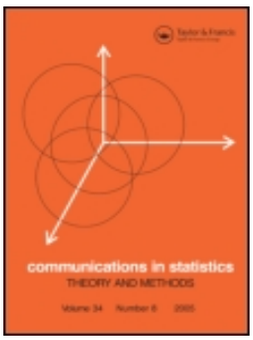

Taylor & Francis
Taylor & Francis Group

Communications in Statistics - Theory and Methods

ISSN: 0361-0926 (Print) 1532-415X (Online) Journal homepage: http://www.tandfonline.com/loi/lsta20

Consistency of information criteria for model selection with missing data

Abdelaziz El Matouat, Freedath Djibril Moussa & Hassania Hamzaoui

To cite this article: Abdelaziz El Matouat, Freedath Djibril Moussa & Hassania Hamzaoui (2016) Consistency of information criteria for model selection with missing data, Communications in Statistics - Theory and Methods, 45:23, 6900-6914, DOI: 10.1080/03610926.2014.972568

To link to this article: http://dx.doi.org/10.1080/03610926.2014.972568

Accepted author version posted online: 28 Jan 2016.
Published online: 28 Jan 2016.

Submit your article to this journal

Article views: 53

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
http://www.tandfonline.com/action/journalInformation?journalCode=lsta20

Download by: [Abdelaziz EL MATOUAT] **Date:** 12 March 2017, At: 14:43

COMMUNICATIONS IN STATISTICS—THEORY AND METHODS
2016, VOL. 45, NO. 23, 6900–6914
http://dx.doi.org/10.1080/03610926.2014.972568

Taylor & Francis
Taylor & Francis Group

Consistency of information criteria for model selection with missing data

Abdelaziz El Matouat[a], Freedath Djibril Moussa[b], and Hassania Hamzaoui[b]

[a]LMAH, University of Le Havre, Le Havre, France; [b]LIM, Math. Dept., Sidi Mohamed Ben Abdellah University, Fez, Morocco

ABSTRACT

In this paper, we investigate the consistency of the Expectation Maximization (*EM*) algorithm-based information criteria for model selection with missing data. The criteria correspond to a penalization of the conditional expectation of the complete data log-likelihood given the observed data and with respect to the missing data conditional density. We present asymptotic properties related to maximum likelihood estimation in the presence of incomplete data and we provide sufficient conditions for the consistency of model selection by minimizing the information criteria. Their finite sample performance is illustrated through simulation and real data studies.

ARTICLE HISTORY
Received 3 October 2013
Accepted 23 September 2014

KEYWORDS
Consistency; *EM* algorithm; Information criteria; Kullback–Leibler divergence; Missing data.

MATHEMATICS SUBJECT CLASSIFICATION
62F12

1. Introduction

Consider observations from an unknown distribution and postulate a set of possible families of distributions for the data generating process. A widely used method for selecting one of the candidate families is the minimization of penalized log-likelihood criteria. This approach was first suggested by Akaike (1973) who derived Akaike information criterion (*AIC*) from the expected Kullback–Leibler discrepancy between the data generating distribution and an approximating one. Schwarz (1978) formulated Bayesian information criterion (*BIC*) in a Bayesian model selection perspective. Several other information criteria have been proposed in the statistical literature (see, e.g., Hannan and Quinn, 1979; Bozdogan, 1988; Hurvich and Tsai, 1989) and they all differ by the penalty factor in the penalization of the log-likelihood. Nishii (1988) gave conditions on the penalty factor to ensure the consistency of the model selected with penalized likelihood criteria.

Typically, these criteria depend on the observed data likelihood function. For missing data settings, trying to compute for example *AIC* or *BIC* can be very challenging due to intractable observed data likelihood. Moreover, the unobserved data are ignored in the model selection procedure with classical criteria. Therefore, specific criteria based on the Expectation Maximization (*EM*) algorithm (Dempster et al., 1977) have been proposed for incomplete data situations. The *EM* algorithm is a tool used to find maximum likelihood estimates (MLEs) of the parameter of a statistical model, in presence of missing data. Shimodaira (1994) introduced a natural extension of *AIC* called Predictive Divergence for Indirect Observation (*PDIO*)

CONTACT Abdelaziz El Matouat abdelaziz.el-matouat@univ-lehavre.fr University of Le Havre, 25 rue Philippe Lebon, BP420 Le Havre 76057, France.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/lsta.
© 2016 Taylor & Francis Group, LLC

models still depending on the observed log-likelihood; but the penalization term in his criterion takes into account the presence of missing data. On the motivation provided by *PDIO*, Cavanaugh and Shumway (1998) derived a variant of *AIC* denoted by AIC_{cd} (cd standing for complete data), which can be evaluated using only complete-data tools. To simplify the computation of traditional criteria like *AIC* or *BIC* in missing data problems, Ibrahim et al. (2008) defined a class of criteria called IC_Q (*IC* for information criterion). AIC_{cd} and IC_Q criteria are formulated as a penalization of the expected complete data log-likelihood with respect to the missing data conditional density and given the observed data.

In this paper, we consider a class IC_{cd} of criteria that includes AIC_{cd} for missing data settings. In IC_{cd} formula, we use the well-known general form of the penalty factor for information criteria, in model selection theory. Given incomplete observations from an unknown distribution, we consider a family of parametric distributions which provide approximation of the true distribution and we investigate the consistency of model selection procedure using the IC_Q and IC_{cd}criteria. We start by studying the asymptotic behavior of the expected complete data log-likelihood found with the *EM* algorithm. An important property of this algorithm is that the parameter estimate obtained is the MLE for the observed data. Hence, in missing data situations, the classical results on maximum likelihood estimation (in absence of missing data) are observed data-based inference results. The asymptotic results on maximum likelihood estimation with missing data are used to derive sufficient conditions for consistent model selection with IC_Q and IC_{cd} criteria.

The paper is organized as follows: in Section 2 we describe the *EM* algorithm, we list all the assumptions on the log-likelihood, and we present asymptotic properties related to the order of consistency of the maximum likelihood estimation with missing data. Section 3 is devoted to model selection. We present the IC_{cd} and IC_Q criteria and we give sufficient conditions for their consistency. In Section 4, simulation results for autoregressive model selection from incomplete data, linear regression with missing covariate, and real data analysis are presented to illustrate the finite sample performance of the criteria.

2. Maximum likelihood estimation with missing data

Let X be a random variable with unknown probability density function g and $\underline{x} = (x_{obs}, x_{mis})$ be an incomplete sample of size n of X where x_{obs} denotes the observed data and x_{mis} the missing data. Consider a parametric family of candidate densities:

$$\mathcal{M} = \{f(.,\theta), \theta \in \Theta\}$$

where Θ is a convex subset of $\mathbb{R}^p$ and the parameter vector θ is supposed unknown. Assume that $\int |\log g(x)| g(x) dx < \infty$. The Kullback–Leibler divergence between g and $f(.,\theta)$ is

$$KL(g;\, f, \theta) = \int g(x) \log \frac{g(x)}{f(x,\theta)} dx$$

Define the quasi true parameter θ_g as

$$\theta_g = \arg\min_{\theta \in \Theta} KL(g;\, f, \theta)$$

The density $f(.,\theta_g)$ is called the quasi true density. If g lies in the candidate family then $f(.,\theta_g) = g$ since the Kullback–Leibler divergence is always non negative and is zero only if the two densities are identical almost everywhere.

Let $e(g;\ f, \theta) = \int g(x) \log f(x, \theta) dx$. Then, we have

$$e(g;\ f, \theta_g) \geq\ e(g;\ f, \theta)\ \forall\, \theta \in \Theta$$

Set the quasi log-likelihood

$$l(\theta, \underline{x}) = \log f(\underline{x}, \theta) \quad \text{and} \quad l(\theta, x_{obs}) = \log f(x_{obs}, \theta)$$

Since $\underline{x}$ is not fully observed, the quasi maximum likelihood estimator (QMLE) $\hat{\theta}$ of θ is obtained via the *EM* algorithm.

2.1 The EM algorithm

The *EM* algorithm (Dempster et al., 1977) is an iterative method to find MLEs from incomplete data. In many cases, the maximization of the observed data log-likelihood $l(\theta, x_{obs})$ is more difficult than that of the complete data log-likelihood $l(\theta, \underline{x})$. The idea behind the *EM* algorithm is to obtain the MLE based on x_{obs} by iteratively maximizing the conditional expectation Q of $l(\theta, \underline{x})$ with respect to the unknown data, given the observed data and the current parameter estimate. The algorithm may be described as follows:

set $\theta^{(0)}$ a starting value of θ. Each iteration $t = 1, 2, \ldots$, consists of two steps.

E step (Expectation): let $\theta^{(t-1)}$ be the current value of θ. Compute

$$Q(\theta; \theta^{(t-1)}) = \int l(\theta, \underline{x}) f(x_{mis}|x_{obs}\theta^{(t-1)}) dx_{mis}$$

M step (Maximization): choose $\theta^{(t)}$ to be any value of $\theta \in \Theta$ that maximizes $Q(\theta; \theta^{(t-1)})$.

The *E* and *M* steps are repeated until the algorithm converges. The QMLE $\hat{\theta}$ satisfies

$$Q(\hat{\theta}; \hat{\theta}) \geq Q(\theta; \hat{\theta}) \qquad \forall \theta \in \Theta$$

Dempster et al. (1977) showed that the observed data log-likelihood $l(\theta, x_{obs})$ does not decrease after an *EM* iteration. Hence the sequence of values $\{l(\theta^{(t)}, x_{obs}), t = 0, 1, \cdots\}$ is monotonically increasing.

Under regularity conditions given by Wu (1983), when the starting value $\theta^{(0)}$ is close to $\hat{\theta}$,

$$\lim_{t\to\infty} \theta^{(t)} = \hat{\theta} \text{ and } \lim_{t\to\infty} l(\theta^{(t)}, x_{obs}) = l(\hat{\theta}, x_{obs})$$

The *EM* algorithm is self-consistent (see, e.g., McLachlan and Krishnan, 1997):

$$\text{for } \hat{\theta} = \lim_{t\to\infty} \theta^{(t)},\ \text{we have}\ \ Q(\hat{\theta}; \hat{\theta}) \geq Q(\theta; \hat{\theta})\ \forall\, \theta\ \in\ \Theta$$

$$\text{implying} \qquad l(\hat{\theta}, x_{obs})\ \geq l(\theta, x_{obs})\ \forall\, \theta\ \in\ \Theta$$

$$\text{Reversely, if} \qquad l(\hat{\theta}, x_{obs}) \geq l(\theta, x_{obs})\ \forall\, \theta \in \Theta$$

$$\hat{\theta}\ \text{satisfies} \qquad Q(\hat{\theta}; \hat{\theta}) \geq Q(\theta; \hat{\theta})\ \forall\, \theta\ \in\ \Theta$$

Next, we present the results related to the order of consistency of based maximum likelihood estimation using the *EM* algorithm.

2.2 Asymptotic results

We make the following assumptions on $(g, \mathcal{M})$ to study the asymptotic behavior of maximum likelihood principle with missing data.

Suppose that the observed data
$x_{obs} = (x_{obs,1}, \ldots, x_{obs,n_{obs}})$ are i.i.d. and the missing data
$x_{mis} = (x_{mis,1}, \ldots, x_{mis,n_{mis}})$ are i.i.d. with $n = n_{obs} + n_{mis}$. We have

$$f(\underset{\sim}{x}, \theta) = f(x_{obs}, \theta) f(x_{mis}|x_{obs}, \theta) \tag{1}$$

where

$$f(x_{obs}, \theta) = \prod_{i=1}^{n_{obs}} f(x_{obs,i}, \theta) \text{ and } f(x_{mis}|x_{obs}, \theta) = \prod_{i=1}^{n_{mis}} f(x_{mis,i}|x_{obs}, \theta)$$

Assumption A1. The quasi true parameter θ_g exists and is a unique point of $\mathring{\Theta}$.

Assumption A2. (a) The functions $l(\theta, x_{obs,i}) = \log f(x_{obs,i}, \theta)$,
$i = 1, \ldots, n_{obs}$, are measurable with respect to $x_{obs,i}$ for each $\theta \in \Theta$, continuous with respect to θ for all θ in Θ and three times differentiable over $\mathring{\Theta}$.

The derivatives $l_\alpha(\theta, x_{obs,i}) = \dfrac{\partial}{\partial\theta_\alpha} l(\theta, x_{obs,i})$, $l_{\alpha\beta}(\theta, x_{obs,i}) = \dfrac{\partial^2 l(\theta, x_{obs,i})}{\partial\theta_\alpha \partial\theta_\beta}$ and $l_{\alpha\beta\gamma}(\theta, x_{obs,i}) = \dfrac{\partial^3 l(\theta, x_{obs,i})}{\partial\theta_\alpha \partial\theta_\beta \partial\theta_\gamma}$, $\alpha, \beta, \gamma = 1, \ldots, p$, are measurable with respect to $x_{obs,i}$ for each θ and continuous with respect to θ for each $x_{obs,i}$.

(b) $|l(x_{obs,i}, \theta)|$, $|l_\alpha(x_{obs,i}, \theta)|$, $|l_\alpha(\theta, x_{obs,i})|^2$, $|l_{\alpha\beta}(\theta, x_{obs,i})|$, $|l_{\alpha\beta}(\theta, x_{obs,i})|^2$, $|l_\alpha(\theta, x_{obs,i}) l_\beta(\theta, x_{obs,i})|$, and $|l_{\alpha\beta\gamma}(x_{obs,i}, \theta)|$ $(\alpha, \beta, \gamma = 1, \cdots, p)$ are dominated by integrable functions with respect to $g(x_{obs,i})$ which do not depend on θ.

Assumption A3. Define $p \times p$ matrices $V(\theta)$ and $W(\theta)$ by

$$V(\theta) = E_g\left[\frac{\partial}{\partial\theta} l(\theta, X) \frac{\partial}{\partial\theta^T} l(\theta, X)\right] \text{ and } W(\theta) = -E_g\left[\frac{\partial^2}{\partial\theta\partial\theta^T} l(\theta, X)\right]$$

where E_g denotes the expectation with respect to the true density g and $l(\theta, X) = \log f(X, \theta)$. $V(\theta_g)$ and $W(\theta_g)$ are definite positive.

Assumption A4. There exists the QMLE $\hat{\theta} \in \mathring{\Theta}$ and $\hat{\theta}$ tends to θ_g almost surely as n_{obs} tends to ∞.

Those assumptions are the regularity conditions to ensure that $\hat{\theta}$ is strongly consistent in the classical theory, and to evaluate its order of consistency.

We now state results on the order of consistency related to the maximum likelihood estimation based on the observed data. By the self-consistency property of the *EM* algorithm (the estimate obtained from the *EM* algorithm is the MLEs for the observed data x_{obs}), the following theorem holds.

Theorem 1. *Under Assumptions A1–A4, we have*

(a) $\hat{\theta} = \theta_g + \mathcal{O}\left(\sqrt{n_{obs}^{-1} \log\log n_{obs}}\right)$ *a.s.,*

(b) $l(\hat{\theta}, x_{obs}) = l(\theta_g, x_{obs}) + \mathcal{O}\left(\log\log n_{obs}\right)$ *a.s.,*

(c) $\dfrac{1}{n_{obs}} l(\hat{\theta}, x_{obs}) = e(g, f, \theta_g) + \mathcal{O}\left(\sqrt{n_{obs}^{-1} \log\log n_{obs}}\right)$ *a.s.*

Proof. See Theorem 1 in Nishii (1988).

In the next section, we present a class of information criteria for model selection with incomplete data and give consistency conditions for missing data information criteria. □

3. Model selection from incomplete data

Let us consider two families of candidate densities:

$$\mathcal{M}_{k_j} = \{f(.,\theta(k_j)), \theta(k_j) \in \Theta_{k_j}\},\ j = 1, 2$$

and suppose both $\left(g, \mathcal{M}_{k_j}\right),\ j = 1, 2$ satisfy Assumptions A1–A4. We want to choose the family that is closer to g with a likelihood ratio approach. Let $\theta_g(k_j)$ be the quasi true parameter, $\hat{\theta}(k_j)$, the QMLE in Θ_{k_j} and set

$$e(g, f, \theta(k_j)) = \int g(x) \log f(x, \theta(k_j))dx$$

which is the maximized log-likelihood in $\mathcal{M}_{k_j}$. Consider the hypothesis test

$$H_0 :\ e(g, f, \theta_g(k_1)) = e(g, f, \theta_g(k_2)) \text{ versus } H_1 :\ e(g, f, \theta_g(k_1)) > e(g, f, \theta_g(k_2)).$$

The likelihood ratio $\lambda_{n_{obs}} = \sum_{i=1}^{n_{obs}} \log \frac{f(x_{obs,i}, \hat{\theta}(k_1))}{f(x_{obs,i}, \hat{\theta}(k_2))}$ can asymptotically find the family that is closer to g. From Nishii (1988), if Assumptions A1–A4 hold, this likelihood ratio test is consistent and under H_1,

$$\frac{1}{n_{obs}}\lambda_{n_{obs}} \to e(g, f, \theta_g(k_1)) - e(g, f, \theta_g(k_2)) > 0\ a.s.\ \text{as}\ n_{obs} \to +\infty$$

But the $\lambda_{n_{obs}}$ may be problematic to compute and generally there are more than two candidates families. So we rather consider the information criteria for incomplete data to choose a model.

3.1 Information criteria for model selection with incomplete data

Consider an incomplete data sample $\underline{x} = (x_{obs}, x_{mis})$ of size n and K candidate models:

$$\mathcal{M}_k = \{f(.,\theta(k)), \theta(k) \in \Theta_k\},\ k = 1, \ldots, K$$

Set

$$I_o(\theta(k)|x_{obs}) = -\frac{\partial^2 \log f(x_{obs}, \theta(k))}{\partial\theta(k)\partial\theta(k)^T} \text{ the observed data information matrix and}$$

$$I_{oc}(\theta(k)|x_{obs}) = \int \left\{-\frac{\partial^2 \log f(\underline{x}, \theta(k))}{\partial\theta(k)\partial\theta(k)^T}\right\} f(x_{mis}|x_{obs}, \theta(k))dx_{mis} \quad \text{the conditional}$$

expectation of the complete data information matrix with respect to the missing data conditional density.

Cavanaugh and Shumway (1998) formulated AIC_{cd} criterion which is an estimation of the expected Kullback-Leibler divergence between a candidate density f and the true unknown density g, based on the complete data.

$$AIC_{cd}(k) = -2Q(\hat{\theta}(k); \hat{\theta}(k)) + 2trace[I_{oc}(\hat{\theta}(k)|x_{obs}).I_o^{-1}(\hat{\theta}(k)|x_{obs})]$$

where k is the number of free parameters of the candidate model $\mathcal{M}_k$ and $\hat{\theta}(k)$ the QMLE in $\mathcal{M}_k$. Using the results in Meng and Rubin (1991),

$$trace[I_{oc}(\hat{\theta}(k)|x_{obs})I_o^{-1}(\hat{\theta}(k)|x_{obs})]$$

can be written as

$$k + trace[DF_k(I_k - DF_k)^{-1}]$$

where DF_k is the Jacobian matrix at $\hat{\theta}(k)$ of the map defined by the *EM* algorithm and I_k is the identity matrix of size k. Hence

$$AIC_{cd}(k) = -2Q(\hat{\theta}(k); \hat{\theta}(k)) + 2\{k + trace[DF_k(I_k - DF_k)^{-1}]\}$$

The matrix DF_k is evaluated by numerical differentiation or using the *SEM* algorithm (Meng and Rubin, 1991). The term $trace[DF_k(I_k - DF_k)^{-1}]$ is always non negative, positive when $\underline{x} \neq x_{obs}$ and is interpreted as a measure of the extent to which the missing data affect the fitted model (Cavanaugh and Shumway, 1998); this term is large when the extent of missingness is substantial and the model is relatively complex.

We consider a general form of criteria that includes AIC_{cd}, defined as

$$IC_{cd}(k) = -2Q(\hat{\theta}(k); \hat{\theta}(k)) + C_n\{k + trace[DF_k(I_k - DF_k)^{-1}]\}$$

where C_n is the general penalty factor, depending on the sample size n, in model selection theory. In the AIC_{cd} formulation, $C_n = 2$. By choosing $C_n = \log n$, we obtain the BIC_{cd} criterion.

Ibrahim et al. (2008) proposed a different class of missing data information criteria denoted IC_Q and written as follows:

$$IC_Q(k) = -2Q(\hat{\theta}(k); \hat{\theta}(k)) + pen(\hat{\theta}(k))$$

where $pen(\hat{\theta}(k))$ is a penalty term that is a function of the data and the fitted model. $pen(\hat{\theta}(k))$ is not specific to missing data problems and can be chosen among the various penalties suggested in the literature on model selection. For instance, $pen(\hat{\theta}(k)) = 2k$ and $pen(\hat{\theta}(k)) = k \log n$ correspond respectively to AIC_Q and BIC_Q. The classical criteria penalization is generally a linear function of the model dimension thus $pen(\hat{\theta}(k)) = kC_n$ where C_n is one of the existing penalty factors. In the remainder, we consider

$$IC_Q(k) = -2Q(\hat{\theta}(k); \hat{\theta}(k)) + kC_n$$

The IC_{cd} and IC_Q criteria share the same goodness of fit term $-2Q(\hat{\theta}_k; \hat{\theta}_k)$ yet differ in their penalization term. The penalization in IC_{cd} is greater than that in IC_Q and the additional term $C_n trace[DF_k(I_k - DF_k)^{-1}]$ reflects the impact of the missing data of the fitted model. Thus IC_{cd} can be viewed as an over penalization of IC_Q according to the impact of missing data on the fitted model. This implies that for a given C_n, the model picked by IC_{cd} will not be larger than the one favored by IC_Q. Ibrahim et al. (2008) mentioned that IC_Q can lead to poor model selection especially when the fraction of missing data is high, and simulation results indicate that for the same penalty factor C_n, IC_{cd} outperforms IC_Q.

As an illustration, we compared the performance of AIC_Q, AIC_{cd}, BIC_Q, and BIC_{cd} criteria using the following model:

$$AR(2): \qquad X_t = 1.4X_{t-1} - .49X_{t-2} + \varepsilon_t \text{ where } \{\varepsilon_t\} \sim \mathcal{N}(0, 1)$$

We considered 100 samples of size $n = 200$ and we created incomplete data sets by discarding at random $P_{mis} = 20\%$ and $P_{mis} = 40\%$ of the simulated data. Candidate models $AR(k)$, $k = 1, \ldots, 10$, are fitted to each sample using the *EM* algorithm and the order is estimated with the criteria. The parameters estimates obtained from the largest block of observed data serve as initial values for the algorithm. The frequencies of the estimated orders are reported in Table 1.

Table 1. Frequencies of estimated orders based on 100 samples of size $n = 200$.

		Orders				
P_{mis}	Criteria	**1**	**2**	**3**	**4**	**5–10**
20%	AIC_Q	0	54	16	9	21
	BIC_Q	0	86	12	1	1
	AIC_{cd}	0	62	17	6	15
	BIC_{cd}	0	90	10	0	0
40%	AIC_Q	0	39	29	11	21
	BIC_Q	3	71	21	3	2
	AIC_{cd}	0	58	27	5	10
	BIC_{cd}	10	79	11	0	0

From these results, we observe that the performances of the criteria are greatly affected by the amount of missingness. As the percentage of missing data in the sample grows, the frequencies of the correct order selection decrease yet remain higher with AIC_{cd} and BIC_{cd} when compared to AIC_Q and BIC_Q, respectively. Hence in practice, IC_{cd} criteria should be considered rather than IC_Q criteria when the fraction of missing data is large.

In absence of missing data ($\underline{x} = x_{obs}$), $Q(\hat{\theta}(k); \hat{\theta}(k)) = \log f(x_{obs}, \hat{\theta}(k))$, $DF_k = 0$, and both IC_Q and IC_{cd} criteria reduce to

$$IC(k) = -2 \log f(x_{obs}, \hat{\theta}(k)) + kC_n$$

which is the general form of traditional information criteria. It is clear that IC_{cd} and IC_Q should possess the same asymptotic properties as IC when they share the same penalty factor C_n and under the assumption that the number n_{mis} of missing data tends to zero as the size n of the whole sample tends to infinity. We next establish the consistency properties of IC_{cd} and IC_Q, when this assumption is not met.

3.2 Consistency conditions

We consider the case where the models $\mathcal{M}_k$ are nested.

$$\mathcal{M}_k = \{f(., \theta(k)), \theta(k) \in \Theta_k\}, \ k = 1, \ldots, K$$

where $\Theta_1 \subset \Theta_2, \ldots, \subset \Theta_K$ such that

$$\Theta_k = \{\theta(k) = (\theta_1(k), \ldots, \theta_k(k), \theta_{k+1}(k), \ldots, \theta_K(k)) \in \Theta_K | \theta_{k+1}(k) = \ldots = \theta_K(k) = 0\}$$

Assume that the unknown quasi true density $f(., \theta_g)$ lies in model $\mathcal{M}_q$, $1 \leq q \leq K$, then $\mathcal{M}_q$ is called the quasi true model.

For $k = 1, \ldots, K$, we suppose that Assumptions A1–A4 hold for $(g, \mathcal{M}_k)$ and also that

Assumption A5. $\forall \theta(k) \in \Theta_k$, we have

$$H(\theta(k), \theta(k)) = \int \log f(x_{mis}|x_{obs}, \theta(k)) f(x_{mis}|x_{obs}, \theta(k)) dx_{mis} < +\infty$$

Set $\theta_g(k)$ the quasi true parameter and $\hat{\theta}(k)$ the QMLE in the model $\mathcal{M}_k$. Set $\lambda_{n_{obs}} = l(x_{obs}, \hat{\theta}_n(k)) - l(x_{obs}, \hat{\theta}_n(q))$ the likelihood ratio for the models $\mathcal{M}_k$ and $\mathcal{M}_q$. Then, we have:

- If the model $\mathcal{M}_k$ includes the quasi true model $\mathcal{M}_q$ ($\Theta_q \subset \Theta_k$), then $\theta_g(q) = \theta_g(k)$, $e(g, f, \theta_g(k)) = e(g, f, \theta_g(q))$ and $\lambda_{n_{obs}} \geq 0$.

- When $\mathcal{M}_k$ is included in $\mathcal{M}_q$ $(\Theta_k \subset \Theta_q)$ then $e(g, f, \theta_g(k)) < e(g, f, \theta_g(q))$ and $\frac{\lambda_{n_{obs}}}{n_{obs}} \xrightarrow{p.s.} e(g, f, \theta_g(k)) - e(g, f, \theta_g(q)) < 0$.

From the results in Section 2, we derive sufficient conditions for consistent model selection with missing data information criteria.

Theorem 2. *Let $\hat{\mathcal{M}}$ be the model that minimizes the IC_Q criterion among the candidate models $\mathcal{M}_k$, $k = 1, \ldots, K$. Suppose that*

Assumption A6. *There exists $c \in]0, 1[$ such that $c < \frac{n_{obs}}{n} \leq 1$.*

If $\lim_{n\to\infty} \frac{C_n}{n} = 0$ and $\lim_{n_{obs}\to\infty} \frac{C_n}{\log\log n_{obs}} = +\infty$ then $\hat{\mathcal{M}} \xrightarrow{a.s.} \mathcal{M}_q$ as $n_{obs} \to \infty$.

Assumption A6 means that the number of observed data is not insignificant with respect to the size n of the whole sample.

Proof. For a candidate model $\mathcal{M}_k$ and the quasi true model $\mathcal{M}_q$, there are two possible cases: either $\mathcal{M}_q$ is included in $\mathcal{M}_k$ or $\mathcal{M}_k$ is included in $\mathcal{M}_q$. Following Nishii (1988), we study the difference

$$\begin{aligned} \delta IC_Q &= IC_Q(k) - IC_Q(q) \\ &= Q(\hat{\theta}(k); \hat{\theta}(k)) - Q(\hat{\theta}(q)); \hat{\theta}(q)) + C_n(k - q) \\ &= \delta Q + C_n(k - q) \end{aligned} \tag{2}$$

where $\delta Q = Q(\hat{\theta}(k); \hat{\theta}(k)) - Q(\hat{\theta}(q)); \hat{\theta}(q))$. Let

$$\lambda_{n_{obs}} = l(x_{obs}, \hat{\theta}(k)) - l(x_{obs}, \hat{\theta}(q)) \text{ and } \delta H = H(\hat{\theta}(k); \hat{\theta}(k)) - H(\hat{\theta}(q); \hat{\theta}(q)).$$

Then, we have

$$\delta Q = \lambda_{n_{obs}} + \delta H$$

□

First case: If the quasi true model $\mathcal{M}_q$ is included in $\mathcal{M}_k$, then

$$k > q, \qquad \theta_g(q) = \theta_g(k) \quad \text{thus} \quad e(g, f, \theta_g(k)) = e(g, f, \theta_g(q))$$

$\lambda_{n_{obs}} \geq 0$ and by (a) of Theorem 1, we have

$$\begin{aligned} \lambda_{n_{obs}} &= l(x_{obs}, \theta_g(k)) - l(x_{obs}, \theta_g(q)) + \mathcal{O}(\log\log n_{obs}) \text{ a.s.} \\ &= \mathcal{O}(\log\log n_{obs}) \text{ a.s.} \\ &\to +\infty \text{ as } n_{obs} \to \infty \text{ since } \lambda_{n_{obs}} \geq 0 \end{aligned} \tag{3}$$

Using (3) and Assumption A5, we obtain

$$\begin{aligned} \delta Q &= \lambda_{n_{obs}} + \delta H \\ &= \underbrace{\mathcal{O}(\log\log n_{obs})}_{\to+\infty \text{ as } n_{obs}\to\infty} + \underbrace{\delta H}_{<+\infty} \\ &= \mathcal{O}(\log\log n_{obs}) \text{ a.s.} \end{aligned}$$

and $\delta Q \geq 0$. Hence, (2) becomes

$$\begin{aligned} \delta IC_Q &= -2\delta Q + C_n(k - q) \\ &= -\mathcal{O}(\log\log n_{obs}) + C_n(k - q) \text{ a.s.} \end{aligned}$$

$$= \log\log n_{obs}\Big(\frac{C_n(k-q)}{\log\log n_{obs}} - \mathcal{O}(1)\Big) \text{ a.s.}$$

Therefore, $\lim_{n_{obs}\to\infty} IC_Q = +\infty$ a.s. when $\lim_{n_{obs}\to\infty} \frac{C_n}{\log\log n_{obs}} = +\infty$ and

$$IC_Q(k) > IC_Q(q) \text{ for } n_{obs} \text{ large.}$$

Second case: If $\mathcal{M}_k$ is not included in $\mathcal{M}_q$, then $q > k$ and Assertion (c) of Theorem 1 implies that

$$\frac{1}{n_{obs}}\lambda_{n_{obs}} = e(g, f, \theta_g(k)) - e(g, f, \theta_g(q)) + \mathcal{O}\Big(\sqrt{n_{obs}^{-1}\log\log n_{obs}}\Big) \text{ a.s.}$$
$$\underset{n_{obs}\to+\infty}{\longrightarrow} e(g, f, \theta_g(k)) - e(g, f, \theta_g(q)) < 0 \tag{4}$$

Thus by (4), Assumptions A5 and A6, and the condition $\lim_{n\to\infty} \frac{C_n}{n} = 0$, we get

$$\begin{aligned}\delta IC_Q &= -2n\Big[\frac{1}{n}\delta Q - \frac{C_n(k-q)}{2n}\Big] \\ &= -2n\Big[\frac{n_{obs}}{n}\frac{\lambda_{n_{obs}}}{n_{obs}} + \frac{\delta H}{n} - \frac{C_n(k-q)}{2n}\Big] \\ &\to +\infty \ a.s.\end{aligned}$$

as both n and n_{obs} tend to $+\infty$. Then $IC_Q(k) > IC_Q(q)$ and it is $\mathcal{M}_q$ that minimizes the criterion.

Combining the two cases, we conclude that there exists some N, such that for all $n > n_{obs} \geq N$, if $\frac{C_n}{n} \to 0$ and $\frac{C_n}{\log\log n_{obs}} \to +\infty$, then the model $\hat{\mathcal{M}}$ that minimizes IC_Q criterion tends to the quasi true model $\mathcal{M}_q$ almost surely.

By relaxing the condition $\frac{C_n}{\log\log n_{obs}} \to +\infty$, we get the following result.

Corollary 1 *Let $\hat{\mathcal{M}}$ be the model that minimizes the IC_Q criterion among the candidate models $\mathcal{M}_k$. Suppose that Assumption A6 in Theorem 2 holds.*

If $\lim_{n\to\infty} \frac{C_n}{n} = 0$ and $\lim_{n\to\infty} C_n = +\infty$, then $P\big[\lim_{n_{obs}\to\infty} \hat{\mathcal{M}} = \mathcal{M}_q\big] = 1$.

Since the penalization in IC_{cd} criteria is greater than that in IC_Q, Theorem 2 and Corollary 1 hold for model selection with IC_{cd}. The BIC_{cd} and BIC_Q criteria, with $C_n = \log n$, are strongly consistent. The penalty factor $C_n = 2$ for AIC_Q and AIC_{cd} does not meet the consistency conditions.

4. Numerical studies

4.1 Autoregressive model with missing data

In this section, we report on simulation studies conducted with the statistical software R to illustrate the consistency of incomplete data criteria. We generated 100 samples of size $n = 100,\ 200,\ \text{and}\ 500$ from the model:

$$AR(2): \qquad X_t = 1.4X_{t-1} - .49X_{t-2} + \varepsilon_t \text{ where } \{\varepsilon_t\} \sim \mathcal{N}(0, 1)$$

Table 2. Frequencies of estimated orders over 100 replications from the $AR(2)$ model with various sample sizes.

		Orders											
		$n = 100$			$n = 200$			$n = 500$			$n = 1000$		
P_{mis}	Criteria	< 2	2	> 2	< 2	2	> 2	< 2	2	> 2	< 2	2	> 2
10%	BIC_Q	2	88	10	0	95	5	0	96	4	0	99	1
	BIC_{cd}	2	89	9	0	95	5	0	97	3	0	99	1
20%	BIC_Q	9	73	18	0	86	14	0	91	9	0	95	5
	BIC_{cd}	11	77	12	0	92	8	1	94	5	0	97	3
30%	BIC_Q	19	65	16	11	77	12	1	87	12	0	91	9
	BIC_{cd}	24	70	6	12	85	3	1	90	9	0	94	6
40%	BIC_Q	20	54	26	9	70	21	4	80	16	1	88	11
	BIC_{cd}	26	63	11	12	79	9	6	85	9	1	91	8

and we created incomplete samples by discarding at random a percentage P_{mis} of data. To investigate the effect of the amount of missingness, we considered $P_{mis} = 10\%,\ 20\%,\ 30\%,$ and 40%. Then candidate autoregressive models $AR(k), k = 1, \ldots, 10$ were fitted to each sample using the *EM* algorithm and the order was estimated with the BIC_Q and BIC_{cd} criteria. In the *EM* algorithm, the parameters estimates obtained from the largest block of observed data serve as initial values. For autoregressive models, the term $(k + Tr[DF(I - DF)^{-1}])$ in the penalization of BIC_{cd} can have the simpler form kn/n_{obs} (El Matouat et al., in press) that we considered in the numerical tests. In Table 2, the distribution of the selected orders by each criterion over the 100 samples is presented. We group the order selection into three categories: (< 2) – underfitting, (2) – correct order identification, and (> 2) – overfitting.

As the amount of missing data in the sample grows, the frequencies of the correct order selection by BIC_Q and BIC_{cd} criteria decrease. Yet when the size of the sample increases, the performances of the criteria at identifying the true order are improved even for large fractions of missingness. For $n = 100$, BIC_{cd} is more prone to underfitting than BIC_Q particularly as the fractions of missing data are increased ($P_{mis} = 20\%,\ 30\%$, and 40%); but this tendency is lessened for larger samples. When P_{mis} is increased, BIC_Q becomes more prone toward selecting higher order models, compared to BIC_{cd}. In addition, we note that BIC_{cd} outperforms BIC_Q particularly for large fractions of missing data.

4.2 Linear regression with missing covariates

We investigated the performance of the criteria at correctly identifying a linear regression model with missing at random covariate. Following Ibrahim et al. (2008), we simulated data sets consisting of 500 samples of size $n = 100,\ 300$, and 500 from the model:

$$y_i|x_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2) \text{ where } x_i \sim \mathcal{N}(\mu, \tau^2)$$

and $\beta_0 = \beta_1 = \sigma^2 = \mu = \tau^2 = 0.8$. We also generated independent observations from the standard normal distribution $z_1, \ldots, z_n$ that are independent from $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$. For the incomplete data setting, y_i and z_i, $i = 1, \ldots, n$, are supposed to be completely observed while some x_i are omitted according to the missing data mechanism defined as follows:

$$r_i = \begin{cases} 0 & \text{if } x_i \text{ is observed} \\ 1 & \text{if } x_i \text{ is missing} \end{cases}$$

Table 3. Distribution of model selection over 500 replications from the linear regression model with various sample sizes.

		$n = 100$		$n = 300$		$n = 500$	
	Models	BIC_Q	BIC_{cd}	BIC_Q	BIC_{cd}	BIC_Q	BIC_{cd}
P_{mis1}	M1	34	42	4	4	1	1
	M2	435	432	475	479	486	491
	M3	16	13	15	7	6	4
	M4	14	13	6	10	7	4
	M5	1	0	0	0	0	0
P_{mis2}	M1	31	56	3	8	0	0
	M2	423	417	468	476	477	483
	M3	23	16	15	9	9	5
	M4	21	11	12	5	13	11
	M5	2	0	2	2	1	1

$$\text{with } p(r_i = 1|y_i, z_i) = \frac{exp(\phi_0 + \phi_1 y_i)}{1 + exp(\phi_0 + \phi_1 y_i)}$$

We consider $\phi_0 = -4,\ \phi_1 = 1$ yielding a missing data fraction of about 11% for the first mechanism P_{mis1}, and $\phi_0 = -3.5,\ \phi_1 = 1.5$ giving about 29% of missing data in each sample for the second mechanism P_{mis2}. Then, for each incomplete sample, we select one model with BIC_Q and BIC_{cd} criteria among the five candidate models:

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

The stopping criterion used for the EM algorithm was that the difference between two successive iterates was less than 10^{-6}. The results obtained for the distribution of selections by the criteria are presented in Table 3.

BIC_Q and BIC_{cd} criteria perform well at identifying the true model under the two missing data mechanisms and for all the sample sizes considered. For $n = 100$ and P_{mis1}, model M2 is chosen 435 times by BIC_Q and 432 times by BIC_{cd} out of 500. With the samples of sizes $n = 300$ and $n = 500$ under P_{mis1}, the number of times the true model is picked by both criteria is increased. Under the missing data mechanism P_{mis2}, model M2 is selected 423 times by BIC_Q and 417 times by BIC_{cd} for $n = 100$. For larger samples, the number of times M2 is chosen by both criteria is slightly decreased when compared to the results obtained under P_{mis1}, hence the performances of BIC_Q and BIC_{cd} seem to be affected by the fraction of missingness. But the selection of the true model under P_{mis2} is still improved for larger samples. These results imply that the performance of BIC_Q and BIC_{cd} at selecting the true model is improved as the size of the sample grows. Both criteria yield similar results under P_{mis1} (about 11% of missingness), but BIC_{cd} outperforms BIC_Q for larger missing data fraction (about 29% under P_{mis2}).

4.3 Sunspots data

We also propose an application of the criteria to the monthly sunspots numbers 1749–1983 data set available in R. For $n = 500,\ 600,\ 800$, and 1000, we considered the first n sunspots data and we suppose an autoregressive modeling for the data. To serve as a reference, we compute the order estimate using BIC criterion among candidate autoregressive models $AR(k)$,

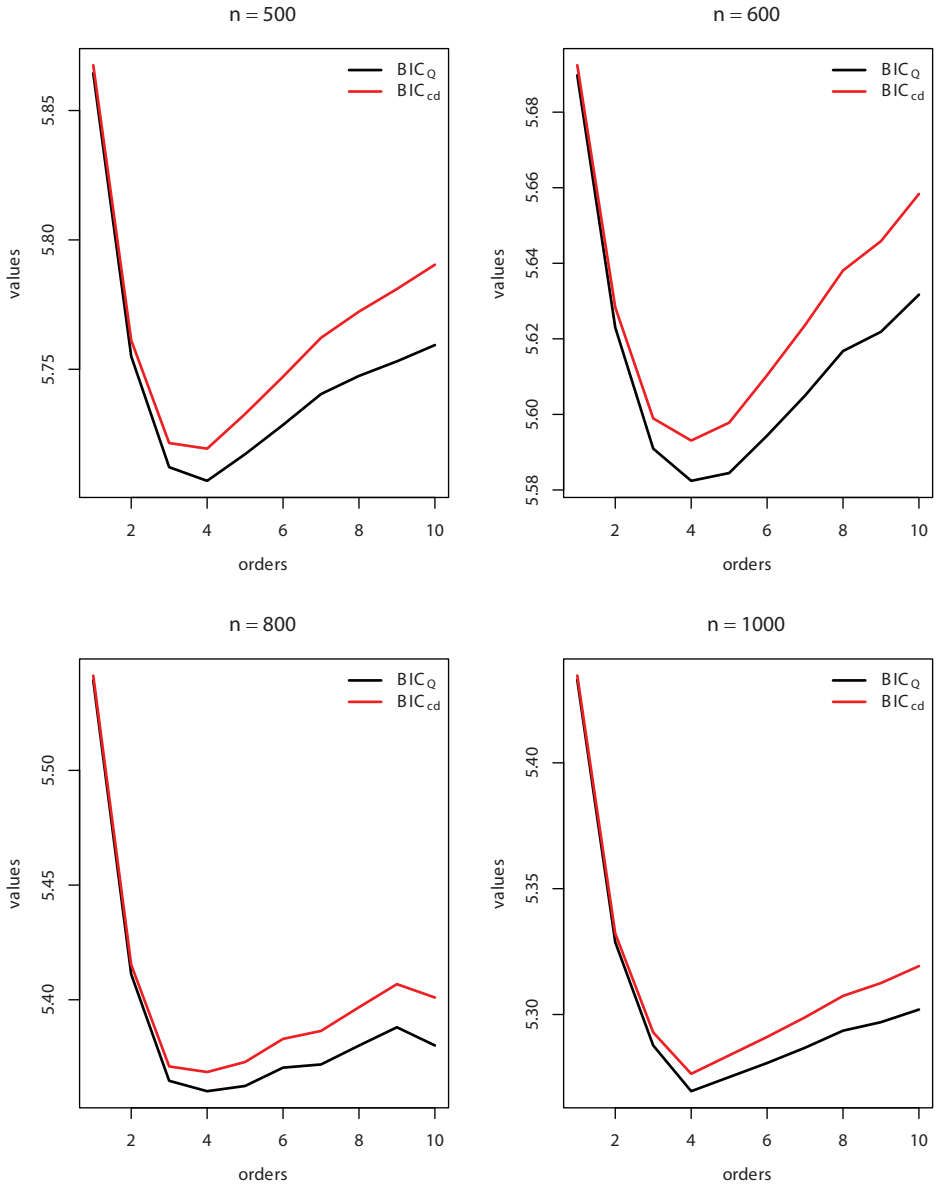

Figure 1. Criteria values from the incomplete samples with 20% of missingness.

$k = 1, \ldots, 10$, and the $AR(4)$ model was selected from all the four samples. Next, we omitted at random some fractions $P_{mis} = 20\%$ and 40% of data to create incomplete samples and we calculated the values of BIC_Q and BIC_{cd}, when fitting the same candidate autoregressive models to the data. Figures 1 and 2 present the plots of the values of these criteria as functions of the order of the candidate model.

When 20% of data are missing in the samples, the $AR(4)$ model is still picked by BIC_Q and BIC_{cd} but the minima of the criteria curves are better defined for $n = 1000$. When the fraction of missingness rises to 40%, underfitting occurs for the samples of size $n = 500$ and $n = 600$ (the $AR(3)$ model is selected). For larger samples ($n = 800,\ 1000$), the order 4, which is the order estimate obtained with the complete samples, is selected by both criteria.

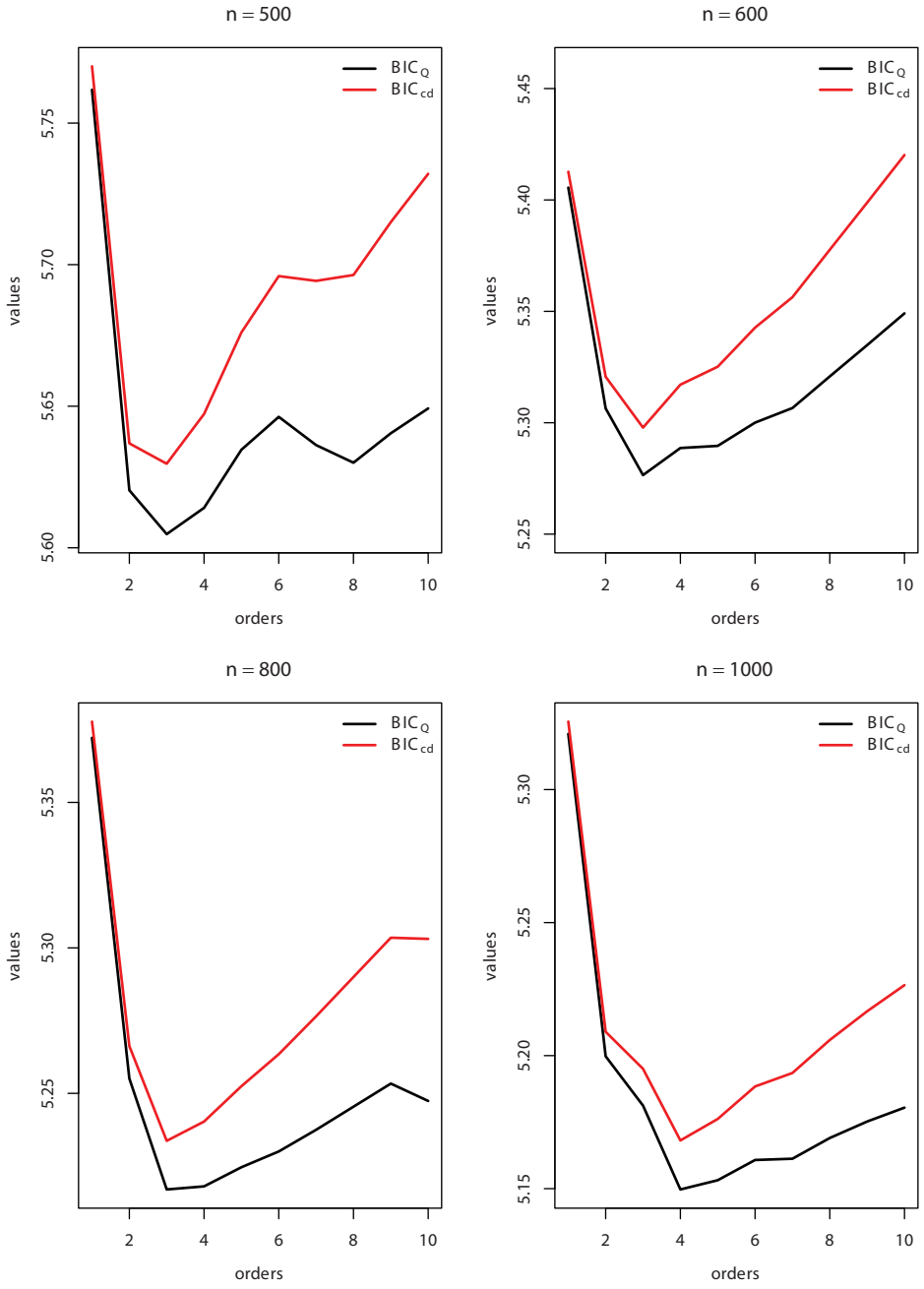

Figure 2. Criteria values from the incomplete samples with 40% of missingness.

4.4 Quakes data

We also apply the criteria to the quakes data set that is available in the data set package of the R statistical software. It consists of observations on 1000 earthquakes that occurred near Fiji, starting in 1964. In the data set, the latitude (lat), the longitude (long), the depth (depth), the Richter magnitude (mag) of each event are included, plus the number of stations that reported the quake. Although these data have no missing covariate, we conducted tests to compare the performance of the criteria based on hypothetical missing data with complete data analysis.

Assuming a linear regression model for the data, the outcome variable is the Richter magnitude (mag) and the predictors variables are the latitude (lat), the longitude (long), and the depth (depth). We consider the following candidate models:

$$M1: \ mag_i = \beta_0 + \beta_1 lat_i + \beta_2 long_i + \varepsilon_i, \ \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$
$$M2: \ mag_i = \beta_0 + \beta_1 lat_i + \beta_2 depth_i + \varepsilon_i, \ \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$
$$M3: \ mag_i = \beta_0 + \beta_1 lat_i + \beta_2 long_i + \beta_3 depth_i + \varepsilon_i, \ \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

Using the *BIC* criterion, the model *M*3 is selected from the complete data and the fitted model is

$$\hat{mag}_i = 6.75327 - 8.939.10^{-3} lat_i - 0.012263 long_i - 3.7464.10^{-4} depth_i + \varepsilon_i$$

with $\hat{\sigma}^2 = 0.148623$.

Next, we generated missing data in the covariate lat with the missing data mechanism

$$r_i = \begin{cases} 0 & \text{if } lat_i \text{ is observed} \\ 1 & \text{if } lat_i \text{ is missing} \end{cases}$$

$$\text{with } p(r_i = 1 | mag_i) = \frac{exp(-0.8 + 1/mag_i)}{1 + exp(-0.8 + 1/mag_i)}$$

giving a fraction of missing data of about 35.5%. We consider the normal distribution $\mathcal{N}(\mu, \tau^2)$ for lat, and we calculated BIC_Q and BIC_{cd} for each candidate model after estimating the parameter with the *EM* algorithm. The same stopping criterion considered for linear regression simulation is used for the algorithm. Model M3 is still picked by both criteria and the estimated model is given by

$$\hat{mag}_i = 6.74317 - 7.7269810^{-3} lat_i - 0.012072 long_i - 3.7443910^{-4} depth_i + \varepsilon_i$$

with $\hat{\sigma}^2 = 0.148828$.

5. Conclusion

We studied the consistency of the *EM* algorithm-based criteria IC_{cd} and IC_Q for model selection with incomplete data. We derived asymptotic results related to maximum likelihood estimation with missing data. This allowed us to provide sufficient conditions for consistent model selection using penalized conditional expected likelihood criteria, and we proved the consistency of BIC_{cd} and BIC_Q. The consistency results require that the fraction of observed data n_{obs}/n does not vanish as both n and n_{obs} grow to infinity, which is a quite legitimate assumption. We also provided numerical results that illustrate the behavior and the consistency of the incomplete data criteria. When there is no missing data, i.e. $n_{obs} = n$, the consistency conditions given above are exactly the conditions by Nishii (1988). Hence, we have extended Nishii's results to the incomplete data case.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In: Petrov, B.N., Csaki, F., eds. *Second International Symposium of Information Entropy Theory* (pp. 267–281). Budapest: Akademia Kiado.

Bozdogan, H. (1988). ICOMP: a new model selection criterion. In: Bock, H.H., ed. *Classification and Related Methods of Data Analysis* (pp. 599–608). Amsterdam: Elsevier.

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

Dempster, A.P., Laird, N.M., Rubin, D. (1977). Maximum likelihood from incomplete data via EM algorithm. *J. R. Stat. Soc. Ser. B* 39(1):1–38.

El Matouat, A., Djibril Moussa, F., Hamzaoui, H. (2015). Resampling for order estimation of autoregressive models with missing data. *Commun. Stat.-Simul. Comput.* 44:1187–1196.

Hannan, E.J., Quinn, B.G. (1979). The determination of the order of an autoregression. *J. Roy. Stat. Soc. B* 41:190–195.

Hurvich, C.M., Tsai, C. (1989). Regression and time series model selection in small samples. *Biometrika* 76(2):297–307.

Ibrahim, J.G., Zhu, H., Tang, N. (2008). Model selection criteria for missing-data problems using the EM algorithm. *J. Am. Stat. Assoc.* 103(484):1648–1658.

McLachlan, G.J., Krishnan, T. (1997). *The EM Algorithm and Extensions*. New York: A Wiley-Interscience.

Meng, X.L., Rubin, D.B. (1991). Using EM algorithm to obtain asymptotic variance-covariance matrices: the SEM algorithm. *J. Am. Stat. Assoc.* 86:899–909.

Nishii, R. (1988). Maximum likelihood principle and model selection when the true model is unspecified. *J. Multivar. Anal.* 27:392–403.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Stat.* 6:461–464.

Shimodaira, H. (1994). A new criterion for selecting models from partially observed data. In: Cheeseman, P., Oldford, R.W., eds. *Selecting Models from Data: Artificial Intelligence and Statistics IV. Lecture Notes in Statistics*. (Vol. 89, pp. 21–29). New York: Springer.

Wu, C.F. Jeff. (1983). On the convergence properties of the EM algorithm. *Ann. Stat.* 11(1):95–103.