	3	24	0	0	1	19	0	0	2	22	0	2	1	26
$\mathrm{IC}_{\tilde{H}(1),Q}$ with 2*d*																								
1	325	25	5	0	323	19	2	0	310	21	1	0	298	49	11	0	314	33	6	0	301	27	6	1
2	7	46	9	1	9	54	13	0	18	49	11	1	8	34	22	3	15	37	16	5	23	37	17	3
3	0	3	47	2	3	3	50	3	1	10	49	4	0	7	36	8	2	4	35	10	6	9	33	6
4	0	0	1	29	0	1	1	19	0	0	2	23	0	1	1	22	0	0	2	21	0	2	2	27

NOTE: Two cases of missing fractions for x_i were included. Three different sample sizes, *n* = 100, 300, and 500 simulated data sets, were used for each case. The columns represent the results from AIC_Q.

Table 2

Comparison of ranks of the true model M1 from various model selection criteria for MAR covariates in GMls

	(I)															**(II)**														
	***n* = 100**					***n* = 300**					***n* = 500**					***n* = 100**					***n* = 300**					***n* = 500**				
Rk	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**
IC $_{\tilde{H}\,(0),Q}$ with 2*d*																														
1	294	22	2	1	0	317	29	2	1	0	316	32	6	0	0	267	34	12	1	0	298	46	6	2	0	305	47	12	4	0
2	8	70	17	1	0	15	53	13	5	0	8	50	19	0	0	12	62	29	5	0	8	37	23	3	0	6	29	36	3	0
3	0	2	61	4	0	0	1	51	11	0	0	0	44	9	0	0	2	47	8	0	0	5	49	8	0	0	6	33	6	0
4	0	0	2	13	2	0	0	2	0	0	0	0	1	15	0	0	0	2	15	2	0	0	2	13	0	0	2	1	10	0
5	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0
IC $_{\tilde{H}\,(1),Q}$ with 2*d*																														
1	290	24	2	1	0	307	26	3	0	0	296	32	5	1	0	265	35	13	1	0	280	37	9	3	0	269	45	14	3	0
2	11	68	13	1	0	22	49	12	5	0	26	44	17	1	0	10	60	21	4	1	24	41	14	3	0	36	27	27	6	0
3	1	2	63	2	0	3	8	47	11	0	1	4	44	11	0	4	3	52	11	0	2	10	50	10	0	5	10	38	6	0
4	0	0	4	14	2	0	0	6	0	0	1	2	4	11	0	0	0	4	11	1	0	0	7	10	0	1	2	3	8	0
5	0	0	0	2	0	0	0	0	12	0	0	0	0	0	0	0	0	1	12	1	0	0	0	0	0	0	0	0	0	0

NOTE: Two cases of missing fractions for x_i were included. Three different sample sizes n = 100, 300, and 500 simulated data sets were used for each case. The columns represent the results from AIC$_Q$.

Table 3

AIDS data: Values of IC_Q and $\mathrm{IC}_{\tilde{H}\,(k),Q}$ for $k = 0$ and 1 with $\hat{c}_n(\hat{\theta}) = 2d$ and $d \log(n)$ for all four models M_0, M_1, M_2, and M_3

	IC_Q		$\mathrm{IC}_{\tilde{H}(0),Q}$		$\mathrm{IC}_{\tilde{H}(1),Q}$	
Model	$\hat{c}_n = 2d$	$\hat{c}_n = d\log(n)$	$\hat{c}_n = 2d$	$\hat{c}_n = d\log(n)$	$\hat{c}_n = 2d$	$\hat{c}_n = d\log(n)$
M_0	34,676.19	34,896.96	32,941.28	30,985.59	35,423.52	33,467.84
M_1	34,680.18	34,905.97	32,961.77	31,017.56	35,709.52	33,765.32
M_2	34,689.32	34,915.11	32,964.85	31,014.59	35,626.51	33,676.26
M_3	34,708.79	34,939.60	32,988.38	31,037.17	35,567.39	33,616.17

Table 4

AIDS data: Selected ML estimates and their standard deviations (SDs) for model M_0

Parameter	ML estimates	SD	Parameter	ML estimates	SD	Parameter	ML estimates	SD
μ_1	3.6362	.0292	ψ_1	.9405	.0765	λ_{21}	.4493	.1124
μ_2	2.5977	.0432	ψ_2	2.2057	.0931	λ_{31}	.7736	.1558
μ_3	3.9725	.0321	ψ_3	.9525	.0464	λ_{52}	1.6294	.1679
μ_4	.0015	.0052	ψ_4	.8665	.0383	λ_{62}	1.1107	.0859
μ_5	.0031	.0323	ψ_5	.6246	.1358	λ_{83}	.4220	.1407
μ_6	.0020	.0092	ψ_6	.8251	.0452	λ_{93}	.7358	.1149
μ_7	4.3696	.0038	ψ_7	.7179	.0783	b_1	−.0579	.0310
μ_8	3.1411	.0431	ψ_8	2.0665	.0900	b_2	.0821	.0290
μ_9	3.7998	.0344	ψ_9	1.4165	.0865	γ_1	−.2711	.0679
φ_{11}	.1410	.0210	ψ_δ	.4059	.0912	γ_2	.2505	.1060
φ_{12}	−.0422	.0090	ψ_x	1.4774	.6778	γ_3	.1897	.1363
φ_{22}	.3819	.0418						