An Akaike Criterion based on Kullback Symmetric Divergence in the Presence of Incomplete-Data

Bezza Hafidi[a] and Abdallah Mkhadri[a][*]

[a] University Cadi-Ayyad, Faculty of sciences Semlalia, Department of Mathematics, PB.2390
Marrakech, Moroco

Abstract

This paper investigates and evaluates an extension of the Akaike information criterion, KIC, which is an approximately unbiased estimator for a risk function based on the Kullback symmetric divergence. KIC is based on the observed-data empirical log-likelihood which may be problematic to compute in the presence of incomplete-data. We derive and investigate a variant of KIC criterion for model selection in settings where the observed-data is incomplete. We examine the performance of our criterion relative to other well known criteria in a large simulation study based on bivariate normal model and bivariate regression modeling.

2000 AMS Subject Classification: 62B10, 94A17, 62E25.
Key-words: Akaike information criterion, Kullback's symmetric divergence, Missing-data, Model selection, EM algorithm, Regression.

1 Introduction

The Akaike information criterion, AIC (Akaike 1973), was designed as an asymptotically unbiased estimator of a variant of Kullback's (1968) directed divergence between the true model and a fitted approximating model. The directed divergence is an asymmetric measure, meaning that an alternative directed divergence may be obtained by reversing the role of the two models in the definition of the measure. The sum of the two directed

[*] The second author is supported by TWAS Research grant 01-159 RG/MATHS/AF/AC.

Received on November 20, 2005. Revised version accepted on February 10, 2007

measures combines the information in both measures. Therefore, it functions as a gauge of model disparity which is arguably more balanced than either of its individual components. In the framework of linear models, Cavanaugh (2004) showed that directed divergence is more sensitive in detecting overfitted models, whereas its counterpart is more sensitive towards detecting underfitted models. Cavanaugh (1999) proposed an Akaike criterion, called KIC (Kullback Information Criterion), as an asymptotically unbiased estimator of Kullback's (1968) symmetric divergence; the sum of the two directed divergences.

However, all these criteria are based on the observed-data empirical log-likelihood, which may be problematic to compute in a large variety of practical problems in presence of missing-data (cf. Cavanaugh & Shumway 1998). As pointed out by Shimodaira (1994), in many applications it may be more natural or desirable to use a criterion based on the complete-data, which assesses the separation between the fitted model and the true model. Cavanaugh & Shumway (1998) provided some arguments to be made in defense of this idea, and proposed a variant of AIC, denoted AICcd, as an approximately unbiased estimator of the asymmetric measure based on complete-data. Another variant, denoted PDIO, was proposed by Shimodaira (1994) which is based on the same principle as AICcd but differs from it in its goodness of fit term.

In the present paper, we propose a novel criterion , which is a variant of KIC and denoted KICcd, for model selection in the presence of incomplete-data. Like AICcd, our criterion KICcd is based on complete-data rather than incomplete-data concepts, but differs from AICcd in its penalty term.

In section 2, we present an overview of criteria based on Kullback's asymmetric and symmetric divergence, AIC and KIC. We derive our criterion KICcd in Section 3. In Section 4, we investigate the performance of our criterion and compare it to AIC, KIC and AICcd in a large simulation study involving modeling bivariate normal data and bivariate regression modeling. We end the paper with a brief discussion in Section 5. *All the tables and the figures can be found at the end of paper in the last appendix.*

2 AIC and KIC criteria

Let $\mathbf{Y}_o$ be the vector of the observed-data or incomplete-data. Let θ_0 be the true parameter vector which is unknown and θ the parameter vector of the candidate model, so that $f(\mathbf{Y}_o|\theta_0)$ and $f(\mathbf{Y}_o|\theta)$ represent the generating and the candidate parametric densities for the incomplete-data $\mathbf{Y}_o$, respectively. We denote by $\hat{\theta}$ the maximum likelihood estimator

(MLE) of θ_0, and by d the dimension of the vector θ. A measure of separation between a candidate model $f(\mathbf{Y}_o|\theta)$ and the generating model $f(\mathbf{Y}_o|\theta_0)$ is defined by (Kullback 1968)

$$I_{\mathbf{Y}_o}(\theta_0,\theta) = D_{\mathbf{Y}_o}(\theta_0,\theta) - D_{\mathbf{Y}_o}(\theta_0,\theta_0), \tag{1}$$

where $D_{\mathbf{Y}_o}(\theta_0,\theta) = \mathbf{E}_{\mathbf{Y}_o|\theta_0}\{-2\ln f(\mathbf{Y}_o|\theta)\}$ and $\mathbf{E}_{\mathbf{Y}_o|\theta_0}$ denotes the expectation under $f(\mathbf{Y}_o|\theta_0)$. The second term of (1) can be discarded, since it does not depend on θ. Therefore, $D_{\mathbf{Y}_o}(\theta_0,\theta)$ which is called a discrepancy, provides a suitable measure of the separation between the two models. Since θ_0 is unknown, the evaluation of this quantity is not possible. Akaike (1973,1974) showed that the criterion

$$AIC = -2\ln L(\mathbf{Y}_o|\hat{\theta}) + 2d, \tag{2}$$

is an asymptotically unbiased estimator of $\Delta_{\mathbf{Y}_o|\theta_0}(d,\theta_0) = \mathbf{E}_{\mathbf{Y}_o|\theta_0}\Big\{D_{\mathbf{Y}_o}(\theta_0,\hat{\theta})\Big\}$, where $L(\mathbf{Y}_o|\hat{\theta})$ represents the empirical likelihood for the incomplete-data and

$$D_{\mathbf{Y}_o}(\theta_0,\hat{\theta}) = \mathbf{E}_{\mathbf{Y}_o|_{\theta_0}}\Big\{-2\ln f(\mathbf{Y}_o|\theta)\Big\}|_{\theta=\hat{\theta}}, \tag{3}$$

is an asymmetric measure of separation between two statistical models.

An alternative directed divergence is Kullback's symmetric divergence, defined as the sum of two directed divergences (Kullback 1968), i.e.

$$J_{\mathbf{Y}_o}(\theta_0,\theta) = I_{\mathbf{Y}_o}(\theta_0,\theta) + I_{\mathbf{Y}_o}(\theta,\theta_0). \tag{4}$$

Note that $J_{\mathbf{Y}_o}(\theta_0,\theta) = J_{\mathbf{Y}_o}(\theta,\theta_0)$, whereas $I_{\mathbf{Y}_o}(\theta_0,\theta) \neq I_{\mathbf{Y}_o}(\theta,\theta_0)$ unless $\theta = \theta_0$. It is well known that $J_{\mathbf{Y}_o}(\theta_0,\theta) \geq 0$ with equality if and only if $\theta = \theta_0$.

Cavanaugh (1999) showed that the criterion defined by

$$KIC = -2\ln L(\mathbf{Y}_o|\hat{\theta}) + 3d, \tag{5}$$

is an asymptotically unbiased estimator of $\Omega_{\mathbf{Y}_o}(d,\theta_0) = \mathbf{E}_{\mathbf{Y}_o|\theta_0}\Big\{K_{\mathbf{Y}_o}(\theta_0,\hat{\theta})\Big\}$, where

$$K_{\mathbf{Y}_o}(\theta_0,\hat{\theta}) = D_{\mathbf{Y}_o}(\theta_0,\hat{\theta}) + \Big\{D_{\mathbf{Y}_o}(\hat{\theta},\theta_0) - D_{\mathbf{Y}_o}(\hat{\theta},\hat{\theta})\Big\}, \tag{6}$$

$D_{\mathbf{Y}_o}(\hat{\theta},\theta_0) = \mathbf{E}_{\mathbf{Y}_o|\theta}\{-2\ln f(\mathbf{Y}_o|_{\theta_0})\}|_{\theta=\hat{\theta}}$ and $D_{\mathbf{Y}_o}(\hat{\theta},\hat{\theta}) = \mathbf{E}_{\mathbf{Y}_o|\theta}\{-2\ln f(\mathbf{Y}_o|_{\theta})\}|_{\theta=\hat{\theta}}$.

Cavanaugh (1999) showed that the criterion KIC outperforms AIC. He suggested that $J_{\mathbf{Y}_o}(\theta_0,\theta)$ is preferable to $I_{\mathbf{Y}_o}(\theta_0,\theta)$ as a directed divergence tool for model selection.

3 Derivation of KICcd

We assume that the vector of complete-data has the following form $\mathbf{Y} = (\mathbf{Y}_o, \mathbf{Y}_m)$, where $\mathbf{Y}_o$ is the vector of the observed-data and $\mathbf{Y}_m$ is the vector of the missing-data. Let $f(\mathbf{Y}|\theta_0)$ and $f(\mathbf{Y}|\theta)$ be the generating and the candidate parametric densities for the complete-data $\mathbf{Y}$, respectively. Following the derivation of KIC, we assume that the candidate class of models includes the true model (Linhart and Zucchini, 1986). In this setting, the complete-data density $f(\mathbf{Y}|\theta)$ can be written as

$$f(\mathbf{Y}|\theta) = f(\mathbf{Y}_o|\theta) f(\mathbf{Y}_m|\mathbf{Y}_o, \theta). \tag{7}$$

So, If the density $f(\mathbf{Y}_m|\mathbf{Y}_o, \theta)$ is substantially affected by deviations of θ from the true parameter θ_0, the model selection based on the discrepancy between the fitting model $f(\mathbf{Y}|\hat{\theta})$ and the generating model $f(\mathbf{Y}|\theta_0)$ would incorporate this information. Therefore, it is not clear that model selection based on the discrepancy of incomplete-data would be the same. In this section, we explore this argument and describe the evaluation of KICcd.

3.1 Complete-data discrepancy for the symmetric divergence

The complete-data discrepancy for the symmetric divergence between a fitting model $f(\mathbf{Y}|\hat{\theta})$ and the generating model $f(\mathbf{Y}|\theta_0)$ is defined by

$$J_{\mathbf{Y}}(\theta_0, \hat{\theta}) \;\; = \;\; I_{\mathbf{Y}}(\theta_0, \hat{\theta}) + I_{\mathbf{Y}}(\hat{\theta}, \theta_0), \tag{8}$$

where $I_{\mathbf{Y}}(\theta_0, \hat{\theta})$ and $I_{\mathbf{Y}}(\hat{\theta}, \theta_0)$ are defined as before when we replace $\mathbf{Y}_o$ by $\mathbf{Y}$.
Although the two measures $I_{\mathbf{Y}}(\theta_0, \hat{\theta})$ and $I_{\mathbf{Y}}(\hat{\theta}, \theta_0)$ judged the dissimilarity between $f(\mathbf{Y}|\theta_0)$ and $f(\mathbf{Y}|\hat{\theta})$, they are not redundant. The measure $I_{\mathbf{Y}}(\theta_0, \hat{\theta})$ evaluates how well the fitted candidate model performs on average on new samples generated under the true model. Whereas, the measure $I_{\mathbf{Y}}(\hat{\theta}, \theta_0)$ evaluates how well the true model conforms on average on new samples generated under the fitted candidate model. Since the symmetric divergence, $J_{\mathbf{Y}}(\theta_0, \hat{\theta})$, reflects the sensitivities of both directed divergences, it may serve as a more balanced discrepancy measure than either of its individual components. Even if this argument is not formal, numerical illustration supporting this fact is provided in Figure 1 and Tables 2 and 4 in the last appendix.
In the presence of missing-data, the following lemma justifies the use of the complete-data discrepancy $K_{\mathbf{Y}}(\theta_0, \hat{\theta})$ defined as in (6) where we replace $\mathbf{Y}_o$ by $\mathbf{Y}$.

Lemma 3.1 *We have*

$$K_{\mathbf{Y}}(\theta_0,\theta) \geq K_{\mathbf{Y}_o}(\theta_0,\theta)+k(\theta_0),$$

where $k(\theta_0)=\mathbf{E}_{\mathbf{Y}_o|\theta_0}\Big\{D_{\mathbf{Y}_m|\mathbf{Y}_o}(\theta_0,\theta_0)\Big\}$ *is independent of* θ.

Proof. Using equation (7), we have

$$D_{\mathbf{Y}}(\theta,\theta_0) \;=\; D_{\mathbf{Y}_o}(\theta,\theta_0)+\mathbf{E}_{\mathbf{Y}|\theta}\Big\{-2\ln f(\mathbf{Y}_m|\mathbf{Y}_o,\theta_0)\Big\}. \tag{9}$$

Similarly, we have

$$D_{\mathbf{Y}}(\theta,\theta) \;=\; D_{\mathbf{Y}_o}(\theta,\theta)+\mathbf{E}_{\mathbf{Y}|\theta}\Big\{-2\ln f(\mathbf{Y}_m|\mathbf{Y}_o,\theta)\Big\}. \tag{10}$$

Let

$$D_{\mathbf{Y}_m|\mathbf{Y}_o}(\theta,\theta_0)=\mathbf{E}_{(\mathbf{Y}_m|\mathbf{Y}_o,\theta)}\Big\{-2\ln f(\mathbf{Y}_m|\mathbf{Y}_o,\theta_0)\Big\}$$

and

$$D_{\mathbf{Y}_m|\mathbf{Y}_o}(\theta,\theta)=\mathbf{E}_{(\mathbf{Y}_m|\mathbf{Y}_o,\theta)}\Big\{-2\ln f(\mathbf{Y}_m|\mathbf{Y}_o,\theta)\Big\},$$

where $\mathbf{E}_{(\mathbf{Y}_m|\mathbf{Y}_o,\theta)}$ denotes the expectation under the density $f(\mathbf{Y}_m|\mathbf{Y}_o,\theta)$.
Again from (7), it can be shown that $\mathbf{E}_{\mathbf{Y}|\theta}\Big\{-2\ln f(\mathbf{Y}_m|\mathbf{Y}_o,\theta_0)\Big\}=\mathbf{E}_{\mathbf{Y}_o|\theta}\Big\{D_{\mathbf{Y}_m|\mathbf{Y}_o}(\theta,\theta_0)\Big\}$, and $\mathbf{E}_{\mathbf{Y}|\theta}\Big\{-2\ln f(\mathbf{Y}_m|\mathbf{Y}_o,\theta)\Big\}=\mathbf{E}_{\mathbf{Y}_o|\theta}\Big\{D_{\mathbf{Y}_m|\mathbf{Y}_o}(\theta,\theta)\Big\}$.
Now, substituting (9) and (10) in $K_{\mathbf{Y}}(\theta_0,\hat{\theta})$ with $\hat{\theta}=\theta$, we obtain

$$\begin{aligned}K_{\mathbf{Y}}(\theta_0,\theta) \;&=\; D_{\mathbf{Y}}(\theta_0,\theta)+D_{\mathbf{Y}_o}(\theta,\theta_0)-D_{\mathbf{Y}_o}(\theta,\theta)\\ &\quad+\Big\{\mathbf{E}_{\mathbf{Y}_o|\theta}\Big\{D_{\mathbf{Y}_m|\mathbf{Y}_o}(\theta,\theta_0)\Big\}-\mathbf{E}_{\mathbf{Y}_o|\theta}\Big\{D_{\mathbf{Y}_m|\mathbf{Y}_o}(\theta,\theta)\Big\}\Big\}.\end{aligned}$$

Furthermore, Cavanaugh and Shumway (1998) showed that

$$D_{\mathbf{Y}}(\theta_0,\theta)\geq D_{\mathbf{Y}_o}(\theta_0,\theta)+k(\theta_0),$$

where $k(\theta_0)=\mathbf{E}_{\mathbf{Y}_o|\theta_0}\Big\{D_{\mathbf{Y}_m|\mathbf{Y}_o}(\theta_0,\theta_0)\Big\}$ is independent of θ. This leads to

$$\begin{aligned}K_{\mathbf{Y}}(\theta_0,\theta) \;&\geq\; D_{\mathbf{Y}_o}(\theta_0,\theta)+k(\theta_0)+D_{\mathbf{Y}_o}(\theta,\theta_0)-D_{\mathbf{Y}_o}(\theta,\theta)\\ &\quad+\;\mathbf{E}_{\mathbf{Y}_o|\theta}\Big\{D_{\mathbf{Y}_m|\mathbf{Y}_o}(\theta,\theta_0)-D_{\mathbf{Y}_m|\mathbf{Y}_o}(\theta,\theta)\Big\}\\ &\geq\; K_{\mathbf{Y}_o}(\theta_0,\theta)+k(\theta_0)+\mathbf{E}_{\mathbf{Y}_o|\theta}\Big\{D_{\mathbf{Y}_m|\mathbf{Y}_o}(\theta,\theta_0)-D_{\mathbf{Y}_m|\mathbf{Y}_o}(\theta,\theta)\Big\}.\end{aligned}$$

Now, let $I_{\mathbf{Y}_m|\mathbf{Y}_o}(\theta,\theta_0)=D_{\mathbf{Y}_m|\mathbf{Y}_o}(\theta,\theta_0)-D_{\mathbf{Y}_m|\mathbf{Y}_o}(\theta,\theta)$. Since $I_{\mathbf{Y}_m|\mathbf{Y}_o}(\theta,\theta_0)\geq 0$ with equality if and only if $\theta=\theta_0$, then we have for any θ

$$K_{\mathbf{Y}}(\theta_0,\theta)\geq K_{\mathbf{Y}_o}(\theta_0,\theta)+k(\theta_0),$$

which ends the proof. □

From the inequality in lemma 3.1, we can conclude that the complete-data discrepancy $K_{\mathbf{Y}}(\theta_0, \theta)$ is always at least as great as the incomplete-data discrepancy $K_{\mathbf{Y}_o}(\theta_0, \theta)$. This implies that in the presence of missing-data, an estimator of the expected complete-data discrepancy may be preferable to an estimator of the expected incomplete-data discrepancy as a model selection criterion. Numerical illustration supporting this fact is provided in Figures 2 and 3 in the last appendix.

3.2 Derivation of KICcd

Our objective is to propose a version of KIC that will have an approximately unbiased estimator of the expected complete-data discrepancy $\Omega_{\mathbf{Y}}(d, \theta_0) = \mathbf{E}_{\mathbf{Y}_o|\theta_0}\Big\{K_{\mathbf{Y}}(\theta_0, \hat{\theta})\Big\}$. We assume that θ_0 is an interior point of the parameter space for the candidate model, and that the usual regularity conditions needed to ensure the consistency and asymptotic normality of $\hat{\theta}$ are satisfied. Let

$$\begin{array}{rcl} \mathbf{I}_o(\theta|\mathbf{Y}_o) & = & \dfrac{-\partial^2 \ln f(\mathbf{Y}_o|\theta)}{\partial\theta\partial\theta^t}, \mathbf{I}_o(\theta|\mathbf{Y}) = \dfrac{-\partial^2 \ln f(\mathbf{Y}|\theta)}{\partial\theta\partial\theta^t} \\ Q(\theta_1|\theta_2) & = & \displaystyle\int_{\mathbf{Y}_m} \{\ln f(\mathbf{Y}|\theta_1)\} f(\mathbf{Y}_m|\mathbf{Y}_o, \theta_2)\mathbf{dY}_m \quad \text{and} \\ \mathbf{I}_{oc}(\theta|\mathbf{Y}_o) & = & \mathbf{E}_{(\mathbf{Y}_m|\mathbf{Y}_o,\theta_0)}\left\{\dfrac{-\partial^2 \ln f(\mathbf{Y}|\theta)}{\partial\theta\partial\theta^t}\right\}. \end{array}$$

Cavanaugh and Shumway (1998) established that the criterion defined by

$$AICcd = -2Q(\hat{\theta}, \hat{\theta}) + 2\text{trace}\{\mathbf{I}_{oc}(\hat{\theta}|\mathbf{Y}_o)\mathbf{I}_o^{-1}(\hat{\theta}|\mathbf{Y}_o)\} \tag{11}$$

is an asymptotically unbiased estimator of $\mathbf{E}_{\mathbf{Y}_o|\theta_0}\Big\{D_{\mathbf{Y}}(\theta_0, \hat{\theta})\Big\}$.

The following proposition is an adaptation of this result for Kullback's symmetric divergence in the presence of missing-data.

Proposition 3.2 *In the presence of missing-data, an asymptotically unbiased estimator of* $\Omega_{\mathbf{Y}}(d, \theta_0)$ *is given by*

$$KICcd = -2Q(\hat{\theta}, \hat{\theta}) + 3\text{trace}\{\mathbf{I}_{oc}(\hat{\theta}|\mathbf{Y}_o)\mathbf{I}_o^{-1}(\hat{\theta}|\mathbf{Y}_o)\}. \tag{12}$$

Proof. From (6) when we replace $\mathbf{Y}_o$ by $\mathbf{Y}$, we have

$$\Omega_{\mathbf{Y}}(d, \theta_0) = \mathbf{E}_{\mathbf{Y}_o|\theta_0}\Big\{D_{\mathbf{Y}}(\theta_0, \hat{\theta})\Big\} + \mathbf{E}_{\mathbf{Y}_o|\theta_0}\Big\{D_{\mathbf{Y}}(\hat{\theta}, \theta_0) - D_{\mathbf{Y}}(\hat{\theta}, \hat{\theta})\Big\}. \tag{13}$$

Next, taking a second-order expansion in the second term of $D_{\mathbf{Y}}(\hat{\theta}, \theta_0)$ about $\hat{\theta}$, one can establish

$$D_{\mathbf{Y}}(\hat{\theta}, \theta_0) = D_{\mathbf{Y}}(\hat{\theta}, \hat{\theta}) + (\hat{\theta} - \theta_0)^t \mathbf{E}_{\mathbf{Y}_o|_{\hat{\theta}}} \Big\{ \mathbf{I}_{oc}(\hat{\theta}|\mathbf{Y}_o) \Big\} (\hat{\theta} - \theta_0) + r(\theta_0, \hat{\theta}), \tag{14}$$

where $r(\theta_0, \hat{\theta})$ is $O_p(1)$, $()^t$ stands for transpose. For n large, it is justifiable to replace $f(\mathbf{Y}_o|\hat{\theta})$ in the expression of $\mathbf{E}_{\mathbf{Y}_o|_{\hat{\theta}}} \Big\{ \mathbf{I}_{oc}(\hat{\theta}|\mathbf{Y}_o) \Big\}$ with $f(\mathbf{Y}_o|\theta_0)$ as in Cavanaugh and Shumway (1998). This leads to the large-sample approximation

$$\mathbf{E}_{\mathbf{Y}_o|_{\hat{\theta}}} \Big\{ \mathbf{I}_{oc}(\hat{\theta}|\mathbf{Y}_o) \Big\} \approx \mathbf{E}_{\mathbf{Y}_o|_{\theta_0}} \{ \mathbf{I}_{oc}(\hat{\theta}|\mathbf{Y}_o) \}. \tag{15}$$

Taking the expectation of (14) with respect to $f(\mathbf{Y}_o|\theta_0)$ and using (15) yields

$$\mathbf{E}_{\mathbf{Y}_o|_{\theta_0}} \Big\{ D_{\mathbf{Y}}(\hat{\theta}, \theta_0) - D_{\mathbf{Y}}(\hat{\theta}, \hat{\theta}) \Big\} \approx \mathbf{E}_{\mathbf{Y}_o|_{\theta_0}} \Big\{ (\hat{\theta} - \theta_0)^t \mathbf{E}_{\mathbf{Y}_o|_{\theta_0}} \{ \mathbf{I}_{oc}(\hat{\theta}|\mathbf{Y}_o) \} (\hat{\theta} - \theta_0) \Big\},$$

where $\approx$ stands for an asymptotic equality. Since the variance-covariance matrix of $(\hat{\theta} - \theta_0)$ is approximated by $\mathbf{I}_0^{-1}(\hat{\theta}|\mathbf{Y}_o)$ for the large sample, we have

$$\begin{aligned} & \mathbf{E}_{\mathbf{Y}_o|_{\theta_0}} \Big\{ (\hat{\theta} - \theta_0)^t \mathbf{E}_{\mathbf{Y}_o|_{\theta_0}} \{ \mathbf{I}_{oc}(\hat{\theta}|\mathbf{Y}_o) \} (\hat{\theta} - \theta_0) \Big\} \\ = \ & \text{trace} \Big\{ \mathbf{E}_{\mathbf{Y}_o|_{\theta_0}} \{ \mathbf{I}_{oc}(\hat{\theta}|\mathbf{Y}_o) \} \mathbf{E}_{\mathbf{Y}_o|_{\theta_0}} \{ (\hat{\theta} - \theta_0)(\hat{\theta} - \theta_0)^t \} \Big\} \\ \approx \ & \text{trace} \Big\{ \mathbf{E}_{\mathbf{Y}_o|_{\theta_0}} \{ \mathbf{I}_{oc}(\hat{\theta}|\mathbf{Y}_o) \} \mathbf{I}_0^{-1}(\hat{\theta}|\mathbf{Y}_o) \Big\}. \end{aligned} \tag{16}$$

An estimator for (16) is given by trace$\{ \mathbf{I}_{oc}(\hat{\theta}|\mathbf{Y}_o) \mathbf{I}_0^{-1}(\hat{\theta}|\mathbf{Y}_o) \}$. According to the result of (11) and substituting the latter estimator in the second term of (13), we obtain the estimator of $\Omega_{\mathbf{Y}}(d, \theta_0)$ given in the proposition. □

The penalty term of KICcd involves the information matrix $\mathbf{I}_0^{-1}(\hat{\theta}|\mathbf{Y}_o)$ which is often hard to provide. According to Meng and Rubin (1991), Cavanaugh and Shumway (1998) and Shimodaira (1994), the penalty term can be written as

$$3\text{trace}(\mathbf{I} - \mathbf{DM})^{-1} = 3d + 3\text{trace}\{ \mathbf{DM}(\mathbf{I} - \mathbf{DM})^{-1} \},$$

where the matrix $\mathbf{DM}$ is the operator of the EM algorithm, and $\mathbf{I}$ is the identity matrix. Using this expression, (12) can be expressed as

$$KICcd = -2Q(\hat{\theta}, \hat{\theta}) + \{ 3d + 3\text{trace}\{ \mathbf{DM}(\mathbf{I} - \mathbf{DM})^{-1} \} \}. \tag{17}$$

The first term in KICcd is easily evaluated from the EM algorithm. The matrix $\mathbf{DM}$ may be calculated by the SEM algorithm (supplemented EM, Meng and Rubin 1991) or by numerical differentiation methods such as the RDM and FDM algorithms proposed by Jamshidian and Jennrich (1999) (See Appendix for evaluation of the DM matrix and for definition of the EM and SEM algorithms).

To compare the behavior of the KICcd criterion, an evaluation of its expression can serve as a starting point. The expression (17) implies that the penalty term of KICcd is composed of the penalty term of KIC with a term which assesses an additional penalty in accordance with the impact of the missing-data on the fitted model. Since trace$\{\mathbf{DM}(\mathbf{I}-\mathbf{DM})^{-1}\}$ is positive, the penalty term of KICcd is always at least as large as the penalty term of KIC. Moreover, the goodness-of-fit term of KICcd, $-2Q(\theta_1|\theta_2)$, can be written as $-2H(\theta_1|\theta_2) - 2\ln L(\mathbf{Y}_o|\theta_1)$, where $H(\theta_1|\theta_2) = \mathbf{E}_{\mathbf{Y}_m|\mathbf{Y}_o,\theta_2}\{\ln f(\mathbf{Y}_m|\mathbf{Y}_o,\theta_1)\}$.

However, KICcd contains extra components in its goodness-of-fit term. The component $-2H(\hat{\theta}|\hat{\theta})$ provides a missing-data supplement to the goodness-of-fit term of KIC. It is nonetheless difficult to give a general characterization made by the sum of this extra components to KICcd.

On the other hand, our criterion KICcd is different from AICcd in its penalty term. The difference between these terms causes these criteria to behave quite differently, as indicated by the numerical simulations in the next section.

4 Numerical experiments

We carried out a fairly large simulation study of the performance of the KICcd criterion compared to KIC, AIC and AICcd criteria. This simulation study focuses on two important modeling frameworks: the bivariate normal model and bivariate regression models.

4.1 Bivariate normal

We consider the same example used by Cavanaugh and Shumway (1998) in their simulation study to investigate the performance of the AICcd criterion.

Let μ_1, μ_2, σ_1^2, σ_2^2, and σ_{12} be two means, two variances and the covariance of a bivariate normal model respectively. The data set consists of observations on a pair of random variables (y_1, y_2). The candidate class consists of four types of bivariate normal models corresponding to certain parameters and summarized in the following table.

Dimension	Parameter Constraints	Parameter to be estimated
5	None	μ_1, μ_2, σ_1^2, σ_2^2, σ_{12}
4	$\sigma_1^2 = \sigma_2^2 \equiv \sigma^2$	μ_1, μ_2, σ^2, σ_{12}
3	$\sigma_1^2 = \sigma_2^2 \equiv \sigma^2$, $\mu_1 = \mu_2 \equiv \mu$	μ, σ^2, σ_{12}
2	$\sigma_1^2 = \sigma_2^2 \equiv \sigma^2$, $\mu_1 = \mu_2 \equiv \mu$, $\sigma_{12} = 0$	μ, σ^2

In each simulation set, 1000 samples of size 50 are generated using two bivariate normal models of dimension 3 and 4 for the candidate class. In some sets, certain data pairs within each sample are made incomplete by eliminating, according to specified probabilities, either the first or the second observation. Let Pr(y_{1mis}) be the probability that the first observation is missing and the second is observed, and Pr(y_{2mis}) the probability that the second observation is missing and the first is observed. The discard probabilities are fixed at four values: 0.0, 0.15, 0.30 and 0.40. For each generating model, four simulation sets are considered:

Set	True	True
Numbers	Dimension	Parameter Values
1-4	3	$\mu_1 = \mu_2 \equiv \mu = 0$, $\sigma_1^2 = \sigma_2^2 \equiv \sigma^2 = 10$, $\sigma_{12} = 6$
5-8	3	$\mu_1 = \mu_2 \equiv \mu = 0$, $\sigma_1^2 = \sigma_2^2 \equiv \sigma^2 = 10$, $\sigma_{12} = 8$
9-12	4	$\mu_1 = 0$, $\mu_2 = 2$, $\sigma_1^2 = \sigma_2^2 \equiv \sigma^2 = 10$, $\sigma_{12} = 6$
13-16	4	$\mu_1 = 0$, $\mu_2 = 2$, $\sigma_1^2 = \sigma_2^2 \equiv \sigma^2 = 10$, $\sigma_{12} = 8$

For each of the 1000 samples in a set, all four models in the candidate class are fitted to the data using the SEM algorithm. Over the 1000 data sets, the selections are summarized in Table 1.

When there are no missing-data, the criteria KICcd and KIC (respectively AICcd and AIC) give the same selection results (sets 1, 5, 9 and 13). In this setting, KIC outperforms AIC. In the presence of missing-data, and when the correlation between y_1 and y_2 is increased, the selection performance of the criteria improves. Each criterion performs more effectively in sets 5 through 8 than in sets 1 through 4, and more effectively in sets 13 through 16 than in sets 9 through 12. According to Cavanaugh and Shumway (1998), this behavior can be explained by the fact that when there is a large correlation, incomplete-data pairs are less costly since it is possible to accurately impute the missing elements. Moreover, KICcd outperforms AICcd and is more prone than AICcd to underfitting. In every simulation set where data is missing, KICcd overfits to a slightly lesser degree than other criteria. The same remark for AICcd and AIC is formulated by Cavanaugh and

Shumway (1998). Thus, KICcd outperforms in sets 1 through 9, but is outperformed by KIC in the other sets.

The results of Table 2 features the number of correct order selections obtained by each divergence measure in both the first and second set. For both sets, $J_{\mathbf{Y}}(\theta_0, \hat{\theta})$ obtains more correct order selection than $J_{\mathbf{Y}_o}(\theta_0, \hat{\theta})$, $I_{\mathbf{Y}}(\theta_0, \hat{\theta})$ and $I_{\mathbf{Y}_o}(\theta_0, \hat{\theta})$.

Figure 1 provides some insight as to why KICcd tends to outperform AICcd as a selection criterion. The simulated values of $\Omega_{\mathbf{Y}}(d, \theta_0)$ and $\Delta_{\mathbf{Y}}(d, \theta_0)$ for simulation set 3, with the curves which represent the average values of KICcd and AICcd, are plotted against the order d. The shape of the $\Omega_{\mathbf{Y}}(d, \theta_0)$ and $\Delta_{\mathbf{Y}}(d, \theta_0)$ curves indicates that $J_{\mathbf{Y}}(\theta_0, \hat{\theta})$ or $K_{\mathbf{Y}}(\theta_0, \hat{\theta})$ tends to be more effective than $I_{\mathbf{Y}}(\theta_0, \hat{\theta})$ or $D_{\mathbf{Y}}(\theta_0, \hat{\theta})$ in delineating between fitted models of the true order and other fitted models. Thus, $J_{\mathbf{Y}}(\theta_0, \hat{\theta})$ tends to grow to a greater extent than $I_{\mathbf{Y}}(\theta_0, \hat{\theta})$ when the dissimilarity between a fitted model and the true model becomes more pronounced. This explains why model selection criterion based on $J_{\mathbf{Y}}(\theta_0, \hat{\theta})$ may be preferable to one based on $I_{\mathbf{Y}}(\theta_0, \hat{\theta})$.

Figure 2 illustrates the changes in $\Omega_{\mathbf{Y}}(d, \theta_0)$ and $\Omega_{\mathbf{Y}_o}(d, \theta_0)$ as the probabilities of missing-data are increased. To simulate the expected discrepancies, we use the sample from sets 1 through 4. Each curve has been transformed so that its minimum is equal to zero at $d = 3$. The curves in Figure 2 are scaled by dividing each value by the difference between the maximum and the minimum of the $\Omega_{\mathbf{Y}}(d, \theta_0)$ curve from set 1. The curves in Figure 3 are similarly scaled using $\Omega_{\mathbf{Y}}(d, \theta_0)$. probabilities of missing-data are increased, $\Omega_{\mathbf{Y}}(d, \theta_0)$ and $\Omega_{\mathbf{Y}_o}(d, \theta_0)$ decrease for $d = 2$ and increase for $d = 4, 5$. Moreover, $\Omega_{\mathbf{Y}}(d, \theta_0)$ decreases much less quickly than $\Omega_{\mathbf{Y}_o}(d, \theta_0)$ for $d = 2$, and increases much more quickly for $d = 4, 5$. Consequently the minimum of the $\Omega_{\mathbf{Y}}(d, \theta_0)$ curve is well defined, whereas the minimum of the $\Omega_{\mathbf{Y}_o}(d, \theta_0)$ is less pronounced. This explains thatt in the presence of missing-data, an estimator of $\Omega_{\mathbf{Y}}(d, \theta_0)$ may be preferable to an estimator of $\Omega_{\mathbf{Y}_o}(d, \theta_0)$.

Figure 3 summarizes the simulated $\Omega_{\mathbf{Y}}(d, \theta_0)$ curve, for simulation set 3, with the curve which represents the average values of KICcd. This latter curve has the same representation as the simulated $\Omega_{\mathbf{Y}}(d, \theta_0)$ curve.

4.2 Multivariate regression

Another important setting of application of model selection is the multivariate regression model defined by $\mathbf{Y} = \mathbf{X}\beta + \mathbf{U}$, where the rows of $\mathbf{Y}_{n \times p}$ correspond to p response variables on each of n individuals. $\mathbf{X}_{n \times m}$ is a known matrix of covariate values, and $\beta_{m \times p}$ is a matrix of unknown regression parameters. The rows of the error matrix $\mathbf{U}_{n \times p}$ are assumed to be independent, with identical $\mathcal{N}_p(0, \Sigma)$ distribution. The number of unknown parameters in this setting is $d = pm + 0.5p(p+1)$.

We consider a setting where $p = 2$, so that the rows of $\mathbf{Y}$ represent bivariate data pairs. There were eight candidate models stored in an $n \times 8$ matrix $\mathbf{X}$, with a column of ones followed by seven columns of independent standard normal random variables. We consider 1000 samples of size $n = 50$, with $m_0 = 4$ and $m_0 = 6$. Here, $\Sigma = (1-\rho)I_p + \rho J_p$, where J_p is a $p \times p$ matrix of ones, I_p is the identity matrix and the values of ρ are fixed at 0.3, 0.6 and 0.8. For each the p responses, the regression parameters are identical: $(1, 2, 3, 4)$ if for example $m_0 = 4$. Thus, we use the same matrix and values of ρ for each pair (n, p). We obtained the same simulation results for different values of the correlation ρ between responses, so we present only results for $\rho = 0.6$. This example is considered by Bedrick and Tsai (1994) to investigate the performance of the corrected Akaike information criterion.

For each m_0, a collection of five simulation sets are run with the pair of discard probabilities $(\Pr(\mathrm{y}_{1mis}),\ \Pr(\mathrm{y}_{2mis}))$ set at $(0.00, 0.00)$, $(0.00, 0.60)$, $(0.20, 0.40)$, $(0.40, 0.20)$, $(0.60, 0.00)$. The selected dimensions are grouped in three categories:"$< d_0$"(underfitting), "d_0"(correct dimension), and "$> d_0$"(overfitting). Over the 1000 data sets, the selections are summarized in Table 3.

In the presence of incomplete-data, it can be seen that KICcd greatly outperforms all other criteria. Moreover, the tendency to underestimate the correct dimension is zero for all criteria. As in the previous examples, KICcd overfits to a slightly lesser degree than other criteria. Furthermore, in other examples not reported here, we obtained the same results when modifying the values of Σ and $\mathbf{X}$ or the values of β_0 and $\mathbf{X}$. Other examples of simulation on multiple and bivariate regression, not reported here, give the same results as in the preceding example (Hafidi and Mkhadri 2002).

We now present in Table 4 the number of correct order selections obtained by each divergence measure for sets 1 to 5. We see that $J_{\mathbf{Y}}(\theta_0, \hat{\theta})$ obtains more correct order selection than $J_{\mathbf{Y}_o}(\theta_0, \hat{\theta})$, $I_{\mathbf{Y}}(\theta_0, \hat{\theta})$ and $I_{\mathbf{Y}_o}(\theta_0, \hat{\theta})$.

5 Conclusion

In this paper, we have presented and investigated the KICcd criterion for model selection in applications where the observed-data is incomplete. Our criterion estimates the expected complete-data Kullback-Leibler's symmetric discrepancy in the same manner that

KIC estimates the expected incomplete-data symmetric discrepancy. Our simulations indicate that in the presence of incomplete-data KICcd provides better model order choices than other criteria. Moreover, KICcd tends to underfit to a stronger degree than AICcd, and tends to overfit to a lesser degree than KIC. KICcd achieves this performance by incorporating a penalization term for missing information which is lacking in KIC. Thus, AICcd and AIC, respectively, tend to overfit to a stronger degree than KICcd and KIC. The results suggest that the symmetric discrepancy $J_{\mathbf{Y}}(\theta_0, \hat{\theta})$ may provide a foundation for the development of a model selection criteria in the presence of missing-data which is preferable to that provided by the asymmetric discrepancy $I_{\mathbf{Y}}(\theta_0, \hat{\theta})$.
Unlike KIC, the KICcd criterion is based entirely on complete-data tools, and does not require the evaluation of the observed-data empirical log-likelihood, which may be difficult to compute. Moreover, KICcd may be evaluated in this framework by the EM algorithm for assessment of its goodness of fit term and by the SEM algorithm or by other numerical differentiation methods for the estimation of the observed information matrix or its penalty term.

Acknowledgements

The authors wish to thank Professor Cavanaugh J. E. for his help in making the second simulation example by providing us with his Fortran code of the EM algorithm. We also thank Professor Halima Bensmail for reading the paper and for his valuable suggestions.

Appendix: Evaluating DM matrix

EM algorithm

Starting with an initial value $\theta_0 \in \Theta$, the EM algorithm find $\hat{\theta}$ by iterating between the following two steps $(t = 0, 1, ...)$:
E step: Impute the unknown complete-data log-likelihood by its conditional expectation given the current estimate $\theta^{(t)}$:

$$Q(\theta|\theta^{(t)}) = \int_{\mathbf{Y}_m} \{\ln f(\mathbf{Y}|\theta)\} f(\mathbf{Y}_m|\mathbf{Y}_o, \theta^{(t)}) \mathbf{dY}_m.$$

M step: Determine $\theta^{(t+1)}$ by maximizing the imputed log-likelihood $Q(\theta|\theta^{(t)})$:

$$\theta^{(t+1)} = \text{argmax}\, Q(\theta|\theta^{(t)}), \quad \text{forall} \quad \theta \in \Theta.$$

SEM algorithm Let $r_{(i,j)}$ be the (i,j) element of the $d \times d$ matrix **DM** and define $\theta^{(t)}_{(i)}$ as

$$\theta^{(t)}(i) = (\hat{\theta}_{(1)}, ..., \hat{\theta}_{(i-1)}, \theta^{(t)}_{(i)}, \hat{\theta}_{(i+1)}, ..., \hat{\theta}_{(d)}), \qquad i = 1, 2, ..., d. \tag{18}$$

That is, $\theta^{(t)}(i)$ is $\hat{\theta}$ with the ith component active is replaced by the ith component of $\theta^{(t)}$.
Repeat the SEM steps:
INPUT: $\hat{\theta}$ and $\theta^{(t)}$
Repeat step 1 and 2 for for each i

Step 1: Calculate $\theta^{(t)}(i)$ from (18), treat it as input for the EM algorithm, and run one iteration of the EM algorithm (that is, one E step and one M step) to obtain $\tilde{\theta}^{(t+1)}(i)$.

Step 2: Obtain the ratio

$$r^{(t)}_{(i,j)} = \frac{\tilde{\theta}^{(t)}_{(j)}(i) - \hat{\theta}_{(j)}}{\theta^{(t)}_{(i)} - \hat{\theta}_{(i)}} \quad \text{for} \quad j = 1, ..., d$$

OUTPUT: $\theta^{(t+1)}$ and $\{r^{(t)}_{(i,j)}, i, j = 1, ..., d\}$. $\mathbf{DM} = \{r^{\star}_{(i,j)}\}$ where $\{r^{\star}_{(i,j)}\} = \{r^{(t+1)}_{(i,j)}\}$ is such that

$$|r^{(t+1)}_{(i,j)} - r^{(t)}_{(i,j)}| < \epsilon,$$

for some suitable convergence threshold ϵ.

References

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

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **AC-19**, 716-723.

Bedrick, E. J. and Tsai, C. L. (1994). Model selection for multivariate regression in small samples. *Biometrics*, **50**, 226-231.

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.

Cavanaugh, J. E. (1999). A large-sample model selection criterion based on Kullback's symmetric divergence. *Statistics and Probability Letters*, **44**, 333-344.

Cavanaugh, J. E. (2004). Criteria for linear model selection based on Kullback's symmetric divergence. *Australian and New Zealand Journal of Statistics*, **46**, 257-274.

Hafidi, B. and Mkhadri, A. (2002). An Akaike criterion in the presence of incomplete-data. *Communication in Fourth International Conference on Applied Mathematics and Engineering sciences CIMASI. Oct 23-25 2002 Casablanca, Morocco.*

Jamshidian, M. and Jennrich, R. I. (1999). Standard errors for EM estimation. J. R. *Statist. Soc.* B, **62**, 257-270.

Kullback, S. (1968). *Information theory and statistics.*(Dover,New York).

Linhart, H. and Zucchini, W. (1986). *Model selection.* John Wiley, New York.

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

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

Appendix: Tables and figures

Table 1. Selected dimensions for bivariate normal simulations

			Selected dimensions											
Set	d$_0$	P(y$_{1mis}$)	AIC			KIC			AICcd			KICcd		
		P(y$_{2mis}$)	$< d_0$	d_0	$> d_0$	$< d_0$	d_0	$> d_0$	$< d_0$	d_0	$> d_0$	$< d_0$	d_0	$> d_0$
1	3	0.00,0.00	0	766	144	1	880	119	0	766	234	1	880	119
2	3	0.15,0.15	9	754	237	21	862	117	9	802	189	23	889	88
3	3	0.30,0.30	40	718	242	97	778	125	46	780	174	99	828	73
4	3	0.40,0.40	197	573	230	321	574	105	83	745	172	143	766	91
5	3	0.00,0.00	0	768	232	0	876	124	0	768	232	0	876	124
6	3	0.15,0.15	0	756	244	0	874	126	0	790	210	0	904	96
7	3	0.30,0.30	0	746	254	0	867	133	0	800	200	1	902	97
8	3	0.40,0.40	17	723	260	21	810	169	4	739	257	11	827	162
9	4	0.00,0.00	0	831	169	0	917	83	0	831	169	0	917	83
10	4	0.15,0.15	3	819	178	10	891	99	4	812	184	17	878	105
11	4	0.30,0.30	26	792	182	53	848	99	59	752	189	126	767	107
12	4	0.40,0.40	82	700	214	180	724	96	267	593	140	214	700	86
13	4	0.00,0.00	0	843	157	0	919	81	0	843	157	0	919	81
14	4	0.15,0.15	0	824	176	0	892	108	0	816	184	0	888	112
15	4	0.30,0.30	1	811	188	3	885	112	2	801	197	10	869	121
16	4	0.40,0.40	16	801	183	40	855	105	70	759	171	142	758	100

Table 2: Correct order selections for $J_{\mathbf{Y}}(\theta_0,\hat{\theta})$, $I_{\mathbf{Y}}(\theta_0,\hat{\theta})$, $J_{\mathbf{Y}_o}(\theta_0,\hat{\theta})$ and $I_{\mathbf{Y}_o}(\theta_0,\hat{\theta})$ for bivariate normale.

	Divergence							
	Sets 1-4				Sets 5-8			
Pr$_{y_{1mis}}$	$J_{\mathbf{Y}}(\theta_0,\hat{\theta})$	$I_{\mathbf{Y}}(\theta_0,\hat{\theta})$	$J_{\mathbf{Y}_o}(\theta_0,\hat{\theta})$	$I_{\mathbf{Y}_o}(\theta_0,\hat{\theta})$	$J_{\mathbf{Y}}(\theta_0,\hat{\theta})$	$I_{\mathbf{Y}}(\theta_0,\hat{\theta})$	$J_{\mathbf{Y}_o}(\theta_0,\hat{\theta})$	$I_{\mathbf{Y}_o}(\theta_0,\hat{\theta})$
Pr$_{y_{2mis}}$								
0.00,0.00	900	822	900	822	994	988	994	988
0.15,0.15	961	957	959	954	959	954	955	952
0.30,0.30	937	928	917	906	951	945	941	939
0.40,0.40	836	821	799	786	907	883	884	858

Table 3. Selected dimensions for simulation of multivariate regression

			Selected dimensions											
Set	d_0	P_{y_1mis},	AIC			KIC			AICcd			KICcd		
		P_{y_2mis}	$< d_0$	d_0	$> d_0$	$< d_0$	d_0	$> d_0$	$< d_0$	d_0	$> d_0$	$< d_0$	d_0	$> d_0$
1	11	0.0,0.0	0	715	285	0	912	88	0	715	285	0	912	88
2	11	0.0,0.6	0	603	377	0	845	155	0	771	229	0	912	88
3	11	0.2,0.4	0	654	346	0	860	140	0	782	218	0	924	76
4	11	0.4,0.2	0	638	362	0	857	143	0	764	236	0	901	99
5	11	0.6,0.0	0	618	382	0	847	253	0	777	223	0	909	91
6	15	0.0,0.0	0	732	68	0	896	104	0	732	268	0	896	104
7	15	0.0,0.6	0	627	373	0	817	183	0	819	181	2	916	82
8	15	0.2,0.4	0	669	331	0	862	138	0	805	195	1	925	74
9	15	0.4,0.2	0	686	314	0	847	143	0	815	185	0	920	80
10	15	0.6,0.0	0	622	378	0	809	191	0	794	206	2	900	98

Table 4: Correct order selections for $J_{\mathbf{Y}}(\theta_0, \hat{\theta})$, $I_{\mathbf{Y}}(\theta_0, \hat{\theta})$, $J_{\mathbf{Y}_o}(\theta_0, \hat{\theta})$ and $I_{\mathbf{Y}_o}(\theta_0, \hat{\theta})$ for bivariate regression.

	Divergence: Sets 1-5			
Pr_{y_1mis} Pr_{y_2mis}	$J_Y(\theta_0, \hat{\theta})$	$I_Y(\theta_0, \hat{\theta})$	$J_{Y_o}(\theta_0, \hat{\theta})$	$I_{Y_o}(\theta_0, \hat{\theta})$
0.0,0.0	967	860	967	860
0.0,0.6	970	860	913	710
0.2,0.4	974	870	920	715
0.4,0.6	971	870	929	721
0.6,0.0	985	883	912	798

Figure 1. $\Omega_{\mathbf{Y}}(d,\theta_0)$, $\Delta_{\mathbf{Y}}(d,\theta_0)$ and Average values of KICcd and AICcd curves (bivariate normal simulation set 3)

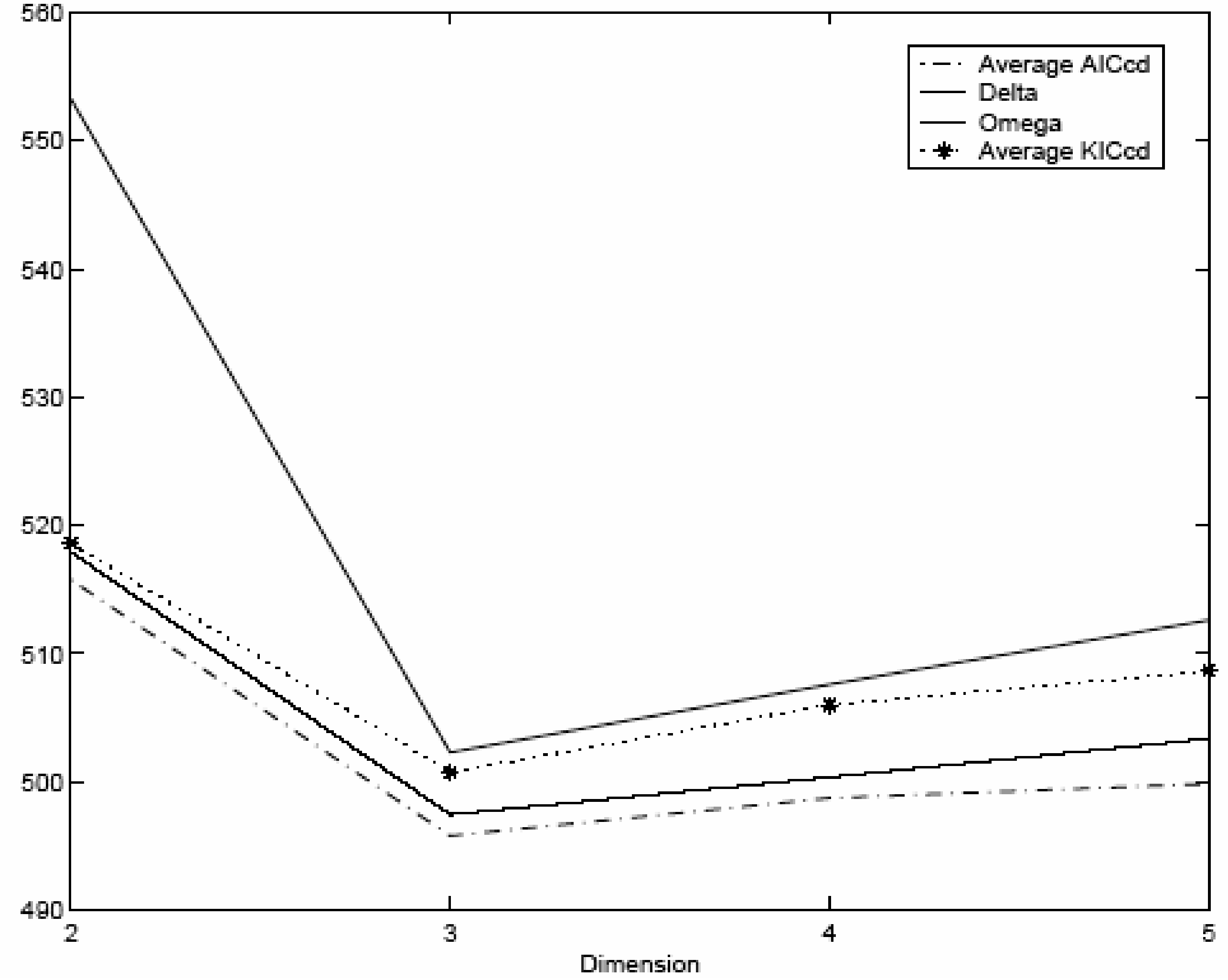


Figure 2 (a). Simulated $\Omega_{\mathbf{Y}}(d, \theta_0)$ and $\Omega_{\mathbf{Y}_o}(d, \theta_0)$ curves for bivariate normal simulation. sets (1,2).

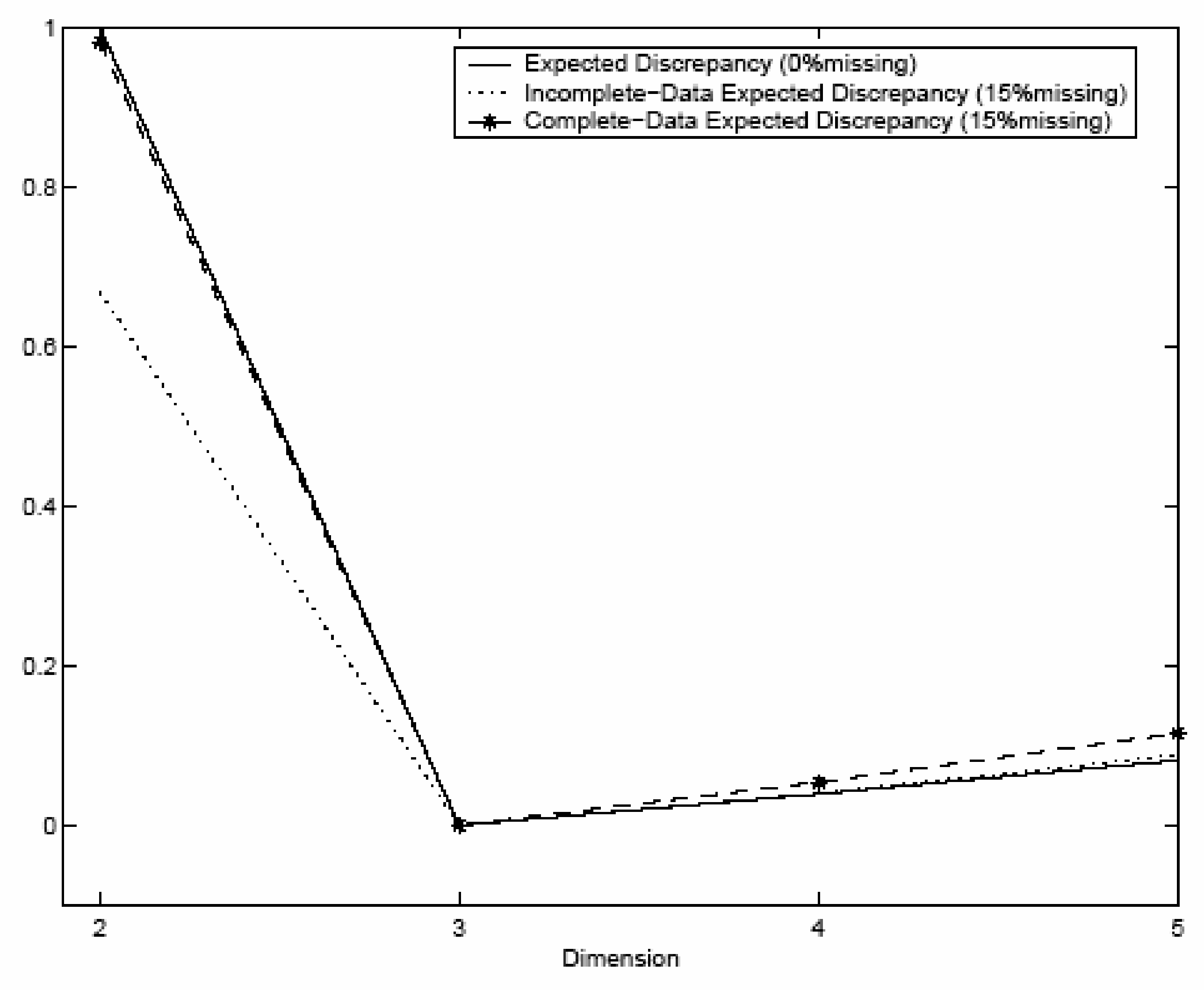


Figure 2 (b). Simulated $\Omega_{\mathbf{Y}}(d, \theta_0)$ and $\Omega_{\mathbf{Y}_o}(d, \theta_0)$ curves for bivariate normal simulation. sets (1,3).

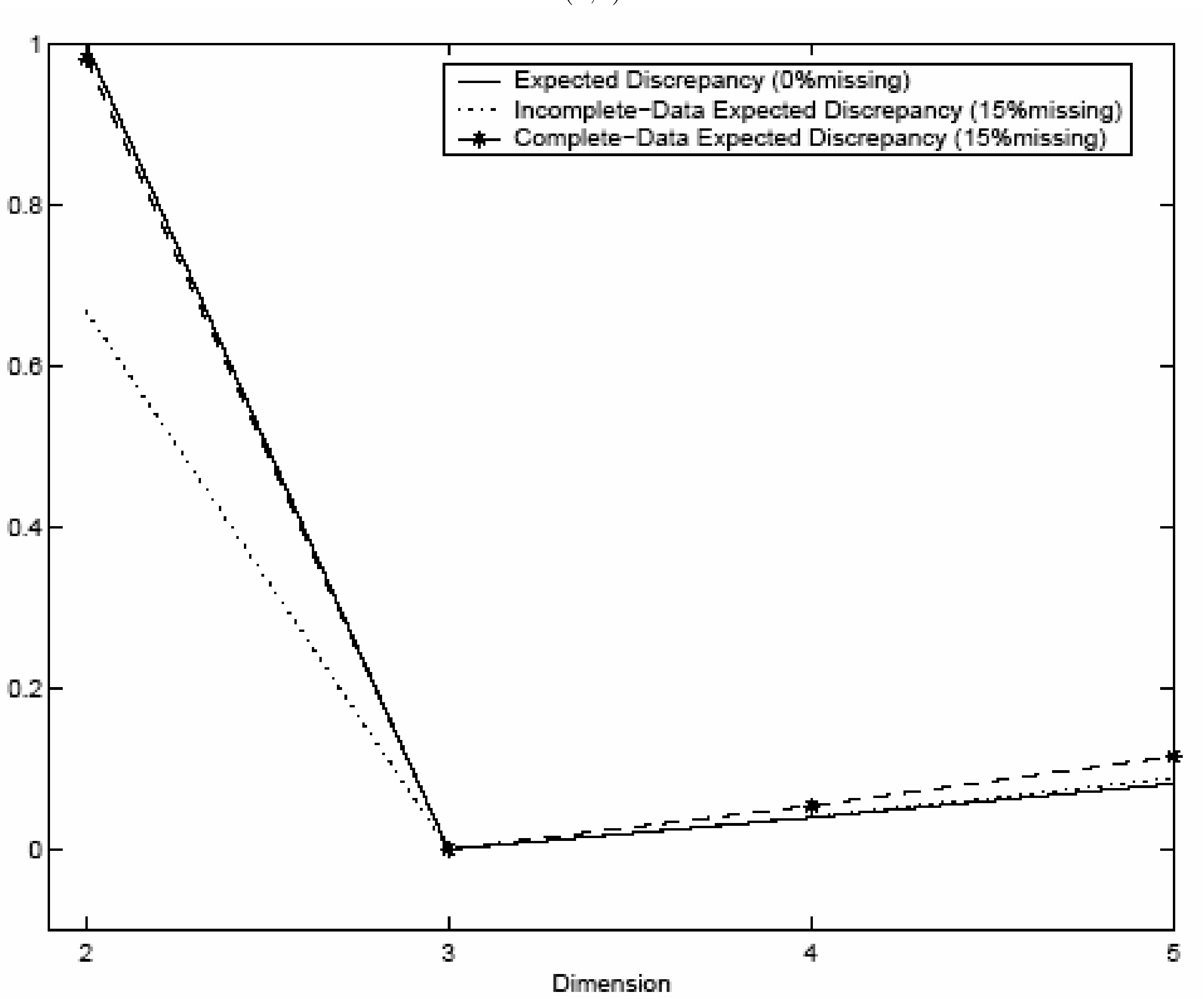


Figure 2 (c). Simulated $\Omega_{\mathbf{Y}}(d,\theta_0)$ and $\Omega_{\mathbf{Y}_o}(d,\theta_0)$ curves for bivariate normal simulation. sets (1,4).

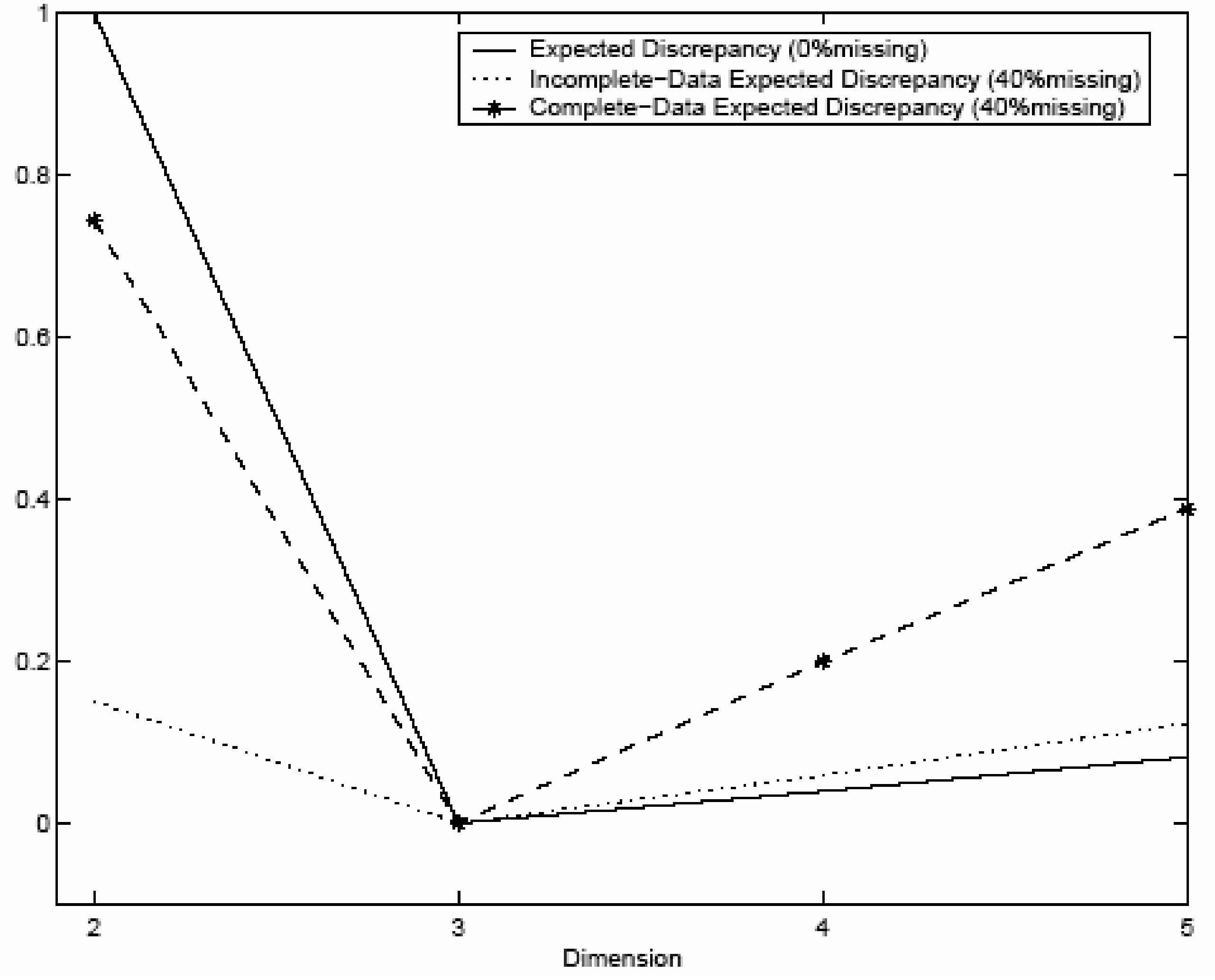
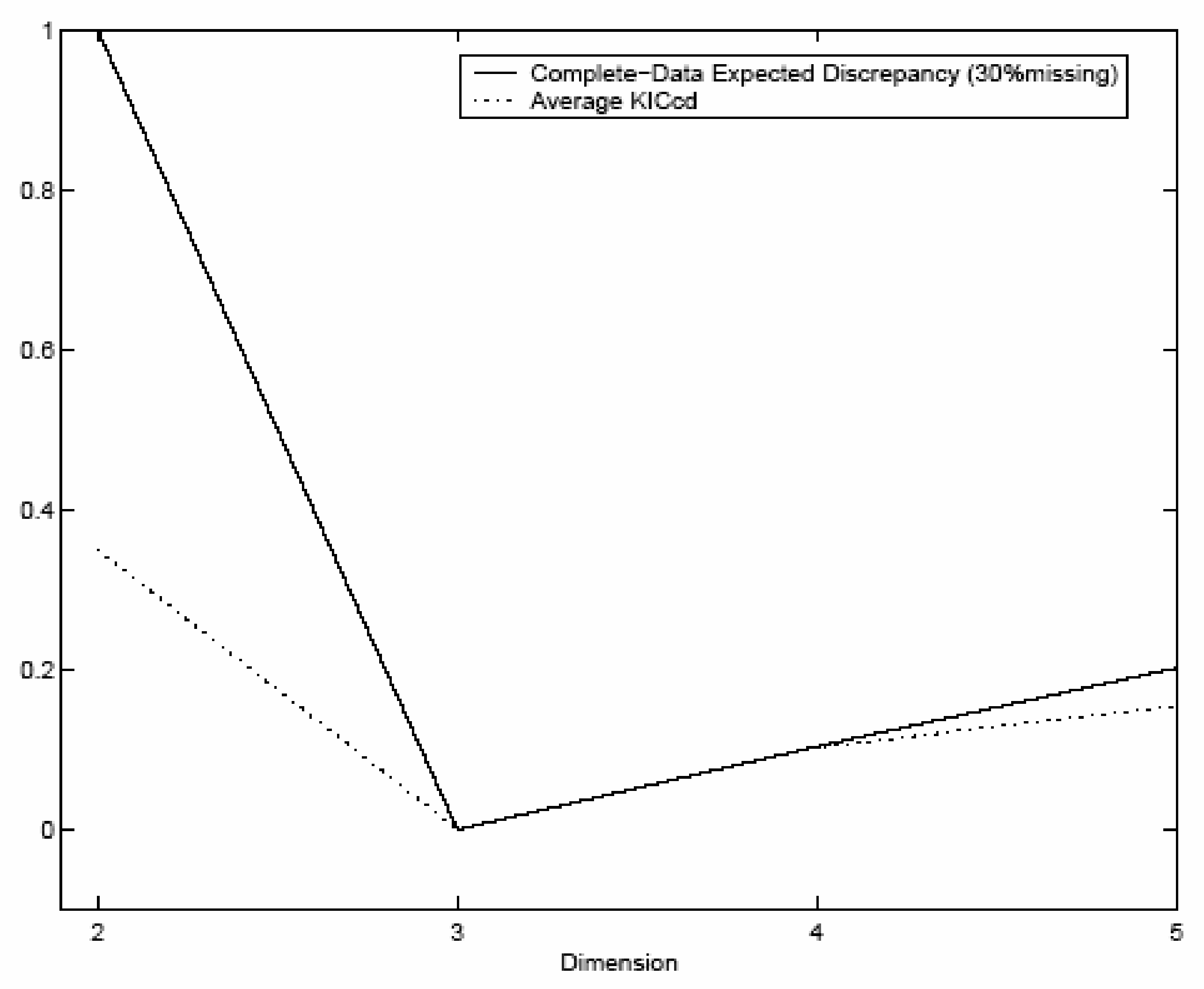


Figure 3. Simulated $\Omega_{\mathbf{Y}}(d, \theta_0)$ and Average KICcd curves
(bivariate normal simulation set 3)


Complete−Data Expected Discrepancy (30%missing)
Average KICcd
Dimension