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

Joseph G. IBRAHIM, Hongtu ZHU, and Niansheng TANG

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 IC\(_{H,\tilde{Q}}\), which yields the Akaike information criterion and the Bayesian information criterion as special cases. The computation of IC\(_{H,\tilde{Q}}\) 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 IC\(_{H(k),\tilde{Q}}\). Theoretical properties of IC\(_{H(k),\tilde{Q}}\), including consistency, are investigated in detail. To eliminate the analytic approximation to the H-function, a computationally simpler approximation to IC\(_{H,\tilde{Q}}\) called IC\(_{\tilde{Q}}\) is proposed, the computation of which depends solely on the Q-function of the EM algorithm. Advantages and disadvantages of IC\(_{H(k),\tilde{Q}}\) and IC\(_{\tilde{Q}}\) 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 IC\(_{H(k),\tilde{Q}}\) and IC\(_{\tilde{Q}}\) in missing-data problems. An AIDS data set also is presented to illustrate the proposed methodology.

KEY WORDS: 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, Thijs, 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 IC\(_{H,\tilde{Q}}\), 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, IC\(_{H(k),\tilde{Q}}\), that depend solely on the EM output. We give a theoretical justification for IC\(_{H(k),\tilde{Q}}\) and derive the asymptotic properties of IC\(_{H(k),\tilde{Q}}\). We also consider another class of model selection criteria, IC\(_{\tilde{Q}}\), which use only the Q-function in their construction and thus omit the H-function in their construction. We also show that compared with IC\(_{H(k),\tilde{Q}}\), IC\(_{\tilde{Q}}\) is an inferior approximation to IC\(_{H,\tilde{Q}}\), 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 IC\(_{H,\tilde{Q}}\), IC\(_{H(k),\tilde{Q}}\), and IC\(_{\tilde{Q}}\). We present three theorems characterizing consistency and asymptotic properties of IC\(_{H(k),\tilde{Q}}\) 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 IC\_H,\_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 $D_{\text{obs}} = (z_{1,\text{obs}}, \ldots, z_{n,\text{obs}})$, the missing data $D_{\text{mis}} = (z_{1,\text{mis}}, \ldots, z_{n,\text{mis}})$, and the complete data $D_{\text{com}} = (z_{1,\text{com}}, \ldots, z_{n,\text{com}})$, in which $z_{i,\text{com}} = (z_{i,\text{mis}}, z_{i,\text{obs}})$ for $i = 1, \ldots, n$. The ITID model assumes that $z_{i,\text{com}}$ and $z_{j,\text{com}}$ are independent for $i \neq j$. Moreover, the dimensions of $z_{i,\text{mis}}$ and $z_{i,\text{obs}}$ 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 Lipsitz 1999, 2001).

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

$$g(D_{\text{obs}}; \theta) = \prod_{i=1}^{n} g(z_{i,\text{obs}}; \theta),$$

$$f(D_{\text{obs}}) = \prod_{i=1}^{n} f(z_{i,\text{obs}}),$$

$$g(D_{\text{com}}; \theta) = \prod_{i=1}^{n} g(z_{i,\text{com}}; \theta),$$

$$f(D_{\text{com}}) = \prod_{i=1}^{n} f(z_{i,\text{com}}),$$

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

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 $s$th step of the EM algorithm, given $\theta^{(s)}$, the E-step involves evaluating the $Q$-function given by

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

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

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

where

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

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: $\hat{\theta}$, $Q(\hat{\theta}|\theta^{(s)})$, and samples drawn from $g(D_{\text{mis}}|D_{\text{obs}}; \hat{\theta})$. 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 $g(D_{\text{obs}}; \theta)$. However, some missing-data problems have very complicated observed data likelihood functions, for which $g(D_{\text{obs}}; \theta)$ has no closed form, so that its direct evaluation is not computationally feasible or computationally accurate. Because

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

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

$$IC_{H, Q} = -2 \log g(D_{\text{obs}}; \hat{\theta}) + \hat{c}_n(\hat{\theta})$$

$$= -2 Q(\hat{\theta}|\hat{\theta}) + 2 H(\hat{\theta}|\hat{\theta}) + \hat{c}_n(\hat{\theta}),$$

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 $\theta$, we obtain the AIC of Akaike (1973), given by $AIC = -2 \log g(D_{\text{obs}}; \hat{\theta}) + 2d$. 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 $g(D_{\text{mis}}|D_{\text{obs}}; \theta)$ 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 $g(D_{\text{mis}}|D_{\text{obs}}; \theta)$ first needs an analytic approximation to allow computation of the $H$-function through the EM output. Once $g(D_{\text{mis}}|D_{\text{obs}}; \theta)$ is analytically approximated (i.e., the integrand of the $H$-function is analytically approximated), the
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 $w(D_{mis}) = \log g(D_{mis}|D_{obs}; \hat{\theta})$ does not have a closed form for most missing-data problems.

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

$$\tilde{g}(zi, mis; \hat{\mu}_i, \hat{\Sigma}_i, \psi_i, k) = P_t(t; \psi_i, k)\phi(zi, mis; \hat{\mu}_i, \hat{\Sigma}_i),$$

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

$$P_t(t; \psi_i, 2) = \frac{1 + \psi_1 t + \psi_2 t^2}{\sqrt{1 + \psi_1^2 + 3\psi_2^2 + 2\psi_2}}.$$  

If $k = 0$, then we obtain $P_t(t; \psi_i, 0) = 1$ and $g(zi, mis; \hat{\mu}_i, \hat{\Sigma}_i, \psi_i, k)$ can be replaced by another density, such as a multivariate $t$, Poisson, or gamma density (Cameron and Johansson 1997; Kim 2007).

We can use $\tilde{g}(zi, mis; \hat{\mu}_i, \hat{\Sigma}_i, \psi_i, k)$ 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, $\{zi, mis^{(s)}: s = 1, \ldots, S_0\}$, from $g(zi, mis|zi, obs; \hat{\theta})$ using MCMC sampling, where $S_0$ is a prefixed number. In step 2 we use the sample mean and covariance matrix of $\{zi, mis^{(s)}: s = 1, \ldots, S_0\}$ to approximate $\hat{\mu}_i$ and $\hat{\Sigma}_i$. In step 3, because $\{zi, mis^{(s)}: s = 1, \ldots, S_0\}$ are observations from $g(zi, mis|zi, obs; \hat{\theta})$, we can then obtain estimators (e.g., ML estimators) of $\psi_i$, 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 $g(zi, mis; \hat{\mu}_i, \hat{\Sigma}_i, \hat{\psi}_i(k), k)$ and $g(zi, mis|zi, obs; \hat{\theta})$, that is,

$$\hat{\psi}_i(k) = \arg \min_{\psi_i} \left\{ \int \log \frac{g(zi, mis|zi, obs; \hat{\theta})}{g(zi, mis; \hat{\mu}_i, \hat{\Sigma}_i, \psi_i(k), k)} \times g(zi, mis|zi, obs; \hat{\theta}) dzi, mis \right\}.$$  

In step 4 we calculate

$$\tilde{H}(k|\hat{\theta}) = S_0^{-1} \sum_{i=1}^{S_0} \sum_{\psi_i(k)} \log g(zi, mis^{(s)}; \hat{\mu}_i, \hat{\Sigma}_i, \hat{\psi}_i(k), k)$$

$$= E \left[ \log \tilde{g}(D_{mis}|D_{obs}; k, \hat{\theta}) \right] + o(1),$$

where $\tilde{g}(D_{mis}|D_{obs}; k, \hat{\theta}) = \prod_{i=1}^{S_0} \tilde{g}(zi, mis; \hat{\mu}_i, \hat{\Sigma}_i, \hat{\psi}_i(k), k)$ 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 $\psi_i(k)$ for each $i$ can be computationally cumbersome when $k$ is relatively large. If we set $k = 0$ at then we can avoid the maximization in step 3.

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

$$IC_{H,Q} = -2Q(\hat{\theta}|\hat{\theta}) + 2\tilde{H}(k|\hat{\theta}) + 2H(\hat{\theta}|\hat{\theta}) - 2\tilde{H}(k|\hat{\theta}) + c_n(\hat{\theta}) \approx IC_{H,Q} = -2Q(\hat{\theta}|\hat{\theta}) + 2\tilde{H}(k|\hat{\theta}) + c_n(\hat{\theta}).$$

Moreover, because $\tilde{H}(k|\hat{\theta}) \leq H(\hat{\theta}|\hat{\theta})$ according to Jensen’s inequality, IC$_{H(Q),k} \leq IC_{H,Q}$. 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 $g(zi, mis|zi, obs; \hat{\theta})$ is quite attractive, because model choice is quite robust with respect to the choice of $k$.

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

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

$$g(k)(D_{obs}; \theta_1, \theta_2) = \exp \left\{ E \left[ \log g(D_{mis}; \theta_1) - \log g(D_{mis}|D_{obs}; k, \theta_1) | D_{obs}; \theta_2 \right] \right\}$$

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 $k \to \infty$, it can be shown that under some conditions, $g(D_{mis}|D_{obs}; k, \theta_1)$ converges to $g(D_{mis}|D_{obs}; \theta_1)$, and thus $g(k)(D_{obs}; \theta_1, \theta_2)$ converges to $g(D_{obs}; \theta_1)$. To develop a general class of model selection criteria, we consider the Kullback–Leibler divergence between $g(k)(D_{obs}; \theta_1, \theta_2)$ and $f(D_{obs})$, defined by

$$K(\theta_1, \theta_2) = - \int f(D_{obs}) \log f(D_{obs}) dD_{obs}$$

$$- \int f(D_{obs}) \log g(k)(D_{obs}; \theta_1, \theta_2) dD_{obs}. $$
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 \( K(\hat{\theta}, \theta^*). \) A simple estimator of \( K(\hat{\theta}, \theta^*) \) 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, \( K(\hat{\theta}, \theta^*) \) can be approximated by

\[
\tilde{K}(\hat{\theta}, \theta^*) = -\log \tilde{g}(k)(D_{obs}; \hat{\theta}, \theta^*). 
\]

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 mis-specified, the ML estimator, \( \hat{\theta} \), converges to the \( \theta_0 \) that minimizes

\[
E[\sum_{i=1}^n \ell(zi_{obs}; \theta)] = \sum_{i=1}^n \int \ell(zi_{obs}; \theta)f(zi_{obs})dzi_{obs}, \quad \ell(zi_{obs}; \theta) = \log \tilde{g}(zi_{obs}; \theta) 
\]

(see, e.g., White 1994). For simplicity, we further assume that \( \theta_0 = \theta_\star \) for all \( n \) and \( E[\theta_0 \ell(zi_{obs}; \theta_0)] = 0 \) for all \( i \).

The conditions are as follows:

\begin{enumerate}
\item [(C1)] \( \theta_0 \) is unique and an interior point of \( \Theta \), where \( \Theta \) is a compact set in \( \mathbb{R}^p \).
\item [(C2)] \( \hat{\theta} \to \theta_0 \) in probability as \( n \to \infty \).
\item [(C3)] For all \( i \), \( \ell(zi_{obs}; \theta) \) is three times continuously differentiable on \( \Theta \), and \( |\partial_i \ell(zi_{obs}; \theta)|^2 \) and \( |\partial_j \partial_i \ell(zi_{obs}; \theta)| \) are dominated by \( B_i(zi_{obs}) \) for all \( j, j', l = 1, 2, \ldots, d \), where \( \partial_j = \partial/\partial \theta_j \). The same smoothness condition also holds for \( h_i(zi_{obs}; \theta) = E[\log \tilde{g}(zi_{mis}; zi_{obs}; k, \theta) \mid zi_{obs}] \).
\item [(C4)] For each \( \epsilon > 0 \), there exists a finite \( C \) such that

\[
\sup_{n \geq 1} n^{-1} \sum_{i=1}^n E[B_i(zi_{obs})1[B_i(zi_{obs}) > C]] < \epsilon 
\]

for all \( n \), where \( 1[B_i(zi_{obs}) > C] \) is the indicator function of \( B_i(zi_{obs}) > C \).
\item [(C5)] \( \lim_{n \to \infty} n^{-1} E \left\{ -\sum_{i=1}^n \partial_\theta \ell(zi_{obs}; \theta_0) \right\} = A(\theta_0) 
\]

and

\[
\lim_{n \to \infty} n^{-1} \sum_{i=1}^n E[\partial_\theta \ell(zi_{obs}; \theta_0) \partial_\theta \tilde{K}(\hat{\theta}, \theta^*)] \left|_{\theta=\theta_0} \right. = B(\theta_0 | \theta^*), 
\]

where \( A(\theta_0) \) 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 \( \ell(zi_{obs}; \theta) \) and \( h_i(zi_{obs}; \theta) \). 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

\[
n^{-1}|\tilde{K}(\hat{\theta}, \theta^*) - E[\tilde{K}(\hat{\theta}, \theta^*)]| + n^{-1}|E[\tilde{K}(\hat{\theta}, \theta^*)] - E[\tilde{K}(\theta_0, \theta^*)]| \to 0 
\]

in probability, where \( E[\tilde{K}(\hat{\theta}, \theta^*)] \) denotes the expectation with respect to the observed data, \( E[\tilde{K}(\hat{\theta}, \theta^*)] \) denotes \( E[\tilde{K}(\hat{\theta}, \theta^*)] \) evaluated at \( \hat{\theta} = \hat{\theta} \), and \( \theta_0 \) is the pseudo true value of \( \theta \) based on \( g(D_{obs}; \theta) \).

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

\[
b(\theta^*) = E_{D_{obs}}[\tilde{K}(\hat{\theta}, \theta^*) - E[\tilde{K}(\hat{\theta}, \theta^*)]] 
\]

\[
= b_1(\theta^*) + O(1), 
\]

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 \( b(\theta^*) \), we can derive an asymptotic bias expression, denoted \( b_1(\theta^*) \).

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

\[
E_{D_{obs}}[\tilde{K}(\hat{\theta}, \theta^*) - E[\tilde{K}(\hat{\theta}, \theta^*)]] 
\]

\[
= b(\theta^*) 
\]

\[
= \text{tr}(A(\theta_0)^{-1}B(\theta_0 | \theta^*)) + O_p(1), 
\]

where \( A(\theta) \) and \( B(\theta | \theta^*) \) are defined in condition (C5).

Theorem 2 provides a theoretical basis for using \( -2\tilde{K}(\hat{\theta}, \theta^*) + b(\theta^*) \) as a model selection criterion, and this quantity is precisely a bias-corrected estimate of \( -2E_{D_{obs}}[\tilde{K}(\hat{\theta}, \theta^*)] \). In particular, if \( \theta^* = \theta_0 \) and \( g(D_{obs}; \theta) \) is specified correctly, then \( A(\theta_0) - B(\theta_0 | \theta_0) \) converges to a zero matrix and \( b(\theta_0) \approx 2d \) as \( k \to \infty \). But because \( \theta^* \) is unknown, we replace \( \theta_0 \) and \( \theta^* \) by \( \hat{\theta} \). In particular, under the correct specification of \( g(D_{obs}; \theta) \), \( b(\theta) \) should be close to \( 2d \) for large \( k \). This leads to an approximation to the AIC as \( \text{AIC}_{\hat{\theta}(k), Q} = -2\tilde{K}(\hat{\theta}, \theta) + 2d \).

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

\[
M_t = \{g(t)(D_{com}; \theta_0) = g(t)(D_{obs}; \theta_0)g(t)(D_{mis} | D_{obs}; \theta_0) : \theta_0 \in \Theta(t) \subset R^d \} 
\]

for \( t = 1, 2 \). For each \( M_t \), the ML estimator \( \hat{\theta}(t) \) converges in probability to the pseudo true value, denoted by \( \theta_0(t) \). To select
a better model, we first calculate
\[
d\text{IC}_{\tilde{H}(k), Q2} = \text{IC}_{\tilde{H}(k), Q2} - \text{IC}_{\tilde{H}(k), Q1} \\
= 2Q(\hat{\theta}(1)|\hat{\theta}(1)) - 2\tilde{H}(k|\hat{\theta}(1)) - \hat{c}_n(\hat{\theta}(1)) \\
- 2\tilde{Q}(\hat{\theta}(2)|\hat{\theta}(2)) + 2\tilde{H}(k|\hat{\theta}(2)) + \hat{c}_n(\hat{\theta}(2)).
\] 
We choose \(M_2\) if \(d\text{IC}_{\tilde{H}(k), Q2} < 0\) and \(M_1\) otherwise. Define
\[
\delta_{21,k} = E\{Q(\theta_{x(1)}|\theta_{x(1)}) - \tilde{H}(k|\theta_{x(1)})\} \\
- E\{Q(\theta_{x(2)}|\theta_{x(2)}) - \tilde{H}(k|\theta_{x(2)})\}
\]
and \(\delta_c = \hat{c}_n(\hat{\theta}(2)) - \hat{c}_n(\hat{\theta}(1))\). 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 \(\delta_{21} = (d_2 - d_1) \log(n)\).

Theorem 3. Suppose that \(M_1\) and \(M_2\) are ITID models and satisfy conditions (C1)–(C5). We then have the following results:

a. If \(\lim inf_n n^{-1}\delta_{21,k} > 0\) and \(\delta_{21} = \sigma_p(n)\), then \(d\text{IC}_{\tilde{H}(k), Q2} > 0\) in probability.

b. Assume that
\[
\lim sup_n n^{-1/2}\{E\{Q(\theta_{x(2)}|\theta_{x(2)}) - \tilde{H}(k|\theta_{x(2)})\} \\
- E\{Q(\theta_{x(1)}|\theta_{x(1)}) - \tilde{H}(k|\theta_{x(1)})\}\} < \infty,
\]
n\(-1/2\{H(\hat{\theta}(1)) - E[H(k|\hat{\theta}(1))]| = \sigma_p(1)\), and \(n^{-1/2} \times \{Q(\hat{\theta}(1)|\hat{\theta}(1)) - E[Q(\theta_{x(2)}|\theta_{x(2)})]\} = \sigma_p(1)\) for \(t = 1, 2\). Then \(d\text{IC}_{\tilde{H}(k), Q2} \leq 0\) in probability as \(n^{-1/2}\delta_{21} \to \infty\).

c. Assume that \(Q(\theta_{x(2)}|\theta_{x(2)}) - Q(\theta_{x(1)}|\theta_{x(1)}) = \sigma_p(1)\) and \(\tilde{H}(k|\hat{\theta}(2)) - \tilde{H}(k|\hat{\theta}(1)) = \sigma_p(1)\). Then \(d\text{IC}_{\tilde{H}(k), Q2} \leq 0\) in probability as \(\delta_{21} \to \infty\).

Theorem 3 has some important implications. Theorem 3a indicates that \(d\text{IC}_{\tilde{H}(k), Q}\) chooses \(M_2\) as \(\lim inf_n n^{-1}\delta_{21,k} > 0\) and \(\delta_{21} = \sigma_p(n)\). 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 \(\delta_{21} = \sigma_p(n)\) (Nishii 1988). The condition \(\lim inf_n n^{-1}\delta_{21,k} > 0\) ensures that \(d\text{IC}_{\tilde{H}(k), Q}\) chooses a model with large \(E[Q(\theta_{x} | \theta_{x}) - \tilde{H}(k|\theta_{x})]\). If \(M_1\) and \(M_2\) have the same average \(n^{-1}E[Q(\theta_{x} | \theta_{x}) - \tilde{H}(k|\theta_{x})]\) (i.e., \(\lim inf_n n^{-1} \times \delta_{21,k} = 0\)), then Theorem 3b and c indicate that \(d\text{IC}_{\tilde{H}(k), Q}\) picks out the “simpler” \(M_1\) when \(\delta_{21,k}\) increases to \(\infty\) at a certain rate \([e.g., \log(n)]\). But \(\hat{c}_n = 2d\) does not satisfy this condition. Thus, because \(d\text{IC}_{\tilde{H}(k), Q}\) 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 \(d\text{IC}_{\tilde{H}(k), Q}\) in the Presence of Nonignorable Missing Data

Although our model selection criteria \(d\text{IC}_{\tilde{H}(k), Q}\) 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 \(d\text{IC}_{\tilde{H}(k), Q}\):

1. \(d\text{IC}_{\tilde{H}(k), Q}\) 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 \(d\text{IC}_Q\)

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 \(\hat{H}(\hat{\theta}|\hat{\theta})\) from (7), leading to the criterion
\[
d\text{IC}_Q = -2Q(\hat{\theta}|\hat{\theta}) + \hat{c}_n(\hat{\theta}).
\] 
Thus \(d\text{IC}_Q\) can be viewed as a crude approximation to \(d\text{IC}_{\tilde{H}(k), Q}\) in which \(\hat{H}(\hat{\theta}|\hat{\theta})\) is omitted. When \(\hat{c}_n(\hat{\theta}) = 2d\) in (18), this leads to the criterion
\[
d\text{AIC}_Q = -2Q(\hat{\theta}|\hat{\theta}) + 2d.
\]

There are clear advantages and disadvantages to using \(d\text{IC}_Q\) instead of \(d\text{IC}_{\tilde{H}(k), Q}\). One advantage of using \(d\text{IC}_Q\) is that it is computationally easier than \(d\text{IC}_{\tilde{H}(k), Q}\), not requiring an approximation to the integrand of the \(H\)-function. But one clear disadvantage of \(d\text{IC}_Q\) 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 \(d\text{IC}_{\tilde{H}(k), Q}\) over \(d\text{IC}_Q\).

3. SIMULATION STUDIES

In this section we report on several simulation studies used to investigate the finite-sample performance of \(d\text{IC}_{\tilde{H}(k), Q}\) and \(d\text{IC}_Q\) in linear models and GLMs with MAR covariates. More specifically, we demonstrate how \(d\text{IC}_{\tilde{H}(k), Q}\) and \(d\text{IC}_Q\) 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), \(d\text{IC}_{\tilde{H}(k), Q}\) has an analytic closed form, and thus \(g(z_{i,\text{mis}}|z_{i,\text{obs}}, \hat{\theta})\) 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 \( \text{IC}_{H(k),Q}^{\hat{\theta}} \) to \( \text{IC}_{H,Q} \), which is analytically equivalent to AIC in this case when \( \hat{c}_n(\theta) = 2d \). We also compare IC\_Q with IC\_H,\_Q.

But for the GLM with MAR covariates neither IC\_H,\_Q nor \( g(z_i,\text{mis}|z_i,\text{obs}; \hat{\theta}) \) has a closed form, and thus both the Hermite approximation and MCMC sampling are needed to compute 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 \( \text{IC}_{H(k),Q}^{\hat{\theta}} \) for \( k = 0,1 \) and IC\_Q through the MCEM algorithm under several values of \( \hat{c}_n(\theta) \).

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 IC\_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 IC\_H,\_Q, (ii) to compare IC\_Q with IC\_M,\_Q, and (iii) to compare the performance of IC\_Q with IC\_H(k),Q. To save space, we focus on \( c_n(\theta) = 2d \) throughout, although several additional simulation results are available for other values of \( c_n(\theta) \) including \( \hat{c}_n(\theta) \).

Consider the true model \( y_i|x_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2) \), where \( x_i \sim \text{Normal} (\mu, \tau^2) \) for \( i = 1, \ldots, n \). We generated the data set \( \{x_i, y_i\} : i = 1, \ldots, n \) as follows. First, we generated \( n \) independent random variables \( x_i \) from a \( \text{Normal} (\mu, \tau^2) \) distribution; and then generated independent responses \( y_i \) from a \( \text{Normal} (\beta_0 + \beta_1 x_i, \sigma^2) \) 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 \( \beta_0 = 3.8, \beta_1 = 8, \sigma^2 = 8, \mu = 8, \tau^2 = 8 \), and \( n = 100, 300, 500 \).

Furthermore, we assume that the response \( y_i \) and the additional covariate \( z_i \) are completely observed for \( i = 1, \ldots, n \), 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:

\[
p(r_i = 1|y_i, z_i) = \frac{\exp(\phi_0 + \phi_1 y_i)}{1 + \exp(\phi_0 + \phi_1 y_i)},
\]

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 \( \phi_0 \) and \( \phi_1 \): (I) \( \phi_0 = -4.0, \phi_1 = 1.0 \) giving an average missingness fraction for \( x_i \) roughly equal to 11\%, and (II) \( \phi_0 = -3.5, \phi_1 = 1.5 \), giving an average missingness fraction of 29\%.

We considered five candidate models:

- Model M1 (true model): \( y_i|x_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2), x_i \sim \text{Normal} (\mu, \tau^2) \)
- Model M2: \( y_i|x_i \sim N(\beta_0, \sigma^2), x_i \sim \text{Normal} (\mu, \tau^2) \)
- Model M3: \( y_i|x_i, z_i \sim N(\beta_0 + \beta_1 x_i + \beta_2 z_i, \sigma^2), x_i \sim \text{Normal} (\mu, \tau^2) \)
- Model M4: \( y_i|x_i, z_i \sim N(\beta_0 + \beta_1 x_i + \beta_2 x_i z_i, \sigma^2), x_i \sim \text{Normal} (\mu, \tau^2) \)
- Model M5: \( y_i|x_i, z_i \sim N(\beta_0 + \beta_1 x_i + \beta_2 z_i + \beta_3 x_i z_i, \sigma^2), x_i \sim \text{Normal} (\mu, \tau^2) \)

We generated \( R = 500 \) simulated data sets from M1 and then calculated \( \text{IC}_{H(k),Q}^{\hat{\theta}} : k = 0,1 \) and IC\_Q for \( \hat{c}_n(\theta) = 2d \), and AIC = IC\_H,\_Q when \( \hat{c}_n(\theta) = 2d \) (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, AIC\_H, and IC\_H,\_Q.

### 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 \( \text{IC}_{H(k),Q}^{\hat{\theta}} : k = 0,1 \) 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 \( \text{Normal}(5,1.0) \) distribution, and responses \( y_1, \ldots, y_n \) are generated independently from a Bernoulli distribution with success probability \( p(y_i = 1|\beta_0, \beta_1, x_i) = \frac{\exp(\beta_0 + \beta_1 x_i)}{1+\exp(\beta_0 + \beta_1 x_i)} \) (see (20)). 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 \( \beta_0 = \beta_1 = -0.8 \) and \( n = 100, 300, 500 \). To investigate the effect of the missingness fraction on our model selection criteria, we again considered two sets of true values for \( \phi_0 \) and \( \phi_1 \): (I) \( \phi_0 = -1.2, \phi_1 = -0.8 \), giving a missingness fraction of about 15\%, and (II) \( \phi_0 = -0.5, \phi_1 = -0.8 \), giving a missingness fraction of about 26\%.
Table 1. Comparison of ranks of the true model M1 from various model selection criteria for MAR covariates in linear models

<table>
<thead>
<tr>
<th></th>
<th>(I)</th>
<th></th>
<th>(II)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n = 100</td>
<td>n = 300</td>
<td>n = 500</td>
<td>n = 100</td>
</tr>
<tr>
<td>Rank</td>
<td>1 2 3 4 5</td>
<td>1 2 3 4 5</td>
<td>1 2 3 4 5</td>
<td>1 2 3 4 5</td>
</tr>
<tr>
<td>IC_{H,Q} with 2d</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>331 25 4 0</td>
<td>329 17 1 0</td>
<td>325 14 1 0</td>
<td>301 51 13 0</td>
</tr>
<tr>
<td>2</td>
<td>1 48 8 0</td>
<td>5 57 12 0</td>
<td>4 63 10 0</td>
<td>5 31 21 4</td>
</tr>
<tr>
<td>3</td>
<td>0 1 49 3 2</td>
<td>1 3 53 2</td>
<td>0 3 50 3</td>
<td>0 9 35 8</td>
</tr>
<tr>
<td>4</td>
<td>0 0 0 1 29</td>
<td>0 0 0 20 0</td>
<td>0 0 2 25</td>
<td>0 0 1 21</td>
</tr>
<tr>
<td>IC_{H(0),Q} with 2d</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>329 23 5 0</td>
<td>328 16 3 0</td>
<td>322 14 0</td>
<td>302 49 10 0</td>
</tr>
<tr>
<td>2</td>
<td>3 50 8 1</td>
<td>6 57 12 0</td>
<td>7 59 10 1</td>
<td>4 34 25 5</td>
</tr>
<tr>
<td>3</td>
<td>0 1 48 1</td>
<td>1 4 51 2</td>
<td>0 7 50 3</td>
<td>0 8 34 9</td>
</tr>
<tr>
<td>4</td>
<td>0 0 0 1 30</td>
<td>0 0 0 20 0</td>
<td>0 0 3 24</td>
<td>0 0 1 19</td>
</tr>
<tr>
<td>IC_{H(1),Q} with 2d</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>325 25 5 0</td>
<td>323 19 2 0</td>
<td>310 21 1 0</td>
<td>298 49 11 0</td>
</tr>
<tr>
<td>2</td>
<td>7 46 9 1</td>
<td>9 54 13 0</td>
<td>18 49 11 1</td>
<td>8 34 22 3</td>
</tr>
<tr>
<td>3</td>
<td>0 3 47 2</td>
<td>3 3 50 3</td>
<td>1 10 49 4</td>
<td>0 7 36 8</td>
</tr>
<tr>
<td>4</td>
<td>0 0 0 1 29</td>
<td>0 0 1 1 9</td>
<td>0 0 2 23</td>
<td>0 1 1 22</td>
</tr>
</tbody>
</table>

As in Section 3.1, we considered five candidate models:

- Model M1 (true model): logit(p_i) = β_0 + β_1 x_i, x_i ~ N(μ, τ^2)
- Model M2: logit p_i = β_0, x_i ~ N(μ, τ^2)
- Model M3: logit(p_i) = β_0 + β_1 x_i, x_i ~ N(μ, τ^2)
- Model M4: logit(p_i) = β_0 + β_1 x_i + β_2 z_i, x_i ~ N(μ, τ^2)
- Model M5: logit(p_i) = β_0 + β_1 x_i + β_2 z_i, x_i = β_3 z_i, x_i ~ N(μ, τ^2)

We simulated 500 data sets and then calculated IC_{H(k),Q} : k = 0, 1 and IC_{H,Q} for ˆc_n(ˆθ) = 2d for each simulated data set. Table 2 shows the number of times out of 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 criterion. 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_{H(0),Q} with ˆc_n(ˆθ) = 2d, and 317 times by IC_{H(1),Q} with ˆc_n(ˆθ) = 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_{H(k),Q} : k = 0, 1| for ˆc_n(ˆθ) = 2d perform reasonably well even for large missingness fractions. Moreover, increasing k does not seem to improve the

Table 2. Comparison of ranks of the true model M1 from various model selection criteria for MAR covariates in GMIs

<table>
<thead>
<tr>
<th></th>
<th>(I)</th>
<th></th>
<th>(II)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n = 100</td>
<td>n = 300</td>
<td>n = 500</td>
<td>n = 100</td>
</tr>
<tr>
<td>Rk</td>
<td>1 2 3 4 5</td>
<td>1 2 3 4 5</td>
<td>1 2 3 4 5</td>
<td>1 2 3 4 5</td>
</tr>
<tr>
<td>IC_{H,Q} with 2d</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>294 22 2 1</td>
<td>317 29 2 1</td>
<td>316 32 6</td>
<td>267 34 12 1 0</td>
</tr>
<tr>
<td>2</td>
<td>8 70 17 1 0</td>
<td>15 53 13 3 5</td>
<td>8 50 19</td>
<td>12 62 29 5 0</td>
</tr>
<tr>
<td>3</td>
<td>0 2 61 4 0</td>
<td>0 1 51 11 0</td>
<td>0 0 44 9</td>
<td>0 2 47 8 0</td>
</tr>
<tr>
<td>4</td>
<td>0 0 0 2 13 2</td>
<td>0 0 2 0 0</td>
<td>0 0 1 15</td>
<td>0 0 2 15 2</td>
</tr>
<tr>
<td>5</td>
<td>0 0 0 0 0</td>
<td>0 0 0 0 0</td>
<td>0 0 0 0 0</td>
<td>0 0 0 0 0</td>
</tr>
<tr>
<td>IC_{H(1),Q} with 2d</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>290 24 2 1</td>
<td>307 26 3 0</td>
<td>296 32 5 1</td>
<td>265 35 13 1 0</td>
</tr>
<tr>
<td>2</td>
<td>11 68 13 1 0</td>
<td>22 49 12 5</td>
<td>26 44 17 1</td>
<td>10 60 21 4 1</td>
</tr>
<tr>
<td>3</td>
<td>1 2 63 2 0</td>
<td>3 8 47 11 0</td>
<td>1 4 44 11</td>
<td>4 3 52 11 0</td>
</tr>
<tr>
<td>4</td>
<td>0 0 4 14 2</td>
<td>0 0 6 0 0</td>
<td>1 2 4 11</td>
<td>0 0 4 11 1</td>
</tr>
<tr>
<td>5</td>
<td>0 0 0 2 0</td>
<td>0 0 0 12 0</td>
<td>0 0 0 0 0</td>
<td>0 0 1 12 1</td>
</tr>
</tbody>
</table>

NOTE: Two cases of missing fractions for s_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}. 
performance of $IC_{\mathcal{H}(k),Q}$. Again, the $\{IC_{\mathcal{H}(k),Q} : k = 0, 1\}$ 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 covariates 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$_{\mathcal{H}(k),Q}$ 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-efficacy 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 $y_i = (y_{i1}, \ldots, y_{i9})^T$, 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 $\xi_1$ and $\xi_2$, which can be interpreted as “aggressiveness of the sex worker” and “worry of contracting AIDS.” Specifically, to identify the relationship between the manifest variables $y_i$ and the latent variables $\omega_i = (\eta_i, \xi_1, \xi_2)^T$, we consider the following measurement equation:

$$y_i = \mu + \Lambda \omega_i + e_i,$$

where $\mu = (\mu_1, \ldots, \mu_9)^T$ is a vector of intercepts, $(\xi_1, \xi_2)$ is independent of the measurement error vector $e_i$, $(\xi_1, \xi_2) \sim N(0, \Phi)$, and $e_i \sim N(0, \Psi)$, in which $\Psi = \text{diag}(\Psi_1, \ldots, \Psi_9)$ and $\Phi = (\phi_{ij})$ is a $2 \times 2$ covariance matrix. We also assume the following structure for $\Lambda$:

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

where $0^*$ and $1.0^*$ are regarded as fixed values to identify the scale of the latent factor. We let $r_{ij} = 1$ if $y_{ij}$ is observed and $r_{ij} = 0$ if $y_{ij}$ is missing and $r_{x1i} = 0$ if $x_{1i}$ is missing and $r_{x1i} = 1$ if $x_{1i}$ 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 $x_{1i}$:

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

and

$$M_z: \quad \logit(Pr(r_{x1i} = 1|\varphi)) = \varphi_0 + \varphi_1 y_{1i} + \cdots + \varphi_9 y_{9i},$$

where $\tau$ is a vector of logistic regression coefficients, $y_{io}^* = (1, y_{io}^*)^T$ in which $y_{io}$ is a vector corresponding to the observed data of $y_i$, and $\varphi = (\varphi_0, \varphi_1, \ldots, \varphi_9)^T$. Because $x_{1i}$ may be missing, we need to specify its distribution. For simplicity, we assume that $x_{1i} \sim N(0, \Psi_{x})$.

To study the relationship between $\eta$ and $(x_{1i}, x_{2i}, \xi_1, \xi_2)$, we consider four nonlinear structural equations models, $M_0: \quad \eta_i = b_1 x_{1i} + b_2 x_{2i} + \gamma_1 \xi_1 + \gamma_2 \xi_2 + \gamma_3 \xi_1^2 + \xi_2^2 + \delta_i; \quad M_1: \quad \eta_i = b_1 x_{1i} + b_2 x_{2i} + \gamma_1 \xi_1 + \gamma_2 \xi_2 + \gamma_3 \xi_1 \xi_2 + \delta_i; \quad M_2: \quad \eta_i = b_1 x_{1i} + b_2 x_{2i} + \gamma_1 \xi_1 + \gamma_2 \xi_2 + \gamma_3 \xi_1^2 + \gamma_4 \xi_2^2 + \delta_i; \quad M_3: \quad \eta_i = b_1 x_{1i} + b_2 x_{2i} + \gamma_1 \xi_1 + \gamma_2 \xi_2 + \gamma_3 \xi_1^2 + \gamma_4 \xi_2^2 + \gamma_5 \xi_1 \xi_2 + \gamma_6 \xi_1 \xi_2^2 + \delta_i,$

where $\delta_i \sim N(0, \Psi_{x})$. Clearly, all four models include the linear effect of “aggressiveness,” $\xi_1$, and “worry,” $\xi_2$ 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 $\xi_1$ and $\xi_2$, it may be considered the “full model.”

We calculated values of $\{IC_{\mathcal{H}(k),Q} : k = 0, 1\}$ and IC$_Q$ with $\hat{c}_n(\hat{\theta}) = 2d$ and $d \log(n)$ for all four models (Table 3). The calculation of $\{IC_{\mathcal{H}(k),Q} : k = 0, 1\}$ 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 $\hat{\eta}_i = -0.0579 x_{i1} + 0.0821 x_{i2} - 0.2711 \xi_1 + 0.2505 \xi_2 + 0.1897 \xi_1 \xi_2$. We note the fact that comparatively large (positive) values of $(\eta_i, x_{i2})$ (or $x_{i1}, \xi_1$) and $\xi_2$ 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. $b_1 = -0.0579$ indicates that the longer sex workers are in their jobs, the less threat they feel from AIDS, and $b_2 = 0.0821$ implies that the more they think that they know about AIDS, the more threat they feel from AIDS.
2. \( \hat{\gamma}_1 = -0.2711 \) shows that the more aggressive the sex workers are, the less threat they feel from AIDS, and \( \hat{\gamma}_2 = 0.2505 \) shows that sex workers who are more worried about contracting AIDS feel more of a threat from AIDS.

3. \( \hat{\gamma}_3 = 0.1897 \) indicates that \( \xi_1 \) and \( \xi_2 \) have a positive interaction effect on “threat of AIDS.”

It is easily seen from the foregoing analysis that introducing an interaction term in an nonlinear structural equation model to interpret the relationship between \( \eta_i \) and \( \xi_1, \xi_2 \) is very necessary, and we could get various different effects for different cases. The estimated correlation between “aggressiveness,” \( \xi_1 \), and “worry,” \( \xi_2 \), is -0.1819, which indicates that they are negatively correlated.

### 4. DISCUSSION

We have proposed a general class of model selection criteria, IC\( \bar{H}(k),Q \), for missing-data problems. The computation of IC\( \bar{H}(k),Q \) can be obtained directly from the EM output. The theory of IC\( \bar{H}(k),Q \) is quite general and can be applied to various types of missing-data models for which the EM algorithm is applicable. Moreover, IC\( \bar{H}(k),Q \) 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\( \bar{H}(k),Q \) and IC\( Q \). Compared with \( \hat{c}_n(\hat{\theta}) = 2d \), the use of the penalty \( d \log(n) \) for IC\( \bar{H}(k),Q \) 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 IC\( \bar{H}(k),Q \) 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

### Table 3. AIDS data: Values of IC\( Q \) and IC\( \bar{H}(k),Q \) 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 \( M_3 \)

<table>
<thead>
<tr>
<th>Model</th>
<th>IC( Q )</th>
<th>IC( \bar{H}(0),Q )</th>
<th>IC( \bar{H}(1),Q )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( M_0 )</td>
<td>34,676.19</td>
<td>34,896.96</td>
<td>32,941.28</td>
</tr>
<tr>
<td>( M_1 )</td>
<td>34,680.18</td>
<td>34,905.97</td>
<td>32,961.77</td>
</tr>
<tr>
<td>( M_2 )</td>
<td>34,689.32</td>
<td>34,915.11</td>
<td>32,964.85</td>
</tr>
<tr>
<td>( M_3 )</td>
<td>34,708.79</td>
<td>34,939.60</td>
<td>32,988.38</td>
</tr>
</tbody>
</table>

### Table 4. AIDS data: Selected ML estimates and their standard deviations (SDs) for model \( M_0 \)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>ML estimates</th>
<th>SD</th>
<th>Parameter</th>
<th>ML estimates</th>
<th>SD</th>
<th>Parameter</th>
<th>ML estimates</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \mu_1 )</td>
<td>3.6362</td>
<td>0.0292</td>
<td>( \psi_1 )</td>
<td>0.9405</td>
<td>0.0765</td>
<td>( \lambda_{21} )</td>
<td>0.4493</td>
<td>0.1124</td>
</tr>
<tr>
<td>( \mu_2 )</td>
<td>2.5977</td>
<td>0.0432</td>
<td>( \psi_2 )</td>
<td>2.2057</td>
<td>0.0931</td>
<td>( \lambda_{31} )</td>
<td>0.7736</td>
<td>0.1558</td>
</tr>
<tr>
<td>( \mu_3 )</td>
<td>3.9725</td>
<td>0.0321</td>
<td>( \psi_3 )</td>
<td>0.9525</td>
<td>0.0464</td>
<td>( \lambda_{52} )</td>
<td>1.6294</td>
<td>0.1679</td>
</tr>
<tr>
<td>( \mu_4 )</td>
<td>0.0015</td>
<td>0.0052</td>
<td>( \psi_4 )</td>
<td>0.8665</td>
<td>0.0383</td>
<td>( \lambda_{62} )</td>
<td>1.1107</td>
<td>0.0859</td>
</tr>
<tr>
<td>( \mu_5 )</td>
<td>0.0031</td>
<td>0.0032</td>
<td>( \psi_5 )</td>
<td>0.6246</td>
<td>0.1358</td>
<td>( \lambda_{83} )</td>
<td>0.4220</td>
<td>0.1407</td>
</tr>
<tr>
<td>( \mu_6 )</td>
<td>0.0020</td>
<td>0.0092</td>
<td>( \psi_6 )</td>
<td>0.8251</td>
<td>0.0452</td>
<td>( \lambda_{93} )</td>
<td>0.7358</td>
<td>0.1149</td>
</tr>
<tr>
<td>( \mu_7 )</td>
<td>4.3696</td>
<td>0.0038</td>
<td>( \psi_7 )</td>
<td>0.7179</td>
<td>0.0783</td>
<td>( b_1 )</td>
<td>-0.0579</td>
<td>0.0310</td>
</tr>
<tr>
<td>( \mu_8 )</td>
<td>3.1411</td>
<td>0.0431</td>
<td>( \psi_8 )</td>
<td>2.0665</td>
<td>0.0900</td>
<td>( b_2 )</td>
<td>0.0821</td>
<td>0.0290</td>
</tr>
<tr>
<td>( \mu_9 )</td>
<td>3.7998</td>
<td>0.0344</td>
<td>( \psi_9 )</td>
<td>1.4165</td>
<td>0.0865</td>
<td>( \gamma_1 )</td>
<td>-0.2711</td>
<td>0.0679</td>
</tr>
<tr>
<td>( \phi_{11} )</td>
<td>-0.0422</td>
<td>0.0090</td>
<td>( \psi_x )</td>
<td>1.4774</td>
<td>0.6778</td>
<td>( \gamma_2 )</td>
<td>0.2505</td>
<td>0.1060</td>
</tr>
<tr>
<td>( \phi_{22} )</td>
<td>-0.3819</td>
<td>0.0418</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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 4th 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 10−2 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.

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

Proof of Theorem 1

We need only show that sup{(θ1, θ2) ∈ Θ × Θ : n−1|K(k)(θ1, θ2) − E[K(k)(θ1, θ2)]| → 0 in probability and E[K(k)(θ1, θ2)] is continuous in θ1 and θ2 uniformly over Θ × Θ. Conditions (C3) and (C4) are sufficient for assumption W-LIP of Andrews (1992), which ensures the continuity of E[K(k)(θ1, θ2)] and the stochastic equicontinuity (SE) of K(k)(θ1, θ2). Furthermore, conditions (C3) and (C4) ensure pointwise convergence; that is, n−1|K(k)(θ1, θ2) − E[K(k)(θ1, θ2)]| converges to 0 for each θ1 and θ2 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

$$\sqrt{n}(\hat{\theta} - \theta) = A(\theta) + o_p(1).$$

(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 K(k) and E[K(k)] as follows:

$$\tilde{K}(k)(\hat{\theta}, \theta^*) = K(k)(\theta, \theta^*) + [\partial_\theta K(k)(\theta, \theta^*)] \Delta \hat{\theta} + o_p(1),$$

(A.2)

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

$$E_{D_{obs}}[\tilde{K}(k)(\hat{\theta}, \theta^*)] = E_{D_{obs}}[K(k)(\theta, \theta^*)] + E_{D_{obs}}[[\partial_\theta K(k)(\theta, \theta^*)] \Delta \hat{\theta}] + o(1).$$

(A.3)

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

$$E_{D_{obs}}[\tilde{K}(k)(\hat{\theta}, \theta^*)] = \mathbb{E}_{D_{obs}}[\partial_\theta K(k)(\theta, \theta^*)] + o(1)$$

$$= tr[A(\theta)^{-1} B(\theta_0(\theta^*))] + o(1).$$

(A.4)

Proof of Theorem 3

Based on Theorem 1 and $\delta_{c_{21}} = o_p(n)$, we have

$$n^{-1} dIC_h(k, Q_{21}) = 2n^{-1} \delta_{Q_{21}} + o_p(1),$$

which yields Theorem 3a.

Theorem 3b can be proved by noting that $n^{-1/2} dIC_h(k, Q_{21})$ can be written as the sum of

$$-2n^{-1/2} \mathbb{E} \left[ Q^{(\theta_1)}(\theta_2) - \hat{h}(k|\theta_2) \right]$$

and

$$2n^{-1/2} \mathbb{E} \left[ Q^{(\theta_1)}(\theta_2) - \mathbb{E} \left[ Q^{(\theta_1)}(\theta_2) \right] \right].$$

Note that for $t = 1, 2$, $Q^{(\theta_1)}(\hat{\theta}(t))$ can be written as

$$Q^{(\theta_1)}(\hat{\theta}(t)) = Q^{(\theta_1)}(\theta_1) + 5(\theta_2 - \theta_1) \Delta \theta + o_p(1).$$

Because $\hat{\theta}(t) - \theta_1 = O_p(n^{-1/2})$, $Q^{(\theta_1)}(\hat{\theta}(t)) = Q^{(\theta_1)}(\theta_1) + O_p(1)$. Thus $dIC_h(k, Q_{21})$ can be written as

$$2Q^{(\theta_1)}(\theta_1) - 2Q^{(\theta_1)}(\hat{\theta}_2) + \delta_{c_{21}} + O_p(1).$$

Theorem 3c can be proved by noting that $Q^{(\theta_1)}(\hat{\theta}_1) - Q^{(\theta_1)}(\hat{\theta}_2) = O_p(1)$ and $\delta_{c_{21}} \rightarrow \infty$.

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.

$y_{10}$ (item 37): How long (in months) have you been working at a job where people pay to have sex with you?
REFERENCES


Huang, L., Chen, M. H., and Ibrahim, J. G. (2005), “Bayesian Analysis for Gen-


This article has been cited by:


