Unifying the Derivations of Kullback Information Criterion and Corrected Versions

Warangkhana Keerativibool
Department of Mathematics and Statistics, Faculty of Science, Thaksin University,
Phatthalung 93110, Thailand.
E-mail: warang27@gmail.com

Received: 12 June 2013
Accepted: 12 September 2013

Abstract

The Kullback information criterion (KIC) was proposed by Cavanaugh (1999) to serve as an asymptotically unbiased estimator of a variant of Kullback’s symmetric divergence between the true and fitted candidate models. It was arguably more sensitive than the criterion based on the directed divergence. However, for a small sample size or if the dimension of candidate model is large relative to the sample size, it displayed a large negative bias. Many authors, Cavanaugh (2004), Seghouane and Bekara (2004), Hafidi and Mkhadri (2006), proposed the criteria to correct this bias, i.e., the corrected versions of KIC called, respectively, in this paper KICcC, KICcSB, and KICcHM. Because they have multiple formulas, the aims of this paper are to unify and examine the performance of them relative to the AIC family of criteria, using theoretical and extensive simulation study methods. The unifications of the criteria based on Kullback’s symmetric divergence show that KICcC is closest to the expected estimated symmetric discrepancy and has the strongest penalty function under the condition $(1 - p/n)\exp(p/n) < 1$, followed, respectively, by KICcSB, KICcHM, and KIC. This result makes KICcC has the highest efficiency even though it is likely to select an underfitted model.
Keywords: KIC, KICc, Kullback’s directed divergence, Kullback’s symmetric divergence, model selection.

1. Introduction

The Kulback information criterion (KIC) by Cavanaugh [1] and the corrected versions (KICc) by Cavanaugh [2] called KICcC, by Seghouane and Bekara [3] called KICcSB, and by Hafidi and Mkhadri [4] called KICcHM were designed based on Kullback’s symmetric divergence, also known as the J-divergence, in order to assess the dissimilarity between the model generating the data and a fitted candidate model. However, when the dimension of candidate model increases compared to the sample size, the corrected version of the model selection criterion was better than the original version because it produced a bias reduction and strongly improved model selection [2-9]. Although KIC, KICcC, KICcSB, and KICcHM share the same fundamental objective, the justifications of the criteria proceed along different directions, making it difficult to reconcile how the different corrected versions of KIC refine the approximations used to establish KIC in the setting of linear regression model. With this motivation, the aims of this paper are to unify the derivations of KIC and the corrected versions in order to link the justifications of these criteria and the performance of them is then examined by the extensive simulation study. The remainder of this paper is organized as follows. In Section 2, we review the model selection criteria based on Kullback’s directed and symmetric divergences. In Section 3, we show the unifications for the derivations of KIC and the corrected versions. Simulation study for 1,000 realizations of multiple regression models to examine the performance of the AIC and KIC families of criteria is shown in Section 4. Finally, Section 5 is the conclusions, discussion, and further study.

2. A review of model selection criteria based on Kullback’s directed and symmetric divergences

Suppose that the true and the candidate models are, respectively, given by

\[
y = X_0 \beta_0 + \mathbf{e}_0, \quad \mathbf{e}_0 \sim N_n \left( 0, \sigma_0^2 I_n \right),
\]

\[
y = X \beta + \mathbf{e}, \quad \mathbf{e} \sim N_n \left( 0, \sigma^2 I_n \right),
\]

where \( y \) is an \( n \times 1 \) dependent random vector of observations, \( X \) is an \( n \times p \) matrix of independent variables with full-column rank, \( \beta_0 \) and \( \beta \) are \( p \times 1 \) parameter vectors of regression coefficients, \( \mathbf{e}_0 \) and \( \mathbf{e} \) are \( n \times 1 \) noise vectors. The true model is assumed to
be correctly specified or overfitted by all the candidate models. This means that $\beta_0$ has $p_0$ nonzero entries with $0 < p_0 \leq p$ and the rest of the $(p - p_0)$ entries are equal to zero. The $(p + 1) \times 1$ vector of parameters is $\theta_0 = [\beta_0' \; \sigma_0^2]'$ and the maximum likelihood estimator of $\theta_0$ is $\hat{\theta} = [\hat{\beta}' \; \hat{\sigma}^2]'$ where

$$\hat{\beta} = (X'X)^{-1}X'y$$
and

$$\hat{\sigma}^2 = \frac{(y - X\hat{\beta})' (y - X\hat{\beta})}{n}.$$  \hspace{1cm} (3)

The minus twice log likelihood of the candidate model in (2) when replacing the dependent vector $y$ in (1) is defined by

$$-2\log L(\theta | y) = n \log 2\pi + n \log \sigma^2 + \frac{1}{\sigma^2} e_0'e_0 + \frac{1}{\sigma^2} (\beta_0 - \beta)' X'X(\beta_0 - \beta) + \frac{2}{\sigma^2} e_0'X(\beta_0 - \beta).$$ \hspace{1cm} (4)

A well-known measure to separate the discrepancy between two models is given by Kullback’s directed divergence or l-divergence [10],

$$2I(\theta_0, \theta) = E_{\theta_0} \left\{ 2\log \frac{L(\theta_0 | y)}{L(\theta | y)} \right\} = d(\theta_0, \theta) - d(\theta_0, \theta_0),$$

where

$$d(\theta_0, \theta) = E_{\theta_0} \left\{ -2\log L(\theta | y) \right\}, \quad d(\theta_0, \theta_0) = E_{\theta_0} \left\{ -2\log L(\theta_0 | y) \right\},$$

and the expectation $E_{\theta_0}$ is taken with respect to the true model. Because $d(\theta_0, \theta_0)$ does not depend on $\theta$, any ranking of the candidate models according to $2I(\theta_0, \theta)$ would be identical to ranking them according to $d(\theta_0, \theta)$. Given a set of maximum likelihood estimators $\hat{\theta}$ in (3), the estimated directed measure $d(\theta_0, \theta)$ is

$$d(\theta_0, \hat{\theta}) = E_{\theta_0} \left\{ -2\log L(\theta | y) \right\}_{\theta = \hat{\theta}}$$

$$= n \log 2\pi + n \log \hat{\sigma}^2 + \frac{n\sigma_0^2}{\hat{\sigma}^2} + \frac{1}{\hat{\sigma}^2} (\beta_0 - \hat{\beta})' X'X(\beta_0 - \hat{\beta}).$$ \hspace{1cm} (5)

However, the evaluation in (5) is not possible because it requires the knowledge of $\theta_0$, Akaike [11-12] proposed an asymptotically unbiased estimator of

$$\Delta(\theta_0, p) = E_{\theta_0} \left\{ d(\theta_0, \hat{\theta}) \right\} \hspace{1cm} (6)$$
as
$AIC = n \log \hat{\sigma}^2 + 2(p + 1), \quad (7)$

i.e., $E_{\theta_0} \{ AIC \} + o(1) = \Delta(\theta_0, p)$.

Because of a large negative bias of AIC when the sample size is small or the dimension of candidate model is large relative to the sample size, Hurvich and Tsai [5] proposed an exactly unbiased estimator of (6) as follows:

$$AICc = n \log \hat{\sigma}^2 + \frac{2n(p + 1)}{n - p - 2}, \quad (8)$$

i.e., $E_{\theta_0} \{ AICc \} = \Delta(\theta_0, p)$.

Cavanaugh [1], Seghouane and Bekara [3], Seghouane [13] summarized that the directed divergence produced too underfitted value of model selection, and then it tended to be large for overparameterized models. An alternate measure to prevent both overfitting and underfitting problems is obtained by reversing the roles of two models in the definition of the measure, called Kullback’s symmetric divergence or J-divergence,

$$2J(\theta_0, \theta) = 2I(\theta_0, \theta) + 2I(\theta_0, \theta_0) = \left[ d(\theta_0, \theta) - d(\theta_0, \theta_0) \right] + \left[ d(\theta, \theta_0) - d(\theta, \theta) \right]$$,

where

$$d(\theta_0, \theta) = E_{\theta_0} \left\{ -2 \log L(\theta | y) \right\}, \quad d(\theta_0, \theta_0) = E_{\theta_0} \left\{ -2 \log L(\theta_0 | y) \right\},$$

$$d(\theta, \theta_0) = E_\theta \left\{ -2 \log L(\theta_0 | y) \right\}, \quad \text{and} \quad d(\theta, \theta) = E_\theta \left\{ -2 \log L(\theta | y) \right\}.$$

Dropping $d(\theta_0, \theta_0)$, the ranking of the candidate models according to $2J(\theta_0, \theta)$ is identical to ranking them according to

$$K(\theta_0, \theta) = d(\theta_0, \theta) + d(\theta, \theta) - d(\theta_0, \theta)$$.

Given a set of maximum likelihood estimators $\hat{\theta}$ in (3), the estimated symmetric measure $K(\theta_0, \theta)$ is

$$K(\theta_0, \hat{\theta}) = d(\theta_0, \hat{\theta}) + d(\hat{\theta}, \theta_0) - d(\theta_0, \theta_0)$$, \quad (9)

where $d(\theta_0, \hat{\theta})$ is exhibited in (5),

$$d(\hat{\theta}, \theta_0) = E_{\theta_0} \left\{ -2 \log L(\theta_0 | y) \right\}|_{\theta = \hat{\theta}}$$

$$= n \log 2\pi + n \log \sigma_0^2 + \frac{n\hat{\sigma}^2}{\sigma_0^2} + \frac{1}{\sigma_0^2} \left( \hat{\beta} - \beta_0 \right)' X' X (\hat{\beta} - \beta_0). \quad (10)$$

and
\[ d(\hat{\theta}, \tilde{\theta}) = E_\theta \{ -2 \log L(\theta | y) \} \bigg|_{\theta = \tilde{\theta}} = n \log 2\pi + n \log \hat{\sigma}^2 + n. \]  

(11)

Yet, evaluating \( K(\theta_0, \tilde{\theta}) \) in (9) requires \( \theta_0 \), Cavanaugh [1] proposed an asymptotically unbiased estimator of

\[ \Omega(\theta_0, p) = E_{\theta_0} \{ K(\theta_0, \tilde{\theta}) \} \]  

(12)

as

\[ KIC = n \log \hat{\sigma}^2 + 3(p + 1), \]  

(13)

i.e., \( E_{\theta_0} \{ KIC \} + o(1) = \Omega(\theta_0, p) \).

Seghouane and Bekara [3] proposed an exactly unbiased estimator of (12) in order to correct a large negative bias of KIC in (13) as follows:

\[ KICc = n \log \hat{\sigma}^2 + \frac{2n(p + 1)}{n - p - 2} - n\psi\left(\frac{n - p}{2}\right) + n \log \left(\frac{n}{2}\right), \]  

(14)

i.e., \( E_{\theta_0} \{ KICc \} = \Omega(\theta_0, p) \).

Because the phi \( (\psi) \) or digamma function in KICc has no closed-form solution, Cavanaugh [2], Seghouane and Bekara [3], Hafidi and Mkhadri [4] gave the asymptotically unbiased estimators of (12) called, respectively, in this paper KICcC, KICcSB, and KICcHM,

\[ KICcC = n \log \hat{\sigma}^2 + n \log \left(\frac{n}{n - p}\right) + \frac{n\left[(n - p)(2p + 3) - 2\right]}{(n - p - 2)(n - p)}, \]  

(14)

\[ KICcSB = n \log \hat{\sigma}^2 + \frac{(p + 1)(3n - p - 2)}{n - p - 2} + \frac{p}{n - p}, \]  

(15)

\[ KICcHM = n \log \hat{\sigma}^2 + \frac{(p + 1)(3n - p - 2)}{n - p - 2}. \]  

(16)

3. The unified derivations of KIC and KICc

To begin the unification of the derivations KIC, KICcC, KICcSB, and KICcHM, we consider the expectation of the discrepancies in (5), (10), and (11) with respect to the true model [3],

\[ E_{\theta_0} \{ d(\theta_0, \tilde{\theta}) \} = n \log 2\pi + E_{\theta_0} \{ n \log \hat{\sigma}^2 \} + E_{\theta_0} \left\{ \frac{n\sigma^2}{\hat{\sigma}^2} \right\} + E_{\theta_0} \left\{ \frac{1}{\sigma^2} \left( \beta_0 - \hat{\beta} \right)' X X \left( \beta_0 - \hat{\beta} \right) \right\}, \]  

(17)
\[ E_{\theta_0} \left\{ d\left(\hat{\theta}, \theta_0\right) \right\} = n \log 2\pi + n \log \sigma_0^2 + E_{\theta_0} \left\{ \frac{n \sigma^2}{\sigma_0^2} \right\} + E_{\theta_0} \left\{ \frac{1}{\sigma_0^2} (\beta - \beta_0)' X'X (\beta - \beta_0) \right\}. \]  

(18)

\[ E_{\theta_0} \left\{ d\left(\hat{\theta}, \hat{\theta} \right) \right\} = n \log 2\pi + E_{\theta_0} \left\{ n \log \sigma^2 \right\} + n. \]  

(19)

From the fact that the terms \( \frac{n \sigma^2}{\sigma_0^2} \) and \( \frac{1}{\sigma_0^2} (\beta - \beta_0)' X'X (\beta - \beta_0) \) are the independent \( \chi^2 \) distributions with the degrees of freedom which are, respectively, \( n - p \) and \( p \), we have

\[ E_{\theta_0} \left\{ \frac{n \sigma^2}{\sigma_0^2} \right\} = n - p \]  

and

\[ E_{\theta_0} \left\{ \frac{1}{\sigma_0^2} (\beta - \beta_0)' X'X (\beta - \beta_0) \right\} = p. \]  

(20)

Using the facts in (20), we have

\[ E_{\theta_0} \left\{ \frac{n \sigma^2}{\sigma_0^2} \right\} = n - p \]  

and

\[ E_{\theta_0} \left\{ \frac{1}{\sigma_0^2} (\beta - \beta_0)' X'X (\beta - \beta_0) \right\} = p. \]  

(21)

Substituting the results in (21) into the expected discrepancy in (17) leads to

\[ \Delta\left(\theta_0, p\right) = E_{\theta_0} \left\{ d\left(\hat{\theta}, \theta_0\right) \right\} = n \log 2\pi + E_{\theta_0} \left\{ n \log \sigma^2 \right\} + \frac{n^2}{n - p - 2} + \frac{np}{n - p - 2} \]

\[ = n \left( \log 2\pi + 1 \right) + E_{\theta_0} \left\{ \text{AICc} \right\}, \]  

(22)

where AICc is the corrected version of AIC that was exhibited in (8).

Replacing the facts in (20) into the expected discrepancy in (18) yields

\[ E_{\theta_0} \left\{ d\left(\hat{\theta}, \hat{\theta} \right) \right\} = n \left( \log 2\pi + 1 \right) + n \log \sigma^2. \]  

(23)

Using the results in (19), (22), and (23), the expected value of \( K\left(\theta_0, \hat{\theta} \right) \) in (9) becomes

\[ \Omega\left(\theta_0, p\right) = E_{\theta_0} \left\{ K\left(\theta_0, \hat{\theta} \right) \right\} = n \left( \log 2\pi + 1 \right) + E_{\theta_0} \left\{ \text{AICc} \right\} - E_{\theta_0} \left\{ n \log \sigma^2 \right\}, \]  

(24)

where AICc is the corrected version of AIC that was exhibited in (8).
It is noteworthy that, in KIC and various corrected versions derived from \( K(\theta_0, \hat{\theta}) \) in (9), the differences in all formulas come from the last term of the right-hand side in (24). Therefore, in order to show the connections of KIC, KICc, KICcSB, and KICcHLM, we give the following lemmas.

**Lemma 1.** \(- E_{\theta_0} \left\{ n \log \frac{\hat{\sigma}^2}{\sigma_o^2} \right\} = -n \log \left( \frac{n - p}{2} \right) + \frac{n}{n - p} + n \log \left( \frac{n}{2} \right) + o \left( \frac{n}{(n-p)^2} \right) \). (25)

**Proof.** From [14-15] we have, respectively,

\[
E_{\theta_0} \left\{ \log \chi^2_{df} \right\} = \psi \left( \frac{df}{2} \right) + \log 2 \text{ and } \psi (x) = \log x - \frac{1}{2x} + o \left( \frac{1}{x^2} \right) \text{ as } x \to \infty. \quad (26)
\]

Applying the fact \( E_{\theta_0} \left\{ n \hat{\sigma}^2 / \sigma_o^2 \right\} = n - p \) in (20) and the facts in (26), we have

\[
- E_{\theta_0} \left\{ n \log \frac{\hat{\sigma}^2}{\sigma_o^2} \right\} = - E_{\theta_0} \left\{ n \log \frac{n \hat{\sigma}^2}{\sigma_o^2} \right\} + n \log n = - \left[ \psi \left( \frac{n - p}{2} \right) + \log 2 \right] + n \log n
\]

\[
= -n \left[ \log \left( \frac{n - p}{2} \right) - \frac{1}{n - p} + o \left( \frac{1}{(n-p)^2} \right) \right] - n \log 2 + n \log n
\]

\[
= -n \log \left( \frac{n - p}{2} \right) + \frac{n}{n - p} + n \log \left( \frac{n}{2} \right) + o \left( \frac{n}{(n-p)^2} \right).
\]

**Lemma 2.** \(-n \log \left( \frac{n - p}{2} \right) + \frac{n}{n - p} + n \log \left( \frac{n}{2} \right) + o \left( \frac{n}{(n-p)^2} \right) = p + \frac{n}{n - p} + o \left( \frac{p^2}{n} \right) + o \left( \frac{n}{(n-p)^2} \right) \). (27)

**Proof.** Applying the first-order Taylor’s series expansion to expand the term \( \log \left( \frac{n - p}{2} \right) \)

about \( n/2 \), i.e.,

\[
\log \left( \frac{n - p}{2} \right) = \log \left( \frac{n}{2} \right) - \frac{p}{n} + o \left( \frac{p^2}{n} \right),
\]

to obtain the approximation in (27).

**Lemma 3.** \( p + \frac{n}{n - p} + o \left( \frac{p^2}{n} \right) + o \left( \frac{n}{(n-p)^2} \right) = (p + 1) + o(1). \) (28)
Proof. Rearrange \( p+n/(n-p) \) to be \((p+1)+p/(n-p)\). As \( n \to \infty \) and \( p \) is held constant, the term

\[
\frac{p}{n-p} + o\left(\frac{p^2}{n}\right) + o\left(\frac{n}{(n-p)^2}\right)
\]

is \( o(1) \) which yields the approximation in (28).

Applying Lemma 1 into \( \Omega(\theta_0, p) \) in (24), we obtain

\[
\Omega(\theta_0, p) = n \left( \log 2\pi + 1 \right) + E_{\theta_0} \{ \text{AICc} \} - n \log \left( \frac{n-p}{2} \right) + \frac{n}{n-p} + n \log \left( \frac{n}{2} \right) + o \left( \frac{n}{(n-p)^2} \right)
\]

\[
= n \left( \log 2\pi + 1 \right) + E_{\theta_0} \left\{ \text{KICc} + o \left( \frac{n}{(n-p)^2} \right) \right\},
\]

where \( \text{KICc} \) is the corrected version of \( \text{KIC} \) from Cavanaugh [2] that was exhibited in (14).

Applying Lemmas 1 and 2 into \( \Omega(\theta_0, p) \) in (24), we obtain

\[
\Omega(\theta_0, p) = n \left( \log 2\pi + 1 \right) + E_{\theta_0} \{ \text{AICc} \} + p + \frac{n}{n-p} + o \left( \frac{p^2}{n} \right) + o \left( \frac{n}{(n-p)^2} \right)
\]

\[
= n \left( \log 2\pi + 1 \right) + E_{\theta_0} \left\{ \text{KICcSB} + o \left( \frac{p^2}{n} \right) + o \left( \frac{n}{(n-p)^2} \right) \right\},
\]

where \( \text{KICcSB} \) is the corrected version of \( \text{KIC} \) from Seghouane and Bekara [3] that was exhibited in (15).

Applying Lemmas 1, 2, and 3 into \( \Omega(\theta_0, p) \) in (24), we obtain

\[
\Omega(\theta_0, p) = n \left( \log 2\pi + 1 \right) + E_{\theta_0} \{ \text{AICc} \} + (p+1) + o(1)
\]

\[
= n \left( \log 2\pi + 1 \right) + E_{\theta_0} \left\{ \text{KICcHM} + o(1) \right\},
\]

where \( \text{KICcHM} \) is the corrected version of \( \text{KIC} \) from Hafidi and Mkhadri [4] that was exhibited in (16).

The connections of \( \text{KIC}, \text{KICcHM}, \text{KICcSB}, \) and \( \text{KICcC} \) are given by

\[
\text{KICcHM} = \text{KIC} + \frac{2(p+1)(p+2)}{n-p-2},
\]

(29)
\[ KIC_{SB} = KIC_{HM} + \frac{p}{n-p} \] \hspace{1cm} (30)

\[ KIC_C = KIC_{SB} + n \log \left( \frac{n}{n-p} \right) - p. \] \hspace{1cm} (31)

From the connection in (29), we found that the term
\[ \frac{2(p+1)(p+2)}{n-p-2} \] \hspace{1cm} (32)

is not greater than zero if and only if \( n-p > 2 \) and \( p \) belongs to the set of \([-2, -1]\). Whereas, the term of the connection in (30),
\[ \frac{p}{n-p} \] \hspace{1cm} (33)

is not greater than zero if and only if \( n-p > 0 \) and \( p \) belongs to the set of \((-\infty, 0]\). Therefore, we can argue that the terms in (32) and (33) have values of at least zero because \( p \) represents the number of regression coefficients which is an integer that has the value of at least one and both terms in (32) and (33) are very close to zero if the ratio of \( p/n \) tends to zero. This conclusion links to \( KIC_{SB} \geq KIC_{HM} \geq KIC \). While the term
\[ n \log \left( \frac{n}{n-p} \right) - p \] \hspace{1cm} (34)

has the value in the range \([-p, \infty)\) where it is close to the lower bound \(-p\) if the ratio of \( p/n \) tends to zero. If the value of \( p \) is fixed, this term is the decreasing function of \( n \), whereas when the value of \( n \) is fixed, it is the increasing function of \( p \). Whenever \( n-p > 0 \) and the condition
\[ (1-p/n)\exp\left(\frac{p}{n}\right) < 1 \] \hspace{1cm} (35)

is true, we have the term in (34) being greater than zero. This means that the penalty function of \( KIC_C \) is stronger than other criteria, \( KIC_{SB}, KIC_{HM} \), and \( KIC \), under the condition in (35). The strong penalty may cause \( KIC_C \) to have the maximum frequency of the correct order being selected. However, occasionally it causes the model selection criterion to select underparameterized models [14]. This confusion is studied by the extensive simulation in the next section.

4. Simulation study
To examine the model selection criteria performance, we generated 1,000 realizations of true multiple regression models in (1) for four cases as follows. Model I represents a very weakly identifiable true model with large dimension of the model:

\[ y_1 = 1 + 0.5X_2 + 0.1X_3 + 0.05X_4 + 0.01X_5 + 0.005X_6 + 0.001X_7 + 0.0005X_8 + \varepsilon_1. \]

Model II represents a weakly identifiable true model with small dimension of the model:

\[ y_2 = 1 + 0.5X_2 + 0.25X_3 + \varepsilon_2. \]

Model III represents a very strongly identifiable true model with small dimension of the model:

\[ y_3 = 1 + 2X_2 + 3X_3 + 4X_4 + \varepsilon_3. \]

Model IV represents a strongly identifiable true model with large dimension of the model:

\[ y_4 = 1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7 + X_8 + \varepsilon_4. \]

Model I and Model II represent the weakly identifiable true models which mean they are not easily identified compared to the strongly identifiable true models such as Model III and Model IV. From a previous study, Kundu and Murali [16] concluded that the criteria performance did not change much when the true variance \( \sigma_0^2 \) and the distribution of error random variable \( \varepsilon_0 \) in (1) were changed. As a result, we have taken \( \varepsilon_0 \) to be normally distributed with zero mean and \( \sigma_0^2 \) is assumed to be equal to 1. For each model, four different sample sizes are split into two categories: small sample \((n = 15, 25)\) and large sample \((n = 100, 500)\). Ten candidate variables, \( X_1 \) until \( X_{10} \), are stored in an \( n \times 10 \) matrix \( X \) of the candidate model in (2), with a column of ones, followed by nine independent identically distributed normal random variables with zero mean and variance-covariance matrix equal to identity matrix \( I_{10} \). The candidate models include the columns of \( X \) in a sequentially nested fashion; i.e., columns 1 to \( p \) define the design matrix for the candidate model with dimension \( p \). The criteria considered in this simulation are divided into two families. Firstly, is the criteria based on Kullback’s directed divergence: AIC in (7) and AICc in (8). Secondly, is the criteria based on Kullback’s symmetric divergence: KIC in (13), KICcC in (14), KICcSB in (15), and KICcHM in (16). Model selection criteria performance is examined by a consistent measure which is a measure of counting the frequency of order being selected. Particularly for the case
of true model being weakly identifiable, we use an efficient measure which is the observed $L_2$ efficiency. This is a useful measure when the criteria do not select the correct model. The observed $L_2$ distance or squared error distance, scaled by $1/n$, between the true model in (1) and the fitted candidate model in (2) is defined as [8, 17]

$$L_2(p) = \frac{1}{n} (\beta_0 - \hat{\beta})' X^T (X^T (\beta_0 - \hat{\beta})).$$

Observed $L_2$ efficiency is defined by the ratio

$$\text{Observed } L_2 \text{ efficiency} = \frac{\min_{p \in P} L_2(p)}{L_2(p_s)},$$

where $P$ is the class of all possible candidate models, $p$ is the rank of fitted candidate model, and $p_s$ is the model selected by specific model selection criterion. The closer the selected model is to the true model, the higher the efficiency. Therefore, the best model selection criterion will select a model which yields high efficiency even in small samples or if the true model is weakly identifiable. For 1,000 realizations, the results of comparing the model selection criteria performance are shown in Table 1 and 2. Columns “d” and “K” in Table 1 stand for the estimated measures in (5) and (9), respectively. The conclusions of this simulation are as follows. In Table 1, for the small sample size and the true model is somewhat difficult to identify, such as Model I, Model II for $n = 15, 25$, and Model IV for $n = 15$, the original criteria AIC and KIC perform better than their corrected versions. When the sample size is still small but the true model is easily to identify, such as Model III for $n = 15, 25$ and Model IV for $n = 25$, the corrected versions work better. For the large sample size but the true model is very difficult to detect, such as Model I for $n = 100, 500$, the AIC family of criteria performs better than the KIC family. When the sample size is still large and the true model can be specified more easily, such as Model II, Model III, and Model IV for $n = 100, 500$, the KIC family performs the best. This simulation also found that when the true model is very difficult to detect, such as Model I and the sample size is small $n = 15, 25$, the estimated symmetric measure in (9) has the opportunity to cause more underfitted order being selected than the estimated directed measure in (5). This result contributes the criteria in KIC family to having a low frequency of choosing the correct model. In Table 2, the observed $L_2$ efficiency suggests that KICcC in KIC family is the best criterion for all sample sizes of a weakly identifiable true model.
Table 1. Frequency of the model order being selected by each criterion for 1,000 realizations.

<table>
<thead>
<tr>
<th>Model</th>
<th>n</th>
<th>Order</th>
<th>Criteria</th>
<th>AIC</th>
<th>AICc</th>
<th>KIC</th>
<th>KICcHM</th>
<th>KICcSB</th>
<th>KICcC</th>
<th>d</th>
<th>K</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>15</td>
<td>Underfitted</td>
<td></td>
<td>596</td>
<td>1000</td>
<td>837</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
<td>982</td>
<td>986</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Correct</td>
<td></td>
<td>54</td>
<td>0</td>
<td>26</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Overfitted</td>
<td></td>
<td>350</td>
<td>0</td>
<td>137</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>18</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>25</td>
<td>Underfitted</td>
<td></td>
<td>859</td>
<td>998</td>
<td>972</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
<td>987</td>
<td>992</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Correct</td>
<td></td>
<td>39</td>
<td>1</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Overfitted</td>
<td></td>
<td>102</td>
<td>1</td>
<td>17</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>13</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>Underfitted</td>
<td></td>
<td>944</td>
<td>974</td>
<td>993</td>
<td>998</td>
<td>999</td>
<td>999</td>
<td>998</td>
<td>998</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Correct</td>
<td></td>
<td>23</td>
<td>14</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Overfitted</td>
<td></td>
<td>33</td>
<td>12</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>Underfitted</td>
<td></td>
<td>958</td>
<td>962</td>
<td>998</td>
<td>998</td>
<td>998</td>
<td>999</td>
<td>1000</td>
<td>1000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Correct</td>
<td></td>
<td>21</td>
<td>21</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Overfitted</td>
<td></td>
<td>21</td>
<td>17</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>II</td>
<td>15</td>
<td>Underfitted</td>
<td></td>
<td>284</td>
<td>820</td>
<td>542</td>
<td>859</td>
<td>864</td>
<td>875</td>
<td>577</td>
<td>547</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Correct</td>
<td></td>
<td>132</td>
<td>123</td>
<td>148</td>
<td>111</td>
<td>109</td>
<td>105</td>
<td>423</td>
<td>453</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Overfitted</td>
<td></td>
<td>584</td>
<td>57</td>
<td>310</td>
<td>30</td>
<td>27</td>
<td>20</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>25</td>
<td>Underfitted</td>
<td></td>
<td>374</td>
<td>653</td>
<td>575</td>
<td>716</td>
<td>720</td>
<td>727</td>
<td>368</td>
<td>344</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Correct</td>
<td></td>
<td>244</td>
<td>235</td>
<td>264</td>
<td>235</td>
<td>231</td>
<td>226</td>
<td>631</td>
<td>655</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Overfitted</td>
<td></td>
<td>382</td>
<td>112</td>
<td>161</td>
<td>49</td>
<td>49</td>
<td>47</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>Underfitted</td>
<td></td>
<td>109</td>
<td>133</td>
<td>214</td>
<td>230</td>
<td>231</td>
<td>232</td>
<td>34</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Correct</td>
<td></td>
<td>609</td>
<td>642</td>
<td>676</td>
<td>677</td>
<td>678</td>
<td>680</td>
<td>966</td>
<td>974</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Overfitted</td>
<td></td>
<td>282</td>
<td>225</td>
<td>110</td>
<td>93</td>
<td>91</td>
<td>88</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>Underfitted</td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Correct</td>
<td></td>
<td>737</td>
<td>751</td>
<td>890</td>
<td>895</td>
<td>895</td>
<td>896</td>
<td>1000</td>
<td>1000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Overfitted</td>
<td></td>
<td>263</td>
<td>249</td>
<td>110</td>
<td>105</td>
<td>105</td>
<td>104</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
## Table 1. (Continued).

<table>
<thead>
<tr>
<th>Model</th>
<th>Order</th>
<th>Criteria</th>
<th>$n$</th>
<th>$d$</th>
<th>$K$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>AIC</td>
<td>AICc</td>
<td>KIC</td>
<td>KICcHM</td>
</tr>
<tr>
<td>III</td>
<td>15 Underfitted</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Correct</td>
<td>325</td>
<td>939</td>
<td>583</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Overfitted</td>
<td>675</td>
<td>61</td>
<td>417</td>
</tr>
<tr>
<td></td>
<td>25 Underfitted</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Correct</td>
<td>687</td>
<td>756</td>
<td>855</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Overfitted</td>
<td>313</td>
<td>244</td>
<td>145</td>
</tr>
<tr>
<td></td>
<td>100 Underfitted</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Correct</td>
<td>713</td>
<td>731</td>
<td>885</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Overfitted</td>
<td>287</td>
<td>269</td>
<td>115</td>
</tr>
<tr>
<td></td>
<td>500 Underfitted</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Correct</td>
<td>854</td>
<td>864</td>
<td>951</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Overfitted</td>
<td>146</td>
<td>136</td>
<td>49</td>
</tr>
<tr>
<td>IV</td>
<td>15 Underfitted</td>
<td>36</td>
<td>887</td>
<td>94</td>
<td>943</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Correct</td>
<td>444</td>
<td>113</td>
<td>532</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Overfitted</td>
<td>520</td>
<td>0</td>
<td>374</td>
</tr>
<tr>
<td></td>
<td>25 Underfitted</td>
<td>5</td>
<td>31</td>
<td>9</td>
<td>60</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Correct</td>
<td>710</td>
<td>950</td>
<td>840</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Overfitted</td>
<td>285</td>
<td>19</td>
<td>151</td>
</tr>
<tr>
<td></td>
<td>100 Underfitted</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Correct</td>
<td>815</td>
<td>882</td>
<td>925</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Overfitted</td>
<td>185</td>
<td>118</td>
<td>75</td>
</tr>
<tr>
<td></td>
<td>500 Underfitted</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Correct</td>
<td>854</td>
<td>864</td>
<td>951</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Overfitted</td>
<td>146</td>
<td>136</td>
<td>49</td>
</tr>
</tbody>
</table>

**Note:** Boldface type indicates the maximum frequency of correct order being selected.
Table 2. Average and standard deviation of the observed $L_2$ efficiency over 1,000 realizations.

<table>
<thead>
<tr>
<th>Circumstance</th>
<th>Stat.</th>
<th>AIC</th>
<th>AICc</th>
<th>KIC</th>
<th>KICcHM</th>
<th>KICcSB</th>
<th>KICcC</th>
</tr>
</thead>
<tbody>
<tr>
<td>weakly identifiable</td>
<td>Ave. $L_2$ eff.</td>
<td>0.5332</td>
<td>0.7791</td>
<td>0.6826</td>
<td>0.8048</td>
<td>0.8062</td>
<td>0.8098</td>
</tr>
<tr>
<td>(Model I and Model II), small sample size ($n = 15, 25$)</td>
<td>Rank</td>
<td>6</td>
<td>4</td>
<td>5</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>S.D. $L_2$ eff.</td>
<td>0.3598</td>
<td>0.2765</td>
<td>0.3343</td>
<td>0.2480</td>
<td>0.2462</td>
<td>0.2420</td>
</tr>
<tr>
<td></td>
<td>Rank</td>
<td>6</td>
<td>4</td>
<td>5</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>weakly identifiable</td>
<td>Ave. $L_2$ eff.</td>
<td>0.7239</td>
<td>0.7418</td>
<td>0.7771</td>
<td>0.7808</td>
<td>0.7810</td>
<td>0.7817</td>
</tr>
<tr>
<td>(Model I and Model II), large sample size ($n = 100, 500$)</td>
<td>Rank</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>S.D. $L_2$ eff.</td>
<td>0.3096</td>
<td>0.3001</td>
<td>0.2601</td>
<td>0.2563</td>
<td>0.2562</td>
<td>0.2554</td>
</tr>
<tr>
<td></td>
<td>Rank</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>weakly identifiable</td>
<td>Ave. $L_2$ eff.</td>
<td>0.6286</td>
<td>0.7604</td>
<td>0.7299</td>
<td>0.7928</td>
<td>0.7936</td>
<td>0.7958</td>
</tr>
<tr>
<td>(Model I and Model II)</td>
<td>Rank</td>
<td>6</td>
<td>4</td>
<td>5</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>S.D. $L_2$ eff.</td>
<td>0.3347</td>
<td>0.2883</td>
<td>0.2972</td>
<td>0.2522</td>
<td>0.2512</td>
<td>0.2487</td>
</tr>
<tr>
<td></td>
<td>Rank</td>
<td>6</td>
<td>4</td>
<td>5</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

Note: Boldface type indicates the best performance.

5. Conclusions, discussion, and further study

This paper presents the derivations to unify the justifications of the criteria based on Kullback’s symmetric divergence; the Kulback information criterion (KIC) by Cavanaugh [1] and the corrected versions; KICcC by Cavanaugh [2], KICcSB by Seghouane and Bekara [3], and KICcHM by Hafidi and Mkhadri [4]. The results show that KICcC has the strongest penalty function under the condition in (35), followed, respectively, by KICcSB, KICcHM, and KIC. The performance of them is examined by the extensive simulation study relative to the criteria based on Kullback’s directed divergence, AIC and AICc. Our simulation study indicates that, for the small sample size and the true model is somewhat difficult to identify, the performance of the original criteria AIC and KIC is better than their corrected versions. When the sample size is still small but the true model is easily to identify, the corrected versions perform the best. For the large sample size but the true model is very difficult to detect, the AIC family of criteria performs better than the KIC family. When the sample size is still large and the true model can be specified more easily, the KIC family performs the best. This simulation also found that, although the proofs in this study show that the criteria based on Kullback’s symmetric divergence are stronger than the criteria based on the directed divergence, sometimes the performance of them is worse. This result may be because the estimated symmetric measure in (9) contributes to all criteria in KIC family usually having stronger penalty.
functions than the AIC family. This problem causes a greater number of underfitted orders being selected, which then contributes to a low frequency of choosing the correct model. However, when the true model is very difficult to detect, such as Model 1; none of the criteria correctly identify the true model more than 6% of the time. As a result, the frequency of correct order being selected may not be meaningful. For this reason, we have also used the observed $L_2$ efficiency to assess model selection criteria performance. This measure suggests that, in a weakly identifiable true model, whether the sample size is small or large, KICcC is the best criterion because it has highest average value of the observed $L_2$ efficiency and lowest standard deviation. The better performance of KICcC may be because its formula is closer to the expected estimated symmetric discrepancy than other. But, KICcC is more likely to select an underfitted model than other criterion which is because its penalty function is strong. Nevertheless, even if KICcC tends to select underfitted models, these selected models are close to the true model.

In future work, we hope to extend KICcC from Cavanaugh [2] to construct a model selection criterion to serve as an asymptotically unbiased estimator of a variant of Kullback’s symmetric divergence for multivariate regression, seemingly unrelated regression models, and simultaneous equations model. Because, at this time, there exists the multivariate model selection based on the extensions of KICcSB [18] and KICcHM [4], and there exists only the original version of model selection criterion based on the simultaneous equations model [19-20].

**Acknowledgements**

This project is financial supported by The Thailand Research Fund, Office of The Higher Education Commission, and Thaksin University under grant No. MRG5480044. I would like to give special thanks to Prof. Dr. Joseph E. Cavanaugh, Department of Biostatistics, College of Public Health, University of Iowa, USA, Assoc. Prof. Dr. Jirawan Jitthavech, and Assoc. Prof. Dr. Vichit Lorcharachoonkul, National Institute of Development Administration (NIDA), THAILAND for all comments and suggestions.

**References**


