Biometrics 64, 1062–1069
December 2008
DOI: 10.1111/j.1541-0420.2008.01003.x

Variable Selection with Incomplete Covariate Data

Gerda Claeskens* **and Fabrizio Consentino**
K.U. Leuven, ORSTAT and Leuven Statistics Research Center,
Naamsestraat 69, 3000 Leuven, Belgium
**email:* gerda.claeskens@econ.kuleuven.be

Summary. Application of classical model selection methods such as Akaike's information criterion (AIC) becomes problematic when observations are missing. In this article we propose some variations on the AIC, which are applicable to missing covariate problems. The method is directly based on the expectation maximization (EM) algorithm and is readily available for EM-based estimation methods, without much additional computational efforts. The missing data AIC criteria are formally derived and shown to work in a simulation study and by application to data on diabetic retinopathy.

Key words: Akaike information criterion; EM algorithm; Missing covariates; Model selection; Takeuchi's information criterion.

1. Introduction

This article develops a model selection criterion in the spirit of Akaike's (1973) information criterion (AIC), which is usable for missing data situations in parametric regression models where the response is completely observed, though some of the covariates might be incomplete. The dataset considered for discussion is from the Wisconsin Epidemiologic Study of Diabetic Retinopathy (Klein et al., 1984). It provides information to study diabetic retinopathy as a function of several measurements. The binary outcome variable that indicates the presence of moderate to severe nonproliferate retinopathy, or proliferate retinopathy for at least one of the eyes, is completely observed. Of the other variables, x_1, the intraocular pressure in mmHg (maximum of the measurements for both eyes), contains missing values for 51 out of the 996 cases, and variable x_4, the percentage of glycosylated hemoglobin, contains 46 missing values. For three cases both variables are missing. The other variables do not contain missing information. Fitting models that include variables x_1 and/or x_4 requires special attention because of the missingness of observations for some subjects. Application of traditional model selection methods such as AIC (Akaike, 1973) is easy when using the complete cases only, that is, when leaving out the subject information for those cases where information is missing. It is well known that an analysis of the subset of complete cases only may lead to biased results, and obviously in that way we would not be using all available information in the gathered data. Therefore we sought for an extension of the AIC that is readily applicable to datasets with incomplete cases, where the observed data likelihood are not straightforward to obtain.

The proposed criterion is directly based on the expectation maximization (EM) algorithm and does not require much additional programming efforts. To deal with the missing covariates, we follow the approach of Ibrahim, Chen, and Lipsitz (1999). Their procedure obtains parameter estimates via weighting, extending on Ibrahim (1990), and uses Gibbs sampling to draw from the distribution of the missing covariates, given the observed variables in a Monte Carlo EM algorithm. Their missing data mechanism is assumed to be ignorable. An extension to nonignorable missingness is made in Ibrahim, Lipsitz, and Chen (1999). These methods are valid for categorical and continuous variables, as well as a mixture of those, and so is the model selection method that we propose. We phrase the new criterion in terms of the function that is to be minimized in the EM algorithm. Our model selection method is applicable to likelihood-based models, including the class of generalized linear models.

There are connections of our proposed method to the model selection criterion of Cavanaugh and Shumway (1998), though it differs from that in several aspects. First, we do not make the strong assumption that the likelihood model has to be correctly specified. Instead, our derivation makes use of "best approximating" parameter values, which are defined as the best approximations in Kullback–Leibler sense between a true (and unknown) data-generating mechanism, and the likelihood model used in practice. Second, although their application treats missing response data, we explicitly work in a regression setting with missing covariate data. Third, we use the EM algorithm by the methods of weights of Ibrahim, Chen, and Lipsitz (1999) and Ibrahim, Lipsitz, and Chen (1999), which can deal with models with incomplete covariates. Shimodaira (1994) proposed the predictive divergence for indirect observation models (PDIO), which differs from the proposal by Cavanaugh and Shumway (1998) in that it uses the likelihood of the incomplete data as the goodness-of-fit part of the criterion.

A different approach is developed by Hens, Aerts, and Molenberghs (2006) who consider weighting the complete cases by their inverse selection probabilities, following the idea of the Horvitz–Thompson estimator. A drawback of this method is that it requires estimation of the selection

probabilities, which can be done either parametrically, or nonparametrically, the latter requires yet additional smoothing parameters to be determined. Moreover, the model selection part is distinct from the estimation part of the model. Our goal in this article is to directly use the quantities available in the estimation procedure, in particular the EM algorithm, to get information on the model-fitting aspect. One version of our new AIC also takes the complexity in modeling the missingness mechanism into account.

The article proceeds as follows. In Section 2 we define notation and state assumptions. Guided by results on Kullback–Leibler distances in Section 3, the proposed criteria are defined in Section 4. Section 5 presents the results of a simulation study and data application. Some concluding remarks are in Section 6.

2. Assumptions and Notation

Throughout the article we use the following assumptions and notation. Some of the explanatory variables $X_{i1}, \ldots, X_{ip}$ contain missing observations, whereas the response variable Y_i is fully observed. The vectors $(Y_i, X_{i1}, \ldots, X_{ip})$ for $i = 1, \ldots, n$ are independent. Let $\boldsymbol{Y}$ denote the vector of response values of length n, and $\boldsymbol{X}$ the corresponding design matrix of regression variables, of dimension $n \times p$. This matrix is partitioned in two parts, $\boldsymbol{X} = (\boldsymbol{X}_{\text{obs}}, \boldsymbol{X}_{\text{mis}})$, where $\boldsymbol{X}_{\text{obs}}$ contains in its columns those variables $\boldsymbol{X}_j = (X_{1j}, \ldots, X_{nj})'$ that are completely observed for all subjects $i = 1, \ldots, n$. The matrix $\boldsymbol{X}_{\text{mis}}$ represents the variables X_k of which for at least one i, a value X_{ik} is not observed.

The proposed model selection method is likelihood based. The data are modeled by means of a parametric class of likelihood functions. Denote by $f_\theta = f(\boldsymbol{Y}, \boldsymbol{X}; \boldsymbol{\theta})$ the modeled density function for the complete data $(\boldsymbol{Y}, \boldsymbol{X})$, that is, including the variables that are only partly observed, where $\boldsymbol{\theta}$ is the unknown parameter vector. Further, $f(\boldsymbol{Y}, \boldsymbol{X}_{\text{obs}}; \boldsymbol{\theta})$ is the density function for the subset of completely observed data $(\boldsymbol{Y}, \boldsymbol{X}_{\text{obs}})$. When the observations are independent, as we assume here, the full n-dimensional likelihood function is equal to the product of n one-dimensional likelihood functions $f_\theta = \prod_{i=1}^n f(Y_i, X_{i1}, \ldots, X_{ip}; \boldsymbol{\theta})$. The joint distribution of $(\boldsymbol{Y}_i, \boldsymbol{X}_i)$ is modeled by specifying the conditional distribution of $(\boldsymbol{Y}_i \mid \boldsymbol{X}_i)$ and the marginal distribution of $(\boldsymbol{X}_i)$. In this way the model is described as

$$f_\theta = f(\boldsymbol{Y}, \boldsymbol{X}; \boldsymbol{\theta}) = f(\boldsymbol{Y} \mid \boldsymbol{X}; \boldsymbol{\beta}) f(\boldsymbol{X}; \boldsymbol{\alpha}), \tag{1}$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\alpha})$, and where the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are distinct. One common example to describe the distribution of Y_i, given $\boldsymbol{X}_i$ is the class of generalized linear models where the response random variable belongs to the exponential family. For the dataset, for example, we use the logistic regression model. Because in the dataset the two variables with missing observations are continuous, we use a bivariate normal model to regress these two variables on the remaining variables. In our discussion we assume that the missing data are "missing at random" (MAR), as defined by Little and Rubin (2002), which means that the missingness mechanism depends only on the observed values. Considering both the MAR assumption and the distinctness of the parameters, the missingness mechanism is ignorable and it is not necessary to model it. The estimation of the model proceeds using a weighting method as proposed by Ibrahim (1990). In the presence of missing observations the EM algorithm is particularly suitable. The E-step concerns the estimation of the expectation of $f_\theta = f(\boldsymbol{Y}, \boldsymbol{X}; \boldsymbol{\theta})$. The relevant quantity Q that is further maximized in the M-step is defined as $Q = \sum_{i=1}^n Q_i$ with

$$Q_i\big(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)}\big) = \int w_i \log f(y_i, x_i; \boldsymbol{\theta})\, dx_{\text{mis},i}, \tag{2}$$

where $w_i = f(x_{\text{mis},i} \mid x_{\text{obs},i}, y_i; \boldsymbol{\theta}^{(k)})$. Because we do not need to model the completely observed covariates $\boldsymbol{X}_{\text{obs}}$ as their distribution is fixed and does not influence the estimation of the model parameters, it is allowed to condition on the observed covariates and work with the conditional distribution of $\boldsymbol{X}_{\text{mis}} \mid \boldsymbol{X}_{\text{obs}}$ instead of with the full distribution of $\boldsymbol{X} = (\boldsymbol{X}_{\text{obs}}, \boldsymbol{X}_{\text{mis}})$.

A Monte Carlo EM algorithm is used for the evaluation of Q, using the Gibbs sampler along with the adaptive rejection algorithm of Gilks and Wild (1992), in order to sample from $(x_{\text{mis},i} \mid x_{\text{obs},i}, y_i; \boldsymbol{\theta}^{(k)})$. The proposed information criteria do not depend on the particular way of computation and alternatives may be used. Because of the factorization in (1), the function Q_i can be written as

$$\begin{aligned} Q_i(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)}) &= \int w_i \log f(y_i \mid x_i; \boldsymbol{\beta})\, dx_{\text{mis},i} \\ &\quad + \int w_i \log f(x_i; \boldsymbol{\alpha})\, dx_{\text{mis},i} \\ &= Q_i^{(1)}\big(\boldsymbol{\beta} \mid \boldsymbol{\theta}^{(k)}\big) + Q_i^{(2)}\big(\boldsymbol{\alpha} \mid \boldsymbol{\theta}^{(k)}\big). \end{aligned} \tag{3}$$

The separation in two parts plays a crucial role in the derivation of the model selection criterion and in its interpretation (see also Section 4.2).

3. The Kullback–Leibler Distance and the Q Function

The information criterion AIC (Akaike, 1973) originates as an approximation to the expected Kullback–Leibler distance between the true data-generating density g and the model density f_θ that is used for estimating parameters by means of maximum likelihood. For a derivation for the case of completely observed data, see, for example, Burnham and Anderson (2002). In the derivation we assume that there exists a true likelihood function $g(\boldsymbol{Y}, \boldsymbol{X})$ for the complete data, which is, however, not needed to be known in practice.

The Kullback–Leibler divergence is defined as follows:

$$KL(g, f_\theta) = E_g[\log\{g(\boldsymbol{Y}, \boldsymbol{X}) / f(\boldsymbol{Y}, \boldsymbol{X}; \boldsymbol{\theta})\}], \tag{4}$$

where, unless mentioned otherwise, the expectation is with respect to the true density.

Originally, an estimator $\hat{\boldsymbol{\theta}}$ is obtained by maximum likelihood, and the least false parameter value $\boldsymbol{\theta}_0$, also called the best approximating parameter value, is the value of $\boldsymbol{\theta}$ for which $KL(g, f_\theta)$ is as small as possible. Or equivalently, for which $E_g\{\log f(\boldsymbol{Y}, \boldsymbol{X}; \boldsymbol{\theta})\}$ is as large as possible.

In this setting of missing covariate data, the density f_θ cannot be evaluated at $\boldsymbol{Y}, \boldsymbol{X}$. Instead we obtain an estimator by means of a Monte Carlo EM algorithm. The method of weights arrives at a weighted log-density function in the E-step of the algorithm. Let $\hat{\boldsymbol{\theta}}$ be the maximizer found by this

algorithm. Referring to the function Q in (3), the least false parameter value in this situation is the value $\boldsymbol{\theta}_0$ for which $E_g\{\dot{Q}(\boldsymbol{\theta}_0 \,|\, \boldsymbol{\theta}_0)\} = 0$, where $\dot{Q}(\boldsymbol{\theta}_1 \,|\, \boldsymbol{\theta}_2) = \frac{\partial}{\partial \boldsymbol{\theta}_1} Q(\boldsymbol{\theta}_1 \,|\, \boldsymbol{\theta}_2)$. The estimator $\hat{\boldsymbol{\theta}}$ solves the equation $\dot{Q}(\hat{\boldsymbol{\theta}} \,|\, \hat{\boldsymbol{\theta}}) = 0$.

The method of weights assigns weights to the log-likelihood function, and then integrates over the missing covariates. The weights are defined via the density function (or probability mass function for categorial covariates) of the covariates with missing observations, given the observed data. The "adjusted" likelihood function is hence defined as $\tilde{f}_\theta(\boldsymbol{y}, \boldsymbol{x}) = \exp Q(\boldsymbol{\theta} \,|\, \boldsymbol{\theta})$, or $\log \tilde{f}_\theta(\boldsymbol{y}, \boldsymbol{x}) = Q(\boldsymbol{\theta} \,|\, \boldsymbol{\theta})$ where

$$Q(\boldsymbol{\theta}_1 \,|\, \boldsymbol{\theta}_2) = \sum_{i=1}^{n} \int \log f(y_i, x_{\text{obs},i}, x_{\text{mis},i}; \boldsymbol{\theta}_1) \times f(x_{\text{mis},i} \,|\, x_{\text{obs},i}, y_i, \boldsymbol{\theta}_2)\, dx_{\text{mis},i}.$$

When using the integrated weighted log-likelihood function instead of $\log f_\theta$, the relevant Kullback–Leibler distance to minimize is

$$KL(g, \tilde{f}_\theta) = [E_g\{\log g(\boldsymbol{Y}, \boldsymbol{X})\} - E_g\{\log \tilde{f}_\theta(\boldsymbol{Y}, \boldsymbol{X})\}]/n.$$

Because the first term does not depend on $\boldsymbol{\theta}$, we can focus on the second term only. At the estimated parameter value, the relevant quantity to work with is

$$\int g(\boldsymbol{y}, \boldsymbol{x}) \log \tilde{f}(\boldsymbol{y}, \boldsymbol{x}; \hat{\boldsymbol{\theta}})\, d\boldsymbol{y}\, d\boldsymbol{x}/n.$$

Because $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\boldsymbol{Y}, \boldsymbol{X}_{\text{obs}})$, its expected value is equal to

$$K_n = \int g(\boldsymbol{y}, \boldsymbol{x}) \int g(\tilde{\boldsymbol{y}}, \tilde{\boldsymbol{x}}) \log \tilde{f}(\tilde{\boldsymbol{y}}, \tilde{\boldsymbol{x}}; \hat{\boldsymbol{\theta}})\, d\tilde{\boldsymbol{y}}\, d\tilde{\boldsymbol{x}}\, d\boldsymbol{y}\, d\boldsymbol{x}/n.$$

An estimator of K_n is $\hat{K}_n = Q(\hat{\boldsymbol{\theta}} \,|\, \hat{\boldsymbol{\theta}})/n$. The following theorem is the main motivation for the model selection criterion when covariates contain missing values.

THEOREM 1. *Let f be two times continuously differentiable with respect to $\boldsymbol{\theta}$, with bounded expectation of the second derivative in a neighborhood of θ_0, which belongs to the interior of a compact parameter space. If $n(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)'(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ is uniformly integrable,*

$$E_g(\hat{K}_n - K_n) = \text{tr}\{I^{-1}(\boldsymbol{\theta}_0) J(\boldsymbol{\theta}_0)\}/n + o(1/n),$$

where $I(\boldsymbol{\theta}) = E\{-\ddot{Q}(\boldsymbol{\theta} \,|\, \boldsymbol{\theta})/n\}$, and $J(\boldsymbol{\theta}) = \text{var}\{\dot{Q}(\boldsymbol{\theta} \,|\, \boldsymbol{\theta})\}/n$.

An outline of the proof is placed in the Appendix. This motivates the approximation of $E_g(\hat{K}_n - K_n)$ by $\text{tr}\{I^{-1}(\boldsymbol{\theta}_0) J \times (\boldsymbol{\theta}_0)\}/n$. For similar derivations in the case of completely observed data, we refer to Linhart and Zucchini (1986).

4. The Model Selection Criteria

4.1 *A Model-Robust Criterion for Data with Missing Covariates*

Because $\hat{K}_n$ is overestimating K_n, we maximize a bias-corrected version of $\hat{K}_n$. In the spirit of Takeuchi's (1976) information criterion (TIC), we define the model-robust criterion TIC for missing covariate values as

$$\text{TIC} = -2\, Q(\hat{\boldsymbol{\theta}} \,|\, \hat{\boldsymbol{\theta}}) + 2\, \text{tr}\{\hat{J}(\hat{\boldsymbol{\theta}}) \hat{I}^{-1}(\hat{\boldsymbol{\theta}})\}, \tag{5}$$

where

$$\hat{I}(\hat{\boldsymbol{\theta}}) = -\frac{1}{n} \ddot{Q}(\hat{\boldsymbol{\theta}} \,|\, \hat{\boldsymbol{\theta}}) \text{ and } \hat{J}(\hat{\boldsymbol{\theta}}) = \frac{1}{n} \sum_{i=1}^{n} \dot{Q}_i(\hat{\boldsymbol{\theta}} \,|\, \hat{\boldsymbol{\theta}}) \dot{Q}_i(\hat{\boldsymbol{\theta}} \,|\, \hat{\boldsymbol{\theta}})'.$$

The model with the smallest value of TIC is chosen. This criterion consists of two parts. The first part is the "goodness-of-fit" term, whereas the second one is the "penalty" term, representing twice the effective number of parameters in the model. The criterion is called model robust because it allows for the possibility that the model used is not the correct one, as reflected by the use of estimators for the matrices I and J, which are equal in case the model is correct.

The estimation of the penalty term is straightforward to compute; the $\dot{Q}(\hat{\boldsymbol{\theta}} \,|\, \hat{\boldsymbol{\theta}})$ is a $k \times 1$ gradient, whereas the $\ddot{Q}(\hat{\boldsymbol{\theta}} \,|\, \hat{\boldsymbol{\theta}})$ is a $k \times k$ block matrix of the second derivatives, due to the distinctness of the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$.

For this article it is not our concern to provide accurate estimators of the information matrices I and J, the expressions above are mainly used to provide information on the effective number of parameters in the model. The proposed version of TIC is applicable in a wide range of missing data models. If the information matrices are to be used to obtain more precise variance estimators, one could take the properties of the specific EM algorithm into account. Different types of EM algorithms might require different final variance formulae (see, e.g., Louis, 1982; Meng and Rubin, 1991; Nielsen, 2000).

4.2 *AIC for Data with Missing Covariates*

It is important to point out that if the model f_θ is in fact the correct one, that is, $f_\theta = g$, then the matrices I and J are equal, and the penalty in the expression of the TIC reduces to the number of parameters in the model. This simplification can be applied regardless of whether the model holds or not, leading to a version of AIC (Akaike, 1973) suitable for use with missing covariate information:

$$\text{AIC} = -2\, Q(\hat{\boldsymbol{\theta}} \,|\, \hat{\boldsymbol{\theta}}) + 2\, p_\theta, \tag{6}$$

where $p_\theta = \text{length}(\boldsymbol{\theta})$. The model with the smallest value of AIC is selected. We wish to stress that both criteria AIC and TIC use the "full" function Q, including the part on the regression relationship between the response Y and the covariates X_j, as well as the part where missingness is taken care of, leading to $Q_i^{(2)}$. Because of this second component, the AIC and TIC are not directly comparable to their counterparts in models where all variables are observed. This leads to a problem when comparing models in case some of the models contain only variables that are completely observed, whereas other models contain variables with some observations missing. Therefore, if interest is in modeling the regression structure between Y and the covariates $X_1, \ldots, X_p$, we restrict to that part of the Q function dealing with modeling $f(y|x, \beta)$ only. In this case the criterion reads

$$\text{AIC}_1 = -2\, Q^{(1)}(\hat{\boldsymbol{\beta}} \,|\, \hat{\boldsymbol{\theta}}) + 2\, p_\beta, \tag{7}$$

where $p_\beta = \text{length}(\boldsymbol{\beta})$. The value of this criterion is directly comparable to the classical AIC of Akaike (1973) in case there are no missing observations. This is important to compare AIC values across different models. For models S not

containing any of the incompletely observed variables in $\boldsymbol{X}_{\text{obs}}$, we compute the classical

$$\text{AIC}_{\text{classical}} = -2 \log f(\boldsymbol{Y}, \boldsymbol{X}_S; \hat{\boldsymbol{\beta}}_S) + 2\ p_{\beta_S},$$

where $\boldsymbol{X}_S$ denotes a subset of the covariates $X_1, \ldots, X_p$, and $\boldsymbol{\beta}_S$ is the vector of the corresponding model coefficients. When $\boldsymbol{X}_S$ has some variables in common with $\boldsymbol{X}_{\text{mis}}$, we use AIC_1 instead. Indeed, for models not containing variables with missing observations, $\tilde{f}_\theta(\boldsymbol{y}, \boldsymbol{x})$ is equal to $f_\theta(\boldsymbol{y}, \boldsymbol{x})$ and AIC_1 reduces to the classical AIC. This guarantees that values of AIC (for models including subsets of the variables that are completely observed) are immediately comparable to those obtained by AIC_1, which allows a model search amongst *all* of the variables $X_1, \ldots, X_p$.

For TIC, a similar reduced version (denoted TIC_1) is defined using $Q^{(1)}$ and contains as the penalty term the trace of the upper left submatrix of dimension $p_\beta \times p_\beta$ of the matrix $\{\hat{J}(\hat{\boldsymbol{\theta}})\hat{I}^{-1}(\hat{\boldsymbol{\theta}})\}$.

The criterion AIC, which is built with the full function Q, also takes the complexity of the missingness modeling into account. More complex models $f(\boldsymbol{X}; \boldsymbol{\alpha})$ (with a higher-dimensional $\boldsymbol{\alpha}$) will get a heavier punishment. This criterion is interesting to compare different models for $f(\boldsymbol{X}; \boldsymbol{\alpha})$, which otherwise mainly is done via sensitivity studies or based on heuristical arguments. The "full" AIC and TIC are particularly useful in the situation in which one has decided upon a structure of the regression model and wishes to compare different models for $\boldsymbol{X}_{\text{mis}} \,|\, \boldsymbol{X}_{\text{obs}}$.

In the simulation study and example we have used a bivariate normal model for the distribution of the pair of covariates (of which missing records were reported). Alternatively, if it were expected that the tails of the distribution were heavier than expected under normality, we could have used, for example, a bivariate t-distribution (see, e.g., Kotz and Nadarajah, 2004, for an extensive overview of this distribution). Such a t-distribution is often used for robustness reasons, to take possible outlying observations into account. Liu (1995) and Liu and Rubin (1995) used this distribution in the context of missing data imputation. To decide on which one fits the data best, either the bivariate normal distribution or the bivariate t-distribution, either criterion AIC or TIC (or their small sample variants) could be applied, and the model with the best such value would be considered best for the data at hand.

5. Applications

5.1 *Simulation Study*

To examine the validity of the proposed model selection criteria, we performed a simulation study based on a logistic regression model. We consider different simulation settings, related with different sample sizes and different percentages of missingness. The model that we use to simulate data from is given by

$$\log f(y_i \,|\, \boldsymbol{x}_i; \boldsymbol{\theta}) = \sum_{i=1}^{n} \{y_i \boldsymbol{x}_i' \boldsymbol{\beta} - \log(1 + \exp(\boldsymbol{x}_i' \boldsymbol{\beta}))\} \quad (8)$$

with

$$P(Y_i = 1 \,|\, \boldsymbol{x}_i, \boldsymbol{\beta}) = \frac{\exp(\boldsymbol{x}_i' \boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_i' \boldsymbol{\beta})}.$$

The vector of covariates for the ith observation is given by $\boldsymbol{x}_i' = (1,\ x_{i1},\ x_{i2},\ x_{i3},\ x_{i4})$, with $\boldsymbol{\beta}' = (\beta_0,\ \beta_1,\ \beta_2,\ \beta_3,\ \beta_4)$. The true values chosen for the coefficients are $\boldsymbol{\beta} = (1,\ 1,\ 0,\ 0,\ -1)$. The covariates are generated independently from a standard normal distribution. Only $(x_{i1},\ x_{i2})$ contain missing observations; the missingness in both variables is introduced by generating a missing data mechanism that depends on the fully observed variables $(x_{i3},\ x_{i4})$, leading to the MAR assumption. Independent standard normal errors ϵ_{ij} are generated, and a data value x_{ij} is set to be missing when $(x_{i3} - \overline{x_{.3}}) - (x_{i4} - \overline{x_{.4}}) + \epsilon_{ij} \le z_\alpha$, where $\overline{x_{.k}}$ is the sample mean of x_k and z_α is the α-quantile of a $N(0,\ 3)$ random variable with α being the chosen percentage. This scenario is the same for all the different simulation settings used. Two different sample sizes, $n = 50$ and $n = 100$, are considered and for each of them four different percentages are used, (5%, 5%), (10%, 5%), (15%, 15%), and (30%, 5%). For each setting we run $N = 500$ simulations. The method developed by Ibrahim (1990) uses the Monte Carlo EM algorithm for computing the parameter estimates and the Gibbs sampler along with the adaptive rejection algorithm of Gilks and Wild (1992) in order to get sample from $(x_{\text{mis},i} \,|\, x_{\text{obs},i},\ y_i;\ \boldsymbol{\theta}^{(k)})$, which is valid because of the log concavity of the conditional distribution of $\boldsymbol{Y}$, given $\boldsymbol{X}$ within the exponential family. This leads to the Q function, the gradient $\dot{Q}$, and the Hessian $\ddot{Q}$. In particular, when the ith observation is not completely observed, a sample $z_{i1}, \ldots, z_{im}$ of size m is taken from the distribution of $(x_{\text{mis},i} \,|\, x_{\text{obs},i},\ y_i;\ \boldsymbol{\theta}^{(k)})$; each z_{ij} depends on the iteration number. The E-step for the ith observation at the $(k+1)$th iteration is

$$Q_i(\boldsymbol{\theta} \,|\, \boldsymbol{\theta}^{(k)}) = \frac{1}{m} \sum_{j=1}^{m} \log f(z_{ij}, x_{\text{obs},i}, y_i; \boldsymbol{\theta}).$$

It is straightforward to compute the first and the second derivatives of $\log f(z_{ik},\ x_{\text{obs},i},\ y_i;\ \boldsymbol{\theta})$. A bivariate normal regression model is used for the variables $(x_1,\ x_2)$ that contain missing observations,

$$X_{i1} = \alpha_{10} + \alpha_{11} x_{i3} + \alpha_{12} x_{i4} + \varepsilon_{i1}$$

$$X_{i2} = \alpha_{20} + \alpha_{21} x_{i3} + \alpha_{22} x_{i4} + \varepsilon_{i2},$$

where

$$\begin{pmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \end{pmatrix} \sim N_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma \right)$$

and Σ is a 2×2 covariance matrix. The number of Monte Carlo iterations within each iteration of the EM algorithm is set to $m = 500$. Before choosing this value, we have compared different Monte Carlo sample sizes of 500, 1000, 2000, and 5000, based on limited simulations. In particular, we compared the ratio of the values of the TIC and AIC criteria based on above Monte Carlo sample sizes, that is, 500/5000, 1000/5000, and 2000/5000. For both criteria, the ratios are very close to 1, showing that the chosen sample size of 500 is suitable for the analysis; for instance, the AIC_1 500/5000 ratio is 1.00036, whereas the AIC_1 2000/5000 ratio is 0.99950. Similar values are found for a comparison of the parameter values and the other criteria among which are the full AIC and TIC. The convergence criterion used in the EM algorithm

is the square distance between the tth iteration and the $(t+5)$th iteration that has to be less than 10^{-5} to have reached convergence. For each simulation run, all the possible submodels of $\text{logit}\{P(Y_i = 1 \mid \boldsymbol{x}_i; \boldsymbol{\theta})\} = \boldsymbol{x}_i'\boldsymbol{\beta}$ were fitted. In our situation, this means $2^4 = 16$ models. All programs have been written using the statistical software package R.

For each setting we compared four different model search strategies. Because we focus on the regression part of the model, we use the restricted versions of the criteria, only using $Q^{(1)}$:

(1) the TIC_1 of (5) though based on the $Q^{(1)}$ function and the penalty term with the trace formula adjusted as described in Section 4.2;
(2) the AIC_1 of (7) based on the $Q^{(1)}$ function, with the penalization based on the exact number of parameters;
(3) the classical AIC based on the original data, that is, before introducing missingness;
(4) the classical AIC based on the complete cases only, thus ignoring all observations with missingness.

In Table 1 we display the results of the simulations, where the numbers indicate the percentage of times that a model has been selected; in particular the selection is sorted according to three cases. The first case (C) concerns the correct model; the second case represents the overfitted models (O), which contain more parameters than strictly necessary; the third case represents the underfitted models (U), which do not contain at least one of the true parameters. The first and the second classification can be considered as good models.

Several observations can be made from the results summarized in Table 1. We first consider the AIC based on the subset of complete cases only. For the setting with the smallest sample size it is clear that it does not work properly, and it has the smallest number of correctly selected models. For all cases and both sample sizes it holds that for the subset of complete cases, the sum of the percentages that a correct or overfit model is selected, is the lowest. For the larger sample size the differences are smaller.

For the same setting of percentages of missingness, the total number of good models chosen by all criteria is increasing when the sample size changes from $n = 50$ to $n = 100$ and this aspect is present both in the selection of the correct model and in the selection of the overfitted models; moreover, this is valid for all four criteria analyzed.

When the probability of missingness in the covariates increases, the criteria perform in different ways, in particular the selection of the correct model decreases and this is due to the difficulty of correcting for the missingness; this is particularly clear in the setting with the sample size equal to 50. The observed percentages of correct models chosen by TIC_1 and AIC_1 decrease as the percentages of missingness increase but in a smaller way than for the AIC_{cc}; for instance, for the setting with the smallest sample size and the highest percentage of missingness only 29.8% of correct models are selected by AIC_{cc} in this simulation study.

From the column of good models (correct or overfit) we observe that the TIC_1 performs better than all other criteria, for all settings. The TIC_1 underfits to a lesser degree than AIC_1, though as a consequence, there is more overfitting. The TIC_1 is least likely to miss important variables in the selected model. The penalty term in the TIC_1 formula gives on average a smaller value than just counting the number of parameters

Table 1

Results of a simulation study. The table shows the percentage of times that the different criteria select the exact correct model (C), that an overfit model is selected (O), that it is a model that contains the correct model plus some additional variables, and (U) the underfit models. The percentage of good models is the sum of the values of correct and overfit models.

α-values		Model selection $n = 50$			Good	Model selection $n = 100$			Good
x_1, x_2	Criteria	C	O	U	model	C	O	U	model
5%, 5%	TIC_1	0.428	0.424	0.148	0.852	0.530	0.454	0.016	0.984
	AIC_1	0.476	0.338	0.186	0.814	0.602	0.378	0.020	0.980
	AIC_{orig}	0.564	0.286	0.150	0.850	0.676	0.312	0.012	0.988
	AIC_{cc}	0.436	0.198	0.366	0.634	0.628	0.290	0.082	0.918
10%, 5%	TIC_1	0.406	0.430	0.164	0.836	0.534	0.446	0.020	0.980
	AIC_1	0.430	0.350	0.220	0.780	0.590	0.384	0.026	0.974
	AIC_{orig}	0.560	0.288	0.152	0.848	0.676	0.312	0.012	0.988
	AIC_{cc}	0.390	0.196	0.414	0.586	0.590	0.306	0.104	0.896
15%, 15%	TIC_1	0.384	0.450	0.166	0.834	0.458	0.524	0.018	0.982
	AIC_1	0.440	0.322	0.238	0.762	0.530	0.442	0.028	0.972
	AIC_{orig}	0.560	0.280	0.160	0.840	0.676	0.312	0.012	0.988
	AIC_{cc}	0.318	0.152	0.530	0.470	0.560	0.284	0.156	0.844
30%, 5%	TIC_1	0.370	0.428	0.202	0.798	0.470	0.500	0.030	0.970
	AIC_1	0.406	0.322	0.272	0.728	0.532	0.420	0.048	0.952
	AIC_{orig}	0.564	0.266	0.170	0.830	0.676	0.312	0.012	0.988
	AIC_{cc}	0.298	0.158	0.544	0.456	0.590	0.224	0.186	0.814

(results not shown). This difference becomes slightly larger when the percentage of missingness increases and becomes slightly smaller when the sample size increases.

5.2 *Data Application*

The dataset considered for discussion is from the Wisconsin Epidemiologic Study of Diabetic Retinopathy (Klein et al., 1984). It provides information to study diabetic retinopathy as a function of several measurements. The full set of data consists of patient information for 484 women and 512 men. The binary outcome variable $Y = 0$ indicates whether there is no or only mild nonproliferate retinopathy on both of the eyes. An outcome value $Y = 1$ is obtained when there is moderate to severe nonproliferate retinopathy, or proliferate retinopathy for at least one of the eyes. Other variables are: x_1: the intraocular pressure in mmHg (maximum of the measurements for both eyes); x_2: the age of the patient; x_3: the duration of diabetes in years; x_4: the percentage of glycosylated hemoglobin; x_5: gender, using 1 for male and 0 for female, x_6: indicator for the presence of insulin protein; x_7: area of residence (1 = urban, 2 = rural).

The response variable is completely observed, as are variables x_2, x_3, x_5, x_6, and x_7. Variable x_1 contains missing values for 51 out of the 996 cases, and variable x_4 contains 46 missing values. For three cases both variables x_1 and x_4 are missing.

The data are modeled by means of a logistic regression model, the full model takes the form

$$\text{logit}P(Y = 1 \,|\, x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_7 x_7.$$

Our goal is to perform variable selection in the logistic regression model, including all cases. That is, we do not wish to remove the cases with missing values. The model for the joint covariate distribution of x_1, x_4 is given by

$$f(x_{i1}, x_{i4} \,|\, \boldsymbol{v}_i, \boldsymbol{\alpha}) = f(x_{i1} \,|\, x_{i4}, \boldsymbol{v}_i, \alpha_1) f(x_{i4} \,|\, \boldsymbol{v}_i, \alpha_2)$$

with $\boldsymbol{v}_i = (x_2,\ x_3,\ x_5,\ x_6,\ x_7)$. Because both variables are continuous, $(x_{i1},\ x_{i4} \,|\, v_i,\ \alpha)$ is modeled as a bivariate normal distribution with mean $\boldsymbol{\mu}' = (\mu_{i1},\ \mu_{i2})$, $\mu_{it} = \alpha_{t0} + \alpha_{t1}x_{i2} + \alpha_{t2}x_{i3} + \alpha_{t3}x_{i5} + \alpha_{t4}x_{i6} + \alpha_{t5}x_{i7}$, $t = 1, 2$, and Σ an arbitrary covariance matrix. Fitting the full model, including all the variables, it shows that the following variables are significant at the 5% level: age of the patient, the duration of diabetes, the percentage of glycosylated hemoglobin, and gender, with the last three being highly significant; for further discussion we refer to Klein et al. (1984). We wish to possibly reduce the number of variables using model selection criteria by choosing that model among those that optimizes the criteria. With p potential covariates, there are 2^p possible models that can be considered; in our case the total number of all possible submodels amounts to $2^7 = 128$. We carry out the estimation of all the possible models and we compare the results of the AIC_1 and the TIC_1 (both taking care of the missing observations), and the AIC for the complete cases only that ignores the missingness. Tables 2 and 3 show for each criterion the 10 best models. The tables contain the value of the criterion for each of those models, together with the $Q^{(1)}$ function and the penalty term.

Table 2

Results of variable selection for the WESDR data. The table displays the best 10 *models selected by the different criteria, the value of the criterion for each of these models, together with the value of the* $Q^{(1)}$ *function and the penalty used for* (*a*) TIC_1, *and* (*b*) AIC_1. *All models contain an intercept.*

	Variables	Criterion	$Q^{(1)}$ function	Penalty term
(a)		TIC_1		
1.	x_2, x_3, x_4, x_5	932.19	460.76	5.33
2.	x_1, x_2, x_3, x_4, x_5	932.56	459.88	6.40
3.	x_2, x_3, x_4, x_5, x_7	932.86	460.09	6.34
4.	x_2, x_3, x_4, x_5, x_6	933.01	460.17	6.34
5.	x_1, x_2, x_3, x_4, x_5, x_7	933.38	459.28	7.41
6.	x_1, x_2, x_3, x_4, x_5, x_6	933.59	459.40	7.39
7.	x_2, x_3, x_4, x_5, x_6, x_7	933.98	459.63	7.36
8.	x_1, x_2, x_3, x_4, x_5, x_6, x_7	934.30	458.73	8.41
9.	x_1, x_3, x_4, x_5, x_6	937.10	462.20	6.35
10.	x_3, x_4, x_5	937.18	464.28	4.32
(b)		AIC_1		
1.	x_2, x_3, x_4, x_5	931.52	460.76	5
2.	x_1, x_2, x_3, x_4, x_5	931.76	459.88	6
3.	x_2, x_3, x_4, x_5, x_7	932.17	460.09	6
4.	x_1, x_2, x_3, x_4, x_5, x_7	932.56	459.28	7
5.	x_2, x_3, x_4, x_5, x_6	934.33	460.17	7
6.	x_1, x_2, x_3, x_4, x_5, x_6	934.80	459.40	8
7.	x_2, x_3, x_4, x_5, x_6, x_7	935.27	459.63	8
8.	x_1, x_2, x_3, x_4, x_5, x_6, x_7	935.47	458.73	9
9.	x_1, x_3, x_4, x_5	936.46	463.23	5
10.	x_3, x_4, x_5	936.55	464.28	4

Table 3
Results of variable selection for the WESDR data. The table displays the best 10 *models selected by the AIC_{cc} using the subset of complete cases, the value of the criterion for each of these models, together with the likelihood function and the penalty used. All models contain an intercept.*

	Variables	AIC_{cc}	Likelihood	Penalty term
1.	x_2, x_3, x_4, x_5	849.43	419.71	5
2.	x_2, x_3, x_4, x_5, x_7	850.39	419.20	6
3.	x_1, x_2, x_3, x_4, x_5	850.43	419.21	6
4.	x_1, x_2, x_3, x_4, x_5, x_7	851.37	418.69	7
5.	x_2, x_3, x_4, x_5, x_6	852.03	419.01	7
6.	x_2, x_3, x_4, x_5, x_6, x_7	852.98	418.49	8
7.	x_1, x_2, x_3, x_4, x_5, x_6	853.06	418.53	8
8.	x_3, x_4, x_5	853.39	422.69	4
9.	x_1, x_2, x_3, x_4, x_5, x_6, x_7	854.00	418.00	9
10.	x_1, x_3, x_4, x_5	854.06	422.03	5

The best model for all the criteria is the one that includes all the significant variables in the full model, that is, variables x_2, x_3, x_4, and x_5. The three highly significant variables are always present in all the models displayed. Although for the two criteria TIC_1 and AIC_1 the first three models are the same, they differ from the fourth model onwards. For model 4, the TIC_1 adds to the best model, variable x_6, the indicator for insulin protein, whereas the AIC_1 adds variable x_7, the area of residence of the patients. The difference between the TIC_1 and AIC_1 selected models is due to the penalty term used for calculating the criteria, because the $Q^{(1)}$ function is the same. In contrast to the simulation study, the penalty term in the TIC_1 criteria here is slightly larger than the exact number of parameters. The former criteria differ significantly from the AIC_{cc}, showing a different model order, starting with the second-best model. This illustrates that removing missing observations from the dataset could be a risky tool for dealing with missing observations for the purpose of model selection.

6. Discussion

We introduced new criteria for model selection in the presence of missing data, by the use of the EM algorithm and the weighting method of Ibrahim (1990). The new criteria are immediately obtained from the EM algorithm and can be directly compared to the AIC and TIC in case no observations are missing. The validity of the criteria is investigated in a simulation study and through data analysis. The results have confirmed the good performance of the criteria, in particular their efficiency to deal with the missingness. Ignoring the missing cases does not work well for model selection.

Although we focused on missing covariate data with an ignorable missingness mechanism in this article, future work will extend these results to include missing response data and nonignorable missingness schemes.

Hurvich and Tsai (1989) proposed a small sample adjustment to the AIC, which has been shown to better approximate the Kullback–Leibler distance for linear regression and autoregressive time-series models. In our setting of missing covariate data in the situation of normal linear regression, this gives

$$\mathrm{AIC}_{1,C} = -2Q^{(1)}(\hat{\boldsymbol{\beta}} \,|\, \hat{\boldsymbol{\theta}}) + 2\frac{p_\beta n}{n - p_\beta - 1}.$$

A similar derivation of the corrected AIC, starting directly with the Kullback–Leibler distance rather than with Taylor series expansions, has been done for the case of generalized linear models (including logistic regression). These results can be found in Claeskens and Hjort (2008). Such a derivation for the case of missing covariates is an interesting idea for further research.

An alternative of using the function Q for the construction of the AIC or TIC is a direct use of the observed data likelihood, computed via $L(\theta) = \int P(Y, X \,|\, \theta)\, dX_{\mathrm{mis}}$, which is valid under the MAR assumption and distinctness of the parameters. The relation between the function Q and the likelihood $L(\theta)$ is

$$Q(\theta \,|\, \theta_1) = L(\theta) + \sum_{i=1}^{n} E_{\theta_1}[\log f(X_{\mathrm{mis},i} \,|\, x_{\mathrm{obs},i}, y_i, \theta)],$$

where E_{θ_1} denotes taking the expectation using density $f(X_{\mathrm{mis},i} \,|\, x_{\mathrm{obs},i},\ y_i,\ \theta_1)$. A calculation of L is possible in certain cases, such as with monotone patterns of missingness, but not in general. One could obtain the estimator $\hat{\theta}$ using the EM algorithm, as described above, and then proceed by approximating the integral, for which several methods exist (numerical integration, Laplace approximation, bridge sampling, etc.). We avoid this additional step by directly using Q. For $\theta = \theta_1$, the difference between Q and L is equal to the negative of the entropy, $\sum_{i=1}^{n} E_\theta[\log w_i]$. This quantity determines the difference between the two approaches, and a further investigation in the model selection framework seems worthwhile.

An associate editor corresponded to us that independent work related to the results presented in this article is being developed by Ibrahim, Zhu, and Tang (personal communication).

Acknowledgements

The authors wish to express their thanks to Profs. Ibrahim and Chen for providing the Fortran code of their programs and to Dr R. Klein for giving permission to use the WESDR data. They further thank all reviewers of this article for their constructive remarks and suggestions. This research is supported by the Fund for Scientific Research Flanders (G0542.06).

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory*, B. Petrov and F. Csáki (eds), 267–281. Budapest, Hungary: Akadémiai Kiadó.

Burnham, K. P. and Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, 2nd edition. New York: Springer-Verlag.

Cavanaugh, J. E. and Shumway, R. H. (1998). An Akaike information criterion for model selection in the presence of incomplete data. *Journal of Statistical Planning and Inference* **67,** 45–65.

Claeskens, G. and Hjort, N. L. (2008). *Model Selection and Model Averaging*. Cambridge, U.K.: Cambridge University Press.

Gilks, W. R. and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Applied Statistics* **41,** 337–348.

Hens, N., Aerts, M., and Molenberghs, G. (2006). Model selection for incomplete and design-based samples. *Statistics in Medicine* **25,** 2502–2520.

Hurvich, C. M. and Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika* **76,** 297–307.

Ibrahim, J. G. (1990). Incomplete data in generalized linear models. *Journal of the American Statistical Association* **85,** 765–769.

Ibrahim, J. G., Chen, M.-H., and Lipsitz, S. R. (1999). Monte Carlo EM for missing covariates in parametric regression models. *Biometrics* **55,** 591–596.

Ibrahim, J. G., Lipsitz, S. R., and Chen, M.-H. (1999). Missing covariates in generalized linear models when the missing data mechanism is non-ignorable. *Journal of the Royal Statistical Society, Series B* **61,** 173–190.

Klein, R., Klein, B. E. K., Moss, S. E., Davis, M. D., and DeMets, D. L. (1984). The Wisconsin epidemiologic study of diabetic retinopathy: II. Prevalence and risk of diabetic retinopathy when age at diagnosis is less than 30 years. *Archives of Ophthalmology* **102,** 520–526.

Kotz, S. and Nadarajah, S. (2004). *Multivariate t Distributions and Their Applications*. Cambridge, U.K.: Cambridge University Press.

Linhart, H. and Zucchini, W. (1986). *Model Selection*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. New York: Wiley.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. Wiley Series in Probability and Statistics, 2nd edition. Hoboken, New Jersey: Wiley.

Liu, C. (1995). Missing data imputation using the multivariate t distribution. *Journal of Multivariate Analysis* **53,** 139–158.

Liu, C. and Rubin, D. B. (1995). ML estimation of the multivariate t distribution with unknown degrees of freedom. *Statistica Sinica* **5,** 19–39.

Louis, T. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B* **44,** 226–233.

Meng, X. L. and Rubin, D. B. (1991). Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. *Journal of the American Statistical Association* **86,** 899–909.

Nielsen, S. F. (2000). The stochastic EM algorithm: Estimation and asymptotic results. *Bernoulli* **6,** 457–489.

Shimodaira, H. (1994). A new criterion for selecting models from partially observed data. In *Selecting Models from Data: Artificial Intelligence and Statistics IV*, P. Cheeseman and R. W. Oldford (eds), 21–29. New York: Springer.

Takeuchi, K. (1976). Distribution of informational statistics and a criterion of model fitting. *Suri-Kagaku (Mathematical Sciences)* **153,** 12–18. In Japanese.

Received March 2007. *Revised November* 2007. *Accepted December* 2007.

Appendix

Proof of Theorem 1

We use a Taylor expansion of $\log \tilde{f}(\tilde{\boldsymbol{Y}}, \tilde{\boldsymbol{X}}; \hat{\boldsymbol{\theta}})$ about $\boldsymbol{\theta}_0$ and take the expectation with respect to the true distribution of $(\tilde{\boldsymbol{Y}}, \tilde{\boldsymbol{X}})$ to obtain

$$\begin{aligned} &E_g\{\log \tilde{f}(\tilde{\boldsymbol{Y}}, \tilde{\boldsymbol{X}}; \hat{\boldsymbol{\theta}}) \,|\, \hat{\boldsymbol{\theta}}\} \\ &\quad = E_g\{\log \tilde{f}(\tilde{\boldsymbol{Y}}, \tilde{\boldsymbol{X}}; \boldsymbol{\theta}_0)\} + (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)' E_g\left\{ \frac{\partial}{\partial \boldsymbol{\theta}} \log \tilde{f}(\tilde{\boldsymbol{Y}}, \tilde{\boldsymbol{X}}; \boldsymbol{\theta}_0) \right\} \\ &\qquad - \frac{1}{2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)' E_g \left\{ -\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \log \tilde{f}(\tilde{\boldsymbol{Y}}, \tilde{\boldsymbol{X}}; \boldsymbol{\theta}_0) \right\} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \\ &\qquad + o_p\big(\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|^2\big). \end{aligned} \tag{A.1}$$

By definition of the least false parameter $E_g\{\frac{\partial}{\partial \boldsymbol{\theta}} \log \tilde{f}(\tilde{\boldsymbol{Y}}, \tilde{\boldsymbol{X}}; \boldsymbol{\theta}_0)\} = \mathbf{0}$ and for the third term $E_g\{-\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \log \tilde{f}(\tilde{\boldsymbol{Y}}, \tilde{\boldsymbol{X}}_{\text{obs}}; \boldsymbol{\theta}_0)\} = nI(\boldsymbol{\theta}_0)$. Hence (A.1) reduces to $E_g\{\log \tilde{f}(\tilde{\boldsymbol{Y}}, \tilde{\boldsymbol{X}}; \hat{\boldsymbol{\theta}}) \,|\, \hat{\boldsymbol{\theta}}\} = E_g\{\log \tilde{f}(\tilde{\boldsymbol{Y}}, \tilde{\boldsymbol{X}}; \boldsymbol{\theta}_0)\} - \frac{n}{2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)' I(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + o_p(\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|^2)$. After taking expectations and using the uniform integrability assumption, $K_n = E_g\{E_{\hat{g}}\{\log \tilde{f}(\tilde{\boldsymbol{Y}}, \tilde{\boldsymbol{X}}; \boldsymbol{\theta}_0)\}\}/n = E_g\{\log f(\tilde{\boldsymbol{Y}}, \tilde{\boldsymbol{X}}; \boldsymbol{\theta}_0)\}/n - \frac{1}{2} E_g\{(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)' I(\boldsymbol{\theta}_0) \times (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\} + o(1/n)$. The second term hereof is equal to $\text{tr}\{I(\boldsymbol{\theta}_0) \text{Var}_g \hat{\boldsymbol{\theta}}(\boldsymbol{Y}, \boldsymbol{X})\}/2$, where the variance is obtained from a two-step Taylor series expansion of the first derivative of the Q function, which leads to

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = \{-\ddot{Q}(\boldsymbol{\theta}_0 \,|\, \boldsymbol{\theta}_0)/n\}^{-1} \frac{1}{\sqrt{n}} \dot{Q}(\boldsymbol{\theta}_0 \,|\, \boldsymbol{\theta}_0) + o_p(1). \tag{A.2}$$

By definition of the matrices I and J, the trace expression simplifies to $\text{tr}\{J(\boldsymbol{\theta}_0) I^{-1}(\boldsymbol{\theta}_0)\}/(2n)$. This leads to $nK_n = E_g\{\log f(\tilde{\boldsymbol{Y}}, \tilde{\boldsymbol{X}}; \boldsymbol{\theta}_0)\} - \frac{1}{2}\text{tr}\{J(\boldsymbol{\theta}_0) I^{-1}(\boldsymbol{\theta}_0)\} + o(1)$. A similar Taylor series expansion shows that, omitting smaller-order terms that converge to zero, $\hat{K}_n \doteq Q(\boldsymbol{\theta}_0 \,|\, \boldsymbol{\theta}_0)/n + \dot{Q}(\boldsymbol{\theta}_0 \,|\, \boldsymbol{\theta}_0)'(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)/n - \frac{1}{2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)' I(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$. By representation (A.2) and previous arguments, $\hat{K}_n = E_g\{\log \tilde{f}(\tilde{\boldsymbol{Y}}, \tilde{\boldsymbol{X}}; \boldsymbol{\theta}_0)\}/n + \frac{1}{2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)' I(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + r_n + o_p(1/n)$, where r_n is an average of independent and identically distributed (i.i.d.) random variables with mean zero. Taking the expectation of $\hat{K}_n$ and combining the result with that found for K_n yields the stated result.