Thailand Statistician
January 2014; 12(1): 37-53
http://statassoc.or.th
Contributed paper

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 KICc_{C}, KICc_{SB}, and KICc_{HM}. 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 KICc_{C} 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 KICc_{SB}, KICc_{HM}, and KIC. This result makes KICc_{C} 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 KICc_{C}, by Seghouane and Bekara [3] called KICc_{SB}, and by Hafidi and Mkhadri [4] called KICc_{HM} 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, KICc_{C}, KICc_{SB}, and KICc_{HM} 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

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}_0,\ \boldsymbol{\varepsilon}_0 \sim \mathrm{N}_n\left(\mathbf{0},\ \sigma_0^2\mathbf{I}_n\right), \tag{1}$$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},\ \boldsymbol{\varepsilon} \sim \mathrm{N}_n\left(\mathbf{0},\ \sigma^2\mathbf{I}_n\right), \tag{2}$$

where $\mathbf{y}$ is an $n\times 1$ dependent random vector of observations, $\mathbf{X}$ is an $n\times p$ matrix of independent variables with full-column rank, $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}$ are $p\times 1$ parameter vectors of regression coefficients, $\boldsymbol{\varepsilon}_0$ and $\boldsymbol{\varepsilon}$ 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 $\boldsymbol{\beta}_0$ has p_0 nonzero entries with $0 < p_0 \le p$ and the rest of the $(p - p_0)$ entries are equal to zero. The $(p+1)\times 1$ vector of parameters is $\boldsymbol{\theta}_0 = \left[\boldsymbol{\beta}_0' \;\; \sigma_0^2\right]'$ and the maximum likelihood estimator of $\boldsymbol{\theta}_0$ is $\hat{\boldsymbol{\theta}} = \left[\hat{\boldsymbol{\beta}}' \;\; \hat{\sigma}^2\right]'$ where

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \text{ and } \hat{\sigma}^2 = \left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)'\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)/n. \tag{3}$$

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

$$-2\log L(\boldsymbol{\theta}|\mathbf{y}) = n\log 2\pi + n\log\sigma^2 + \frac{1}{\sigma^2}\boldsymbol{\varepsilon}_0'\boldsymbol{\varepsilon}_0 + \frac{1}{\sigma^2}(\boldsymbol{\beta}_0 - \boldsymbol{\beta})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta}_0 - \boldsymbol{\beta}) + \frac{2}{\sigma^2}\boldsymbol{\varepsilon}_0'\mathbf{X}(\boldsymbol{\beta}_0 - \boldsymbol{\beta}). \tag{4}$$

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

$$2I(\boldsymbol{\theta}_0, \boldsymbol{\theta}) = E_{\boldsymbol{\theta}_0}\left\{2\log\frac{L(\boldsymbol{\theta}_0|\mathbf{y})}{L(\boldsymbol{\theta}|\mathbf{y})}\right\} = d(\boldsymbol{\theta}_0, \boldsymbol{\theta}) - d(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0),$$

where

$$d(\boldsymbol{\theta}_0, \boldsymbol{\theta}) = E_{\boldsymbol{\theta}_0}\left\{-2\log L(\boldsymbol{\theta}|\mathbf{y})\right\},\; d(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0) = E_{\boldsymbol{\theta}_0}\left\{-2\log L(\boldsymbol{\theta}_0|\mathbf{y})\right\},$$

and the expectation $E_{\boldsymbol{\theta}_0}$ is taken with respect to the true model. Because $d(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0)$ does not depend on $\boldsymbol{\theta}$, any ranking of the candidate models according to $2I(\boldsymbol{\theta}_0, \boldsymbol{\theta})$ would be identical to ranking them according to $d(\boldsymbol{\theta}_0, \boldsymbol{\theta})$. Given a set of maximum likelihood estimators $\hat{\boldsymbol{\theta}}$ in (3), the estimated directed measure $d(\boldsymbol{\theta}_0, \boldsymbol{\theta})$ is

$$d(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}) = E_{\boldsymbol{\theta}_0}\left\{-2\log L(\boldsymbol{\theta}|\mathbf{y})\right\}\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$$

$$= n\log 2\pi + n\log\hat{\sigma}^2 + \frac{n\sigma_0^2}{\hat{\sigma}^2} + \frac{1}{\hat{\sigma}^2}\left(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}\right)'\mathbf{X}'\mathbf{X}\left(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}\right). \tag{5}$$

However, the evaluation in (5) is not possible because it requires the knowledge of $\boldsymbol{\theta}_0$, Akaike [11-12] proposed an asymptotically unbiased estimator of

$$\Delta(\boldsymbol{\theta}_0, p) = E_{\boldsymbol{\theta}_0}\left\{d(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})\right\} \tag{6}$$

as

$$\mathrm{AIC} = n\log\hat{\sigma}^2 + 2(p+1), \tag{7}$$

i.e., $E_{\boldsymbol{\theta}_0}\{\mathrm{AIC}\} + o(1) = \Delta(\boldsymbol{\theta}_0, p)$.

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:

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

i.e., $E_{\boldsymbol{\theta}_0}\{\mathrm{AICc}\} = \Delta(\boldsymbol{\theta}_0, p)$.

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,

$$2J(\boldsymbol{\theta}_0, \boldsymbol{\theta}) = 2I(\boldsymbol{\theta}_0, \boldsymbol{\theta}) + 2I(\boldsymbol{\theta}, \boldsymbol{\theta}_0) = \left[d(\boldsymbol{\theta}_0, \boldsymbol{\theta}) - d(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0)\right] + \left[d(\boldsymbol{\theta}, \boldsymbol{\theta}_0) - d(\boldsymbol{\theta}, \boldsymbol{\theta})\right],$$

where

$$d(\boldsymbol{\theta}_0, \boldsymbol{\theta}) = E_{\boldsymbol{\theta}_0}\left\{-2\log L(\boldsymbol{\theta}|\mathbf{y})\right\},\ d(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0) = E_{\boldsymbol{\theta}_0}\left\{-2\log L(\boldsymbol{\theta}_0|\mathbf{y})\right\},$$

$$d(\boldsymbol{\theta}, \boldsymbol{\theta}_0) = E_{\boldsymbol{\theta}}\left\{-2\log L(\boldsymbol{\theta}_0|\mathbf{y})\right\}, \text{ and } d(\boldsymbol{\theta}, \boldsymbol{\theta}) = E_{\boldsymbol{\theta}}\left\{-2\log L(\boldsymbol{\theta}|\mathbf{y})\right\}.$$

Dropping $d(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0)$, the ranking of the candidate models according to $2J(\boldsymbol{\theta}_0, \boldsymbol{\theta})$ is identical to ranking them according to

$$K(\boldsymbol{\theta}_0, \boldsymbol{\theta}) = d(\boldsymbol{\theta}_0, \boldsymbol{\theta}) + d(\boldsymbol{\theta}, \boldsymbol{\theta}_0) - d(\boldsymbol{\theta}, \boldsymbol{\theta}).$$

Given a set of maximum likelihood estimators $\hat{\boldsymbol{\theta}}$ in (3), the estimated symmetric measure $K(\boldsymbol{\theta}_0, \boldsymbol{\theta})$ is

$$K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}) = d(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}) + d(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) - d(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}}), \tag{9}$$

where $d(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ is exhibited in (5),

$$d(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) = E_{\boldsymbol{\theta}}\left\{-2\log L(\boldsymbol{\theta}_0|\mathbf{y})\right\}\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$$

$$= n\log 2\pi + n\log\sigma_0^2 + \frac{n\hat{\sigma}^2}{\sigma_0^2} + \frac{1}{\sigma_0^2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)'\,\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0), \tag{10}$$

and

$$d\left(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\theta}}\right) = E_{\boldsymbol{\theta}}\left\{-2\log L\left(\boldsymbol{\theta}\,|\,\mathbf{y}\right)\right\}\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = n\log 2\pi + n\log\hat{\sigma}^2 + n. \tag{11}$$

Yet, evaluating $K\left(\boldsymbol{\theta}_0,\hat{\boldsymbol{\theta}}\right)$ in (9) requires $\boldsymbol{\theta}_0$, Cavanaugh [1] proposed an asymptotically unbiased estimator of

$$\Omega\left(\boldsymbol{\theta}_0, p\right) = E_{\boldsymbol{\theta}_0}\left\{K\left(\boldsymbol{\theta}_0,\hat{\boldsymbol{\theta}}\right)\right\} \tag{12}$$

as

$$\mathrm{KIC} = n\log\hat{\sigma}^2 + 3\left(p+1\right), \tag{13}$$

i.e., $E_{\boldsymbol{\theta}_0}\left\{\mathrm{KIC}\right\} + o(1) = \Omega\left(\boldsymbol{\theta}_0, p\right).$

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:

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

i.e., $E_{\boldsymbol{\theta}_0}\left\{\mathrm{KICc}\right\} = \Omega\left(\boldsymbol{\theta}_0, p\right).$

Because the phi (ψ) 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 $\mathrm{KICc_C}$, $\mathrm{KICc_{SB}}$, and $\mathrm{KICc_{HM}}$,

$$\mathrm{KICc_C} = n\log\hat{\sigma}^2 + n\log\left(\frac{n}{n-p}\right) + \frac{n\left[\left(n-p\right)\left(2p+3\right)-2\right]}{\left(n-p-2\right)\left(n-p\right)}, \tag{14}$$

$$\mathrm{KICc_{SB}} = n\log\hat{\sigma}^2 + \frac{\left(p+1\right)\left(3n-p-2\right)}{n-p-2} + \frac{p}{n-p}, \tag{15}$$

$$\mathrm{KICc_{HM}} = n\log\hat{\sigma}^2 + \frac{\left(p+1\right)\left(3n-p-2\right)}{n-p-2}. \tag{16}$$

3. The unified derivations of KIC and KICc

To begin the unification of the derivations KIC, $\mathrm{KICc_C}$, $\mathrm{KICc_{SB}}$, and $\mathrm{KICc_{HM}}$, we consider the expectation of the discrepancies in (5), (10), and (11) with respect to the true model [3],

$$E_{\boldsymbol{\theta}_0}\left\{d\left(\boldsymbol{\theta}_0,\hat{\boldsymbol{\theta}}\right)\right\} = n\log 2\pi + E_{\boldsymbol{\theta}_0}\left\{n\log\hat{\sigma}^2\right\} + E_{\boldsymbol{\theta}_0}\left\{\frac{n\sigma_0^2}{\hat{\sigma}^2}\right\} + E_{\boldsymbol{\theta}_0}\left\{\frac{1}{\hat{\sigma}^2}\left(\boldsymbol{\beta}_0-\hat{\boldsymbol{\beta}}\right)'\mathbf{X}'\mathbf{X}\left(\boldsymbol{\beta}_0-\hat{\boldsymbol{\beta}}\right)\right\}, \tag{17}$$

$$E_{\boldsymbol{\theta}_0}\left\{ d\left(\hat{\boldsymbol{\theta}},\boldsymbol{\theta}_0\right)\right\} = n\log 2\pi + n\log\sigma_0^2 + E_{\boldsymbol{\theta}_0}\left\{\frac{n\hat{\sigma}^2}{\sigma_0^2}\right\} + E_{\boldsymbol{\theta}_0}\left\{\frac{1}{\sigma_0^2}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\right)' \mathbf{X}'\mathbf{X}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\right)\right\}, \tag{18}$$

$$E_{\boldsymbol{\theta}_0}\left\{ d\left(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\theta}}\right)\right\} = n\log 2\pi + E_{\boldsymbol{\theta}_0}\left\{n\log\hat{\sigma}^2\right\} + n. \tag{19}$$

From the fact that the terms $\frac{n\hat{\sigma}^2}{\sigma_0^2}$ and $\frac{1}{\sigma_0^2}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\right)' \mathbf{X}'\mathbf{X}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\right)$ are the independent χ^2 distributions with the degrees of freedom which are, respectively, $n - p$ and p, we have

$$E_{\boldsymbol{\theta}_0}\left\{\frac{n\hat{\sigma}^2}{\sigma_0^2}\right\} = n-p \text{ and } E_{\boldsymbol{\theta}_0}\left\{\frac{1}{\sigma_0^2}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\right)' \mathbf{X}'\mathbf{X}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\right)\right\} = p\,. \tag{20}$$

Using the facts in (20), we have

$$E_{\boldsymbol{\theta}_0}\left\{\frac{n\sigma_0^2}{\hat{\sigma}^2}\right\} = E_{\boldsymbol{\theta}_0}\left\{\frac{n^2}{n\hat{\sigma}^2/\sigma_0^2}\right\} = \frac{n^2}{n-p-2}$$

and

$$E_{\boldsymbol{\theta}_0}\left\{\frac{1}{\hat{\sigma}^2}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\right)' \mathbf{X}'\mathbf{X}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\right)\right\} = \frac{1}{n}E_{\boldsymbol{\theta}_0}\left\{\frac{n\sigma_0^2}{\hat{\sigma}^2}\right\}E_{\boldsymbol{\theta}_0}\left\{\frac{1}{\sigma_0^2}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\right)' \mathbf{X}'\mathbf{X}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\right)\right\} = \frac{np}{n-p-2}. \tag{21}$$

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

$$\begin{aligned}\Delta\left(\boldsymbol{\theta}_0, p\right) &= E_{\boldsymbol{\theta}_0}\left\{ d\left(\boldsymbol{\theta}_0,\hat{\boldsymbol{\theta}}\right)\right\} = n\log 2\pi + E_{\boldsymbol{\theta}_0}\left\{n\log\hat{\sigma}^2\right\} + \frac{n^2}{n-p-2} + \frac{np}{n-p-2} \\ &= n\left(\log 2\pi + 1\right) + E_{\boldsymbol{\theta}_0}\left\{\text{AICc}\right\},\end{aligned} \tag{22}$$

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

$$E_{\boldsymbol{\theta}_0}\left\{ d\left(\hat{\boldsymbol{\theta}},\boldsymbol{\theta}_0\right)\right\} = n\left(\log 2\pi + 1\right) + n\log\sigma_0^2. \tag{23}$$

Using the results in (19), (22), and (23), the expected value of $K\left(\boldsymbol{\theta}_0,\hat{\boldsymbol{\theta}}\right)$ in (9) becomes

$$\Omega\left(\boldsymbol{\theta}_0, p\right) = E_{\boldsymbol{\theta}_0}\left\{ K\left(\boldsymbol{\theta}_0,\hat{\boldsymbol{\theta}}\right)\right\} = n\left(\log 2\pi + 1\right) + E_{\boldsymbol{\theta}_0}\left\{\text{AICc}\right\} - E_{\boldsymbol{\theta}_0}\left\{n\log\frac{\hat{\sigma}^2}{\sigma_0^2}\right\}, \tag{24}$$

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 $K\left(\boldsymbol{\theta}_0,\hat{\boldsymbol{\theta}}\right)$ 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_{C}, KICc_{SB}, and KICc_{HM}, we give the following lemmas.

Lemma 1. $$-E_{\boldsymbol{\theta}_0}\left\{n\log\frac{\hat{\sigma}^2}{\sigma_0^2}\right\}=-n\log\left(\frac{n-p}{2}\right)+\frac{n}{n-p}+n\log\left(\frac{n}{2}\right)+o\left(\frac{n}{(n-p)^2}\right). \quad (25)$$

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

$$E_{\boldsymbol{\theta}_0}\left\{\log\chi^2_{df}\right\}=\psi\left(\frac{df}{2}\right)+\log 2 \text{ and } \psi(x)=\log x-\frac{1}{2x}+\text{o}\left(\frac{1}{x^2}\right) \text{ as } x\to\infty. \quad (26)$$

Applying the fact $E_{\boldsymbol{\theta}_0}\left\{n\hat{\sigma}^2/\sigma_0^2\right\}=n-p$ in (20) and the facts in (26), we have

$$\begin{aligned}-E_{\boldsymbol{\theta}_0}\left\{n\log\frac{\hat{\sigma}^2}{\sigma_0^2}\right\} &= -E_{\boldsymbol{\theta}_0}\left\{n\log\frac{n\hat{\sigma}^2}{\sigma_0^2}\right\}+n\log n=-n\left[\psi\left(\frac{n-p}{2}\right)+\log 2\right]+n\log n\\ &=-n\left[\log\left(\frac{n-p}{2}\right)-\frac{1}{n-p}+o\left(\frac{1}{(n-p)^2}\right)\right]-n\log 2+n\log n\\ &=-n\log\left(\frac{n-p}{2}\right)+\frac{n}{n-p}+n\log\left(\frac{n}{2}\right)+o\left(\frac{n}{(n-p)^2}\right).\end{aligned}$$

Lemma 2.

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

Proof. Applying the first-order Taylor's series expansion to expand the term

$$\log\left((n-p)/2\right)$$

about $n/2$, i.e.,

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

to obtain the approximation in (27).

Lemma 3. $$p+\frac{n}{n-p}+o\left(\frac{p^2}{n}\right)+o\left(\frac{n}{(n-p)^2}\right)=(p+1)+o(1). \quad (28)$$

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).

Appling Lemma 1 into $\Omega(\boldsymbol{\theta}_0, p)$ in (24), we obtain

$$\Omega(\boldsymbol{\theta}_0, p) = n(\log 2\pi+1)+E_{\boldsymbol{\theta}_0}\{\mathrm{AICc}\}-n\log\left(\frac{n-p}{2}\right)+\frac{n}{n-p}+n\log\left(\frac{n}{2}\right)+o\left(\frac{n}{(n-p)^2}\right)$$

$$= n(\log 2\pi+1)+E_{\boldsymbol{\theta}_0}\left\{\mathrm{KICc}_{\mathrm{C}}+o\left(\frac{n}{(n-p)^2}\right)\right\},$$

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

Appling Lemmas 1 and 2 into $\Omega(\boldsymbol{\theta}_0, p)$ in (24), we obtain

$$\Omega(\boldsymbol{\theta}_0, p) = n(\log 2\pi+1)+E_{\boldsymbol{\theta}_0}\{\mathrm{AICc}\}+p+\frac{n}{n-p}+o\left(\frac{p^2}{n}\right)+o\left(\frac{n}{(n-p)^2}\right)$$

$$= n(\log 2\pi+1)+E_{\boldsymbol{\theta}_0}\left\{\mathrm{KICc}_{\mathrm{SB}}+o\left(\frac{p^2}{n}\right)+o\left(\frac{n}{(n-p)^2}\right)\right\},$$

where $\mathrm{KICc}_{\mathrm{SB}}$ is the corrected version of KIC from Seghouane and Bekara [3] that was exhibited in (15).

Appling Lemmas 1, 2, and 3 into $\Omega(\boldsymbol{\theta}_0, p)$ in (24), we obtain

$$\Omega(\boldsymbol{\theta}_0, p) = n(\log 2\pi+1)+E_{\boldsymbol{\theta}_0}\{\mathrm{AICc}\}+(p+1)+o(1)$$

$$= n(\log 2\pi+1)+E_{\boldsymbol{\theta}_0}\{\mathrm{KICc}_{\mathrm{HM}}+o(1)\},$$

where $\mathrm{KICc}_{\mathrm{HM}}$ is the corrected version of KIC from Hafidi and Mkhadri [4] that was exhibited in (16).

The connections of KIC, $\mathrm{KICc}_{\mathrm{HM}}$, $\mathrm{KICc}_{\mathrm{SB}}$, and $\mathrm{KICc}_{\mathrm{C}}$ are given by

$$\mathrm{KICc}_{\mathrm{HM}} = \mathrm{KIC} + \frac{2(p+1)(p+2)}{n-p-2}, \qquad (29)$$

$$KICc_{SB} = KICc_{HM} + \frac{p}{n-p}, \tag{30}$$

$$KICc_{C} = KICc_{SB} + n\log\left(\frac{n}{n-p}\right) - p. \tag{31}$$

From the connection in (29), we found that the term

$$\frac{2(p+1)(p+2)}{n-p-2} \tag{32}$$

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),

$$\frac{p}{n-p} \tag{33}$$

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 $KICc_{SB} \geq KICc_{HM} \geq KIC$. While the term

$$n\log\left(\frac{n}{n-p}\right) - p \tag{34}$$

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

$$(1-p/n)\exp(p/n) < 1 \tag{35}$$

is true, we have the term in (34) being greater than zero. This means that the penalty function of $KICc_C$ is stronger than other criteria, $KICc_{SB}$, $KICc_{HM}$, and KIC, under the condition in (35). The strong penalty may cause $KICc_C$ 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 σ_0^2 and the distribution of error random variable $\boldsymbol{\varepsilon}_0$ in (1) were changed. As a result, we have taken $\boldsymbol{\varepsilon}_0$ to be normally distributed with zero mean and σ_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 $\mathbf{X}$ 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 $\mathbf{I}_{10}$. The candidate models include the columns of $\mathbf{X}$ 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), KICc$_C$ in (14), KICc$_{SB}$ in (15), and KICc$_{HM}$ 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]

$$L_2(p) = \frac{1}{n}\left(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}\right)' \mathbf{X}'\mathbf{X}\left(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}\right).$$

Observed L_2 efficiency is defined by the ratio

$$\text{Observed } L_2 \text{ efficiency } = \frac{\min_{1\le p\le P} L_2(p)}{L_2(p_s)},$$

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 KICc_C 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.

Model	*n*	Order	Criteria							
			AIC	**AICc**	**KIC**	**KICcHM**	**KICcSB**	**KICcC**	***d***	***K***
I	15	Underfitted	596	1000	837	1000	1000	1000	982	986
very		Correct	**54**	0	26	0	0	0	0	0
weakly		Overfitted	350	0	137	0	0	0	18	14
identifiable	25	Underfitted	859	998	972	1000	1000	1000	987	992
(true order		Correct	**39**	1	11	0	0	0	0	0
$p_0 = 8$)		Overfitted	102	1	17	0	0	0	13	8
	100	Underfitted	944	974	993	998	999	999	998	998
		Correct	**23**	14	5	2	1	1	0	0
		Overfitted	33	12	2	0	0	0	2	2
	500	Underfitted	958	962	998	998	998	999	1000	1000
		Correct	**21**	**21**	1	1	1	0	0	0
		Overfitted	21	17	1	1	1	1	0	0
II	15	Underfitted	284	820	542	859	864	875	577	547
weakly		Correct	132	123	**148**	111	109	105	423	453
identifiable		Overfitted	584	57	310	30	27	20	0	0
(true order	25	Underfitted	374	653	575	716	720	727	368	344
$p_0 = 3$)		Correct	244	235	**264**	235	231	226	631	655
		Overfitted	382	112	161	49	49	47	1	1
	100	Underfitted	109	133	214	230	231	232	34	26
		Correct	609	642	676	677	678	**680**	966	974
		Overfitted	282	225	110	93	91	88	0	0
	500	Underfitted	0	0	0	0	0	0	0	0
		Correct	737	751	890	895	895	**896**	1000	1000
		Overfitted	263	249	110	105	105	104	0	0

Table 1. (Continued).

Model	*n*	Order	Criteria							
			AIC	**AICc**	**KIC**	**KICcHM**	**KICcSB**	**KICcC**	***d***	***K***
III	15	Underfitted	0	0	0	0	0	0	30	0
very		Correct	325	939	583	964	964	**968**	970	1000
strongly		Overfitted	675	61	417	36	36	32	0	0
identifiable	25	Underfitted	0	0	0	0	0	0	0	0
(true order		Correct	549	855	749	899	904	**920**	1000	1000
$p_0 = 4$)		Overfitted	451	145	251	101	96	80	0	0
	100	Underfitted	0	0	0	0	0	0	0	0
		Correct	687	756	855	874	874	**878**	1000	1000
		Overfitted	313	244	145	126	126	122	0	0
	500	Underfitted	0	0	0	0	0	0	0	0
		Correct	713	731	885	**889**	**889**	**889**	1000	1000
		Overfitted	287	269	115	111	111	111	0	0
IV	15	Underfitted	36	887	94	943	955	969	724	554
strongly		Correct	444	113	**532**	57	45	31	214	377
identifiable		Overfitted	520	0	374	0	0	0	62	69
(true order	25	Underfitted	5	31	9	60	62	67	281	133
$p_0 = 8$)		Correct	710	**950**	840	928	928	925	663	846
		Overfitted	285	19	151	12	10	8	56	21
	100	Underfitted	0	0	0	0	0	0	0	0
		Correct	815	882	925	950	950	**953**	1000	1000
		Overfitted	185	118	75	50	50	47	0	0
	500	Underfitted	0	0	0	0	0	0	0	0
		Correct	854	864	951	**956**	**956**	**956**	1000	1000
		Overfitted	146	136	49	44	44	44	0	0

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.

Circumstance	Stat.	Criteria					
		AIC	**AICc**	**KIC**	**KICc$_{HM}$**	**KICc$_{SB}$**	**KICc$_{C}$**
weakly identifiable (Model I and Model II), small sample size (n = 15, 25)	Ave. L_2 eff.	0.5332	0.7791	0.6826	0.8048	0.8062	**0.8098**
	Rank	6	4	5	3	2	**1**
	S.D. L_2 eff.	0.3598	0.2765	0.3343	0.2480	0.2462	**0.2420**
	Rank	6	4	5	3	2	**1**
weakly identifiable (Model I and Model II), large sample size (n = 100, 500)	Ave. L_2 eff.	0.7239	0.7418	0.7771	0.7808	0.7810	**0.7817**
	Rank	6	5	4	3	2	**1**
	S.D. L_2 eff.	0.3096	0.3001	0.2601	0.2563	0.2562	**0.2554**
	Rank	6	5	4	3	2	**1**
weakly identifiable (Model I and Model II)	Ave. L_2 eff.	0.6286	0.7604	0.7299	0.7928	0.7936	**0.7958**
	Rank	6	4	5	3	2	**1**
	S.D. L_2 eff.	0.3347	0.2883	0.2972	0.2522	0.2512	**0.2487**
	Rank	6	4	5	3	2	**1**

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; KICc$_C$ by Cavanaugh [2], KICc$_{SB}$ by Seghouane and Bekara [3], and KICc$_{HM}$ by Hafidi and Mkhadri [4]. The results show that KICc$_C$ has the strongest penalty function under the condition in (35), followed, respectively, by KICc$_{SB}$, KICc$_{HM}$, 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, $KICc_C$ is the best criterion because it has highest average value of the observed L_2 efficiency and lowest standard deviation. The better performance of $KICc_C$ may be because its formula is closer to the expected estimated symmetric discrepancy than other. But, $KICc_C$ is more likely to select an underfitted model than other criterion which is because its penalty function is strong. Nevertheless, even if $KICc_C$ tends to select underfitted models, these selected models are close to the true model.

In future work, we hope to extend $KICc_C$ 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 $KICc_{SB}$ [18] and $KICc_{HM}$ [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 Lorchirachoonkul, National Institute of Development Administration (NIDA), THAILAND for all comments and suggestions.

References

[1] Cavanaugh, J.E., A large-sample model selection criterion based on Kullback's symmetric divergence, *Statistics & Probability Letters*, 1999; 42: 333-343.

[2] Cavanaugh, J.E., Criteria for linear model selection based on Kullback's symmetric divergence, *Australian & New Zealand Journal of Statistics,* 2004; 46: 257-274.

[3] Seghouane, A.K., and Bekara, M., A small sample model selection criterion based on Kullback's symmetric divergence, *IEEE Transactions on Signal Processing*, 2004; 52: 3314-3323.

[4] Hafidi, B., and Mkhadri, A., A corrected Akaike criterion based on Kullback's symmetric divergence: Applications in Time Series, Multiple and Multivariate Regression, *Computational Statistics & Data Analysis*, 2006; 50: 1524-1550.

[5] Hurvich, C.M., and Tsai, C.L., Regression and time series model selection in small samples, *Biometrika*, 1989; 76: 297-307.

[6] Bedrick, E.J., and Tsai, C.L., Model selection for multivariate regression in small samples, *Biometrics*, 1994; 50: 226-231.

[7] Cavanaugh, J.E., Unifying the derivation of Akaike and corrected information criteria, *Statistics & Probability Letters*, 1997; 33: 201-208.

[8] McQuarrie, A.D., A small-sample correction for the Schwarz SIC model selection criterion, *Statistics & Probability Letters*, 1999; 44: 79-86.

[9] Hafidi, B., A small-sample criterion based on Kullback's symmetric divergence for vector autoregressive modeling, *Statistics & Probability Letters*, 2006; 76: 1647-1654.

[10] Kullback, S., *Information Theory and Statistics*, New York, Dover, 1968.

[11] Akaike, H., Information theory and an extension of the maximum likelihood principle, In *2nd International Symposium on Information Theory*, B.N. Petrov and F. Csaki, eds. Akademiai Kiado, Budapest, 1973: 267-281.

[12] Akaike, H., A new look at the statistical model identification, *IEEE Transactions on Automatic Control*, 1974; 19: 716-723.

[13] Seghouane, A.K., A note on overfitting properties of KIC and KICc, *Signal Process* 2006; 86: 3055-3060.

[14] McQuarrie, A.D., and Tsai, C.L., Regression and time series model selection, Singapore, World Scientific, 1998.

[15] Bernardo, J.M., Psi (digamma) function, *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 1976; 25: 315-317.

[16] Kundu, D., and Murali G., Model selection in linear regression, *Computational Statistics & Data Analysis*, 1996; 22: 461-469.

[17] McQuarrie, A.D., Shumway, R., and Tsai, C.L., The model selection criterion AICu, *Statistics & Probability Letters*, 1997; 34: 285-292.

[18] Seghouane, A.K., Multivariate regression model selection from small samples using Kullback's symmetric divergence, *Signal Processing*, 2006; 86: 2074-2084.

[19] Keerativibool, W., Jitthavech, J., and Lorchirachoonkul, V., Model selection in a system of simultaneous equations model, *Communications in Statistics-Theory and Methods*, 2011; 40: 373-393.

[20] Keerativibool, W., New criteria for selection in simultaneous equations model, *Thailand Statistician: Journal of Thai Statistical Association*, 2012; 10: 163-181.