A SMALL SAMPLE MODEL SELECTION CRITERION BASED ON KULLBACK'S SYMMETRIC DIVERGENCE

Abd-Krim Seghouane, Maiza Bekara [*] *and Gilles Fleury*

Service des Mesures - SUPELEC
3, rue Joliot Curie
91192 Gif-sur-Yvette, Cedex France
firstame.lastname@supelec.fr

ABSTRACT

The Kullback information criterion KIC is a recently developed tool for statistical model selection [1]. KIC serves as an *asymptotically* unbiased estimator of a variant of the Kullback symmetric divergence, known also as J-divergence. In this paper a bias correction of the Kullback symmetric information criterion is derived for linear models. The correction is of particular use when the sample size is small or when the number of fitted parameters is of moderate to large fraction of the sample size. For linear regression models, the corrected method called KICc is an *exactly* unbiased estimator of a variant of the Kullback symmetric divergence between the true unknown model and the candidate fitted model. Furthermore KICc is found to provide better model order choice than any other asymptotically efficient methods in an application to autoregressive time series models.

1. INTRODUCTION

Model selection is an important area of statistical modeling, and its results are applied to many problems in signal processing [2].
the first model selection criterion to gain widespread acceptance was Akaike information criterion, AIC [3]. Many others criteria have been introduced and studied, including the cross validation, CV, by Stone [4], Bayesian information criterion, BIC, by Schwarz [5], minimum description length, MDL, by Rissanen [6].
AIC serves as an *asymptotically* unbiased estimator of the Kullback's directed divergence between the true model and the fitted approximating model. The directed divergence, also known as the Kullback-Leibler information, relative entropy or the I-divergence, assesses the dissimilarity between two statistical models [7]. As the dimension of the candidate model, increases compared to n, the sample size, AIC becomes a strongly negatively biased estimate of the information. A bias corrected version of AIC was proposed by Sugiura for linear regression models [8]. Later it was successfully applied by Hurvich to nonlinear regression and autoregressive time series [9].
The Kullback's divergence is an asymmetric measure, it means that an alternative directed divergence can be obtained by reversing the roles of the two models in the definition of the measure. A new measure of model's dissimilarity can be obtained by the sum of the two directed divergences, known as the Kullback's symmetric divergence, or J-divergence [10]. Since the symmetric divergence combines information about models dissimilarity through distinct measures, it functions as a gauge of model disparity, which is arguably more sensitive than either of its individual component. Following the above reasoning, Cavanaugh [1] proposed the Kullback information criterion KIC as an asymptotically unbiased estimate of a variant (within a constant) of the J-divergence between the true unknown model and the fitted approximating model.
Motivated by the above developments, we propose a bias corrected version of the KIC for linear regression models. The new criterion is shown to outperform classical criteria in a small sample autoregressive modeling.
The remainder of this paper is organized as follows. In section 2 we present a short overview of Kullback's directed divergence, AIC, its corrected version AICc and KIC. In section 3 we introduce the bias corrected version of KIC. Section 4 presents simulation results for autoregressive model selection. We end up by concluding remarks.

2. REVIEW OF AIC, AICC AND KIC

Suppose a collection of data $\mathbf{y} = (y_1, \dots, y_2)$ has been generated according to an unknown parametric model $p(\mathbf{y}|\theta_0)$. We consider to find a parametric model which provides a suitable approximation for $p(\mathbf{y}|\theta_0)$.
let $\mathcal{M}_k = \{p(\mathbf{y}|\theta_k)|\theta_k \in \Theta_k\}$ denote a k-dimensional parametric family and let $\hat{\theta}_k$ denote the vector of estimate obtained by maximizing the likelihood function $p(\mathbf{y}|\theta_k)$ over Θ_k. For simplicity, we will assume $k = 1, 2, \dots, k_{max}$, so the collection consists of families of dimension 1 through k_{max} [11].
To determine which candidate model best approximates the true unknown model $p(\mathbf{y}|\theta_0)$, we require a measure which provide a suitable reflection of the disparity between $p(\mathbf{y}|\theta_0)$ and an approximating model $p(\mathbf{y}|\theta_k)$. The Kullback's directed divergence is one of such measure.
For the two parametric densities $p(\mathbf{y}|\theta_k)$ and $p(\mathbf{y}|\theta_0)$, the *Kullback's divergence* between $p(\mathbf{y}|\theta_k)$ and $p(\mathbf{y}|\theta_0)$ with respect to $p(\mathbf{y}|\theta_0)$ is defined as

$$\begin{aligned} 2I_n(\theta_0, \theta_k) &= E_{\theta_0}\left\{2\ln\frac{p(\mathbf{y}|\theta_0)}{p(\mathbf{y}|\theta_k)}\right\} \\ &= E_{\theta_0}\{-2\ln p(\mathbf{y}|\theta_k)\} - E_{\theta_0}\{-2\ln p(\mathbf{y}|\theta_0)\} \\ &= d_n(\theta_0, \theta_k) - d_n(\theta_0, \theta_0). \end{aligned}$$

where

$$d_n(\theta_0, \theta_k) = E_{\theta_0}\{-2\ln p(\mathbf{y}|\theta_k)\}. \tag{1}$$

[*] Corresponding author

Since $d_n(\theta_0, \theta_0)$ does not depend on θ_k, any ranking of the candidate models according to $I_n(\theta_0, \theta_k)$ would be identical to ranking them according to $d_n(\theta_0, \theta_k)$.
The above discussion suggests that

$$d_n(\theta_0, \hat{\theta}_k) = E_{\theta_0} \left\{-2 \ln p(\mathbf{y}|\theta_k)\right\} |_{\theta_k = \hat{\theta}_k}$$

would provide a suitable measure of a variant of the directed Kullback divergence between the generating model $p(\mathbf{y}|\theta_0)$ and the candidate model $p(\mathbf{y}|\hat{\theta}_k)$. Yet evaluating $d_n(\theta_0, \hat{\theta}_k)$ is not possible, since doing so requires the knowledge of θ_0.
Akaike suggests that $-2 \ln p(\mathbf{y}|\hat{\theta}_k)$ serves as a biased estimator of $d_n(\theta_0, \hat{\theta}_k)$ and proposes an *asymptotic* bias correction [3] leading to

$$AIC = -2 \ln p(\mathbf{y}|\hat{\theta}_k) + 2k \tag{2}$$

if we denote

$$\Delta_n(k, \theta_0) = E_{\theta_0} \left\{ d_n(\theta_0, \hat{\theta}_k) \right\}.$$

One can establish that [3]

$$\Delta_n(k, \theta_0) = E_{\theta_0} \left\{AIC\right\} + o(1).$$

Hurvich proposes a corrected version [9]

$$AICc = -2 \ln p(\mathbf{y}|\hat{\theta}_k) + 2\frac{(k+1)n}{n-k-2}. \tag{3}$$

that is an exactly unbiased estimator of $d_n(\theta_0, \hat{\theta}_k)$ for linear regression, i.e

$$\Delta_n(k, \theta_0) = E_{\theta_0} \left\{AICc\right\}.$$

Recall that the symmetric divergence is defined as

$$\begin{aligned} 2J_n(\theta_0, \theta_k) &= I_n(\theta_0, \theta_k) + I_n(\theta_k, \theta_0) \\ &= d_n(\theta_0, \theta_k) - d_n(\theta_0, \theta_0) + d_n(\theta_k, \theta_0) \\ &\quad -d_n(\theta_k, \theta_k) \end{aligned}$$

dropping $d_n(\theta_0, \theta_0)$ since it does not depend on k, the quantity

$$K_n(\theta_0, \theta_k) = d_n(\theta_0, \theta_k) + d_n(\theta_k, \theta_0) - d_n(\theta_k, \theta_k) \tag{4}$$

is a substitute measure for $J_n(\theta_0, \theta_k)$. $K_n(\theta_0, \theta_k)$ would lead to an appealing measure of separation between the true unknown model and the fitted candidate model because it includes an additional information about the model's dissimilarity .
Since $K_n(\theta_0, \hat{\theta}_k)$ is inaccessible, Cavanaugh proposes an asymptotically unbiased estimator [1]

$$KIC = -2 \ln p(\mathbf{y}|\hat{\theta}_k) + 3k \tag{5}$$

such that

$$\begin{aligned} \Omega_n(k, \theta_0) &= E_{\theta_0} \left\{ K_n(\theta_0, \hat{\theta}_k) \right\} \\ &= E_{\theta_0} \left\{KIC\right\} + o(1). \end{aligned}$$

KIC is shown to outperform AIC in large sample autoregressive model selection and produces less overfitting selection than AIC [1].

3. DERIVATION OF KICC

Consider the case of linear regression models. Suppose that the generating model for the data is given by

$$\mathbf{y} = X\beta_0 + e, \qquad e \sim N(0, \sigma_0^2 I_n) \tag{6}$$

and that of the candidate model

$$\mathbf{y} = X\beta + e, \qquad e \sim N(0, \sigma^2 I_n) \tag{7}$$

The vector of parameters for the model is $\theta = [\beta \quad \sigma^2]^t$. In what follows, we propose an *exactly* unbiased estimator of $K_n(\theta_0, \hat{\theta}_k)$ for linear regression

Proposition. *Let*

$$\begin{aligned} KICc &= -2 \ln p(\mathbf{y}|\hat{\theta}_k) + 2\frac{(k+1)n}{n-k-2} \\ &\quad -n\psi\left(\frac{n-k}{2}\right) + n \ln \frac{n}{2}. \end{aligned} \tag{8}$$

then KICc is an exactly unbiased estimator of $K_n(\theta_0, \hat{\theta}_k)$,

$$\text{that is :} \qquad \Omega_n(k, \theta_0) = E_{\theta_0} \left\{KICc\right\}$$

where ψ is digamma or psi function defined as [12]:

$$\psi(x) = \frac{d\{\ln \Gamma(x)\}}{dx} = \frac{\Gamma^{'}(x)}{\Gamma(x)}$$

Proof.
We consider the candidate model of Equation (7) then

$$\begin{aligned} d_n(\theta_1, \theta_2) &= E_{\theta_1} \left\{-2 \ln p(\mathbf{y}|\theta_2)\right\} \\ &= n \ln 2\pi + n \ln \sigma_2^2 + n\frac{\sigma_1^2}{\sigma_2^2} \\ &\quad + \frac{1}{\sigma_2^2} \left(\beta_1 - \beta_2\right)^t X^t X \left(\beta_1 - \beta_2\right). \end{aligned}$$

Replacing the above in what it corresponds in Equation (4)

$$\begin{aligned} K_n(\theta_0, \hat{\theta}_k) &= n \ln \hat{\sigma}_k^2 + n \ln 2\pi - n \ln \frac{\hat{\sigma}_k^2}{\sigma_0^2} - n \\ &\quad + \frac{1}{\sigma_0^2} \left(\hat{\beta}_k - \beta_0\right)^t X^t X \left(\hat{\beta}_k - \beta_0\right) \\ &\quad + \frac{1}{\hat{\sigma}_k^2} \left(\hat{\beta}_k - \beta_0\right)^t X^t X \left(\hat{\beta}_k - \beta_0\right) \\ &\quad + n\frac{\sigma_0^2}{\hat{\sigma}_k^2} + n\frac{\hat{\sigma}_k^2}{\sigma_0^2} \end{aligned}$$

We have the following results:

- the term $\frac{n\hat{\sigma}_k^2}{\sigma_0^2}$ has a χ^2 distribution with $n-k$ degree of freedom [13]

$$E_{\theta_0} \left\{ \frac{n\hat{\sigma}_k^2}{\sigma_0^2} \right\} = n - k,$$

- the term $\frac{1}{\sigma_0^2} \left(\hat{\beta}_k - \beta_0\right)^t X^t X \left(\hat{\beta}_k - \beta_0\right)$ has a χ^2 distribution with k degree of freedom [9]

$$E_{\theta_0} \left\{ \frac{1}{\sigma_0^2} \left(\hat{\beta}_k - \beta_0\right)^t X^t X \left(\hat{\beta}_k - \beta_0\right) \right\} = k,$$

- $\hat{\sigma}_k^2$ and $\left(\hat{\beta}_k - \beta_0\right)^t X^t X \left(\hat{\beta}_k - \beta_0\right)$ are independent random variables [9],

$$E_{\theta_0}\left\{\frac{1}{\hat{\sigma}_k^2}\left(\hat{\beta}_k - \beta_0\right)^t X^t X \left(\hat{\beta}_k - \beta_0\right)\right\} = \frac{nk}{n-k-2},$$

- if μ is a χ^2 random variable with f degree of freedom, then the expectation of $\frac{1}{\mu}$ is given by $\frac{1}{f-2}$,

$$E_{\theta_0}\left\{n\frac{\sigma_0^2}{\hat{\sigma}_k^2}\right\} = \frac{n^2}{n-k-2}.$$

Using the above results, it is straightforward to show

$$\begin{aligned}
\Omega_n(k,\theta_0) &= E_{\theta_0}\left\{n\ln\hat{\sigma}_k^2 + n\ln 2\pi\right\} + \frac{n^2}{n-k-2} + k \\
&\quad n - k + \frac{nk}{n-k-2} - E_{\theta_0}\left\{n\ln\frac{\hat{\sigma}_k^2}{\sigma_0^2}\right\} - n \\
&= E_{\theta_0}\left\{n\ln\hat{\sigma}_k^2 + n\ln 2\pi + n\right\} + 2\frac{(k+1)n}{n-k-2} \\
&\quad -E_{\theta_0}\left\{n\ln\frac{\hat{\sigma}_k^2}{\sigma_0^2}\right\} \\
&= E_{\theta_0}\left\{-2\ln p(\mathbf{y}|\hat{\theta}_k)\right\} + 2\frac{(k+1)n}{n-k-2} \\
&\quad -nE_{\theta_0}\left\{\ln n\frac{\hat{\sigma}_k^2}{\sigma_0^2}\right\} + n\ln n
\end{aligned}$$

after Koltz [14] we have

$$E_{\theta_0}\left\{\ln n\frac{\hat{\sigma}_k^2}{\sigma_0^2}\right\} = \psi\left(\frac{n-k}{2}\right) + \ln 2$$

then

$$\begin{aligned}
\Omega_n(k,\theta_0) &= E_{\theta_0}\left\{-2\ln p(\mathbf{y}|\hat{\theta}_k)\right\} + 2\frac{(k+1)n}{n-k-2} \\
&\quad -n\psi\left(\frac{n-k}{2}\right) + n\ln\frac{n}{2} \\
&= E_{\theta_0}\left\{KICc\right\}
\end{aligned}$$

For ease of computation, it is possible to approximate the digamma function using the following [12]

$$\psi(x) \simeq \ln x - \frac{1}{2x} + o\left(\frac{1}{x^2}\right)$$

$$\psi\left(\frac{n-k}{2}\right) \simeq \ln\left(\frac{n-k}{2}\right) - \frac{1}{n-k} + o\left(\frac{1}{(n-k)^2}\right)$$

A second order Taylor expansion of $\ln(n-k)$ leads to

$$\psi\left(\frac{n-k}{2}\right) \simeq \ln\left(\frac{n}{2}\right) - \frac{k}{n} - \frac{1}{n-k} + o\left(\frac{1}{(n-k)^2}\right)$$
$$+o\left(\left(\frac{k}{n}\right)^2\right)$$

an approximate KICc is given by

$$\begin{aligned}
KICc &\simeq -2\ln p(\mathbf{y}|\hat{\theta}_k) + \frac{(k+1)(3n-k-2)}{n-k-2} \\
&\quad + \frac{k}{n-k} \qquad (9)
\end{aligned}$$

It is worth to mention that asymptotically KICc will converge to KIC. This motivate the use of KICc for small sample applications.

4. SIMULATION RESULTS

A univariate autoregressive process of order k, AR(k), can be represented as

$$\begin{aligned}
y_t &= \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_k y_{t-k} + \varepsilon_t \\
\varepsilon_t &\sim N(0,\sigma^2).
\end{aligned}$$

Given a set of observations $\mathbf{y} = (y_1, y_2, \ldots, y_n)$, from such process, suppose our objective is to determine an appropriate order k of the autoregressive model.

In order to investigate the small sample performance of KICc, two sample size are used: $n = 23$ and $n = 30$ and 1000 realizations were generated from two models [9]:

$$\begin{array}{ll}
\text{Model 1} & y_t = 0.95y_{t-1} + \varepsilon_t \\
\text{Model 2} & y_t = 0.99y_{t-1} - 0.8y_{t-2} + \varepsilon_t \\
t = 1, \ldots, n. &
\end{array}$$

with ε_t independent and identically distributed standard normal. For each realization, Levinson-Durbin method was used to fit candidate autoregressive model of orders 1 to 20 and various criteria are used to select from among the candidate model.

The other criteria considered in our simulation sets are AIC, AICc , KIC, FPE [15] and BIC.

Table 1 gives the frequency of model orders selected by the different criteria for two sample sizes. It is clear that KICc performs best, closely followed by BIC, while other criteria perform less effectively. This improved selection property of KICc is due to two factors. The first is due to the additional information (about model's dissimilarity) in the J-divergence compared with I-divergence. The other is its finite sample bias correction.

Table 1. Frequency of the model order selected by each criterion for 1000 realizations.

Set	$\mathcal{M}$	n	Order	AIC	AICc	FPE	BIC	KIC	KICc
1	1	23	$< k_0$	0	0	0	0	0	0
1	1	23	$= k_0$	863	932	867	949	944	970
1	1	23	$> k_0$	137	68	133	51	56	30
2	1	30	$< k_0$	0	0	0	0	0	0
2	1	30	$= k_0$	835	895	837	952	925	962
2	1	30	$> k_0$	165	105	163	48	75	38
3	2	23	$< k_0$	25	36	25	50	45	71
3	2	23	$= k_0$	820	899	824	897	890	903
3	2	23	$> k_0$	155	65	151	53	65	26
4	2	30	$< k_0$	4	4	4	6	6	7
4	2	30	$= k_0$	827	908	829	950	926	961
4	2	30	$> k_0$	169	88	167	44	68	32

Figure 1 provides some insight as why KICc tends to outperform KIC as a selection criterion. Consider the fourth set of simulation based on generating Model 2 with sample size $n = 30$. Simulated values of $E_{\theta_0}\left\{KICc\right\}$, as given by Equation (9) and $E_{\theta_0}\left\{KIC\right\}$ are obtained by averaging KICc, and KIC, respectively over the 1000 replications. $\Omega(k,\theta_0)$ is obtained by averaging the exact expression of KICc given in Equation (8) using the digamma function. The average values for each of $\Omega(k,\theta_0)$, KICc and KIC are also plotted versus k. Since $\Omega(k,\theta_0)$, KICc and KIC are obtained by adding a non-stochastic penalty term to the log

likelihood, the three criteria have the same variance. This is why in comparison we emphasize only on mean values [1].
We note that KIC is a negatively biased estimator of $\Omega(k, \theta_0)$ specially when k increases. This negative bias is the major factor for the bad performance of KIC compared with KICc. The approximation made to derive KICc are quite reasonable and lead to results which are not far from that obtained with $\Omega(k, \theta_0)$ as presented in Table 2.

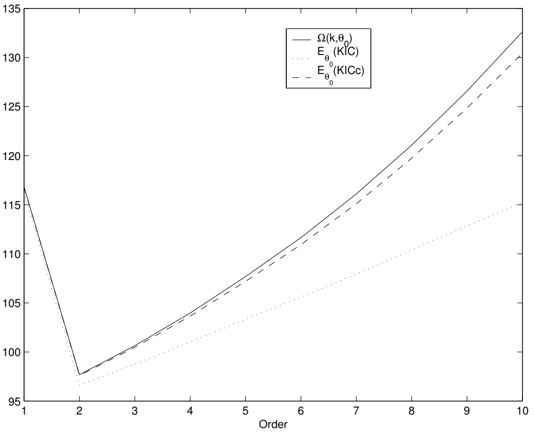

Fig. 1. Averages of KIC, KICc and $\Omega(k, \theta_0)$ for AR(2), $n = 30$

Table 2. Frequency of the model order selected by each criterion for 1000 realizations.

Set	$\mathcal{M}$	n	Order	KICc	$\Omega(k, \theta_0)$
1	1	23	$< k_0$	0	0
1	1	23	$= k_0$	970	972
1	1	23	$> k_0$	30	28
2	1	30	$< k_0$	0	0
2	1	30	$= k_0$	962	965
2	1	30	$> k_0$	38	35
3	2	23	$< k_0$	71	72
3	2	23	$= k_0$	903	901
3	2	23	$> k_0$	26	25
4	2	30	$< k_0$	7	7
4	2	30	$= k_0$	961	964
4	2	30	$> k_0$	32	29

5. CONCLUSION

The results in section 4 suggest that KICc should function as an effective model selection criterion in small sample applications. KICc has two major strength: firstly, it is based on $2J_n(\theta_0, \theta_k)$, which provides an additional information about model's dissimilarity compared with $2I_n(\theta_0, \theta_k)$, originally used to derive AIC-based criteria. Secondly KICc is an unbiased estimator of $2J_n(\theta_0, \hat{\theta}_k)$ rather than an asymptotically unbiased as for KIC. This makes KICc outperforming KIC in small sample cases.
The approximations made to get KICc of Equation (9) is very reasonable as tested by simulation. Even though it produces a slight bias, computational savings and performance relative to exact criterion justifies the use of such an approximation.
The type of bias adjustment proposed in this article is based on assuming a particular modeling framework of the candidate family $\mathcal{M}_k$, and using the characteristics of the framework to derive either an exact expression or a more precise approximation for the bias refinement. In future work, we aim to use an other alternative type of refinement based on using the bootstrap to approximate the bias correction.

6. REFERENCES

[1] J. E. Cavanaugh, "A large-sample model selection criterion based on kullback's symmetric divergence," *Statistics and Probability Letters*, vol. 42, pp. 333–343, 1999.

[2] P. M. Djuric, "A model selection rule for sinusoids in white gaussian noise," *IEEE Transaction on Signal Processing*, vol. 44, pp. 1744–1751, 1996.

[3] H. Akaike, "A new look at the statistical model identification," *IEEE Transaction on Automatic and Control*, vol. 19, pp. 716–723, 1974.

[4] M. Stone, "Cross validatory choice and assesment of statistical predictions (with discussion)," *Journal of the Royal Statistical Society series B*, vol. 36, pp. 111–147, 1978.

[5] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, pp. 461–464, 1978.

[6] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, pp. 465–471, 1978.

[7] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematics Statistics*, vol. 22, pp. 76–86, 1951.

[8] N. Sugiura, "Further analysis of the data by akaike's information criterion and the finite corrections," *Communication in Statistics*, vol. A 7, pp. 13–26, 1987.

[9] C. M. Hurvich and C. L. Tsai, "Regression and time series model selection in small samples," *Biometrika*, vol. 76, pp. 297–307, 1989.

[10] H. Jeffreys, "An invariant form of the prior probability in estimation problems," *The Royal Statistical Society*, vol. A 186, pp. 453–469, 1946.

[11] H. Akaike, "Infrmation theory and an extension of the maximum likelihood principle," *in Proceeding of the 2nd International Symposum of Information theory*, pp. 267–281, 1972.

[12] J. M. Bernardo, "Psi (digamma) function," *Applied Statistics*, vol. 25, pp. 315–317, 1976.

[13] A. R. Gallant, *Nonlinear Statistical Models*, New York Wiley, 1986, pp. 17.

[14] S. Kotz, N. L. Johnson, and C B Read eds, *Encyclopedia of Statistical Sciences*, vol. 2, New York Wiley, 1982, pp. 373.

[15] H. Akaike, "Statistical predictor identification," *Annals of the Institute of Statistical Mathematics*, vol. 22, pp. 203–217, 1970.

[1] An unbiased estimator of the information with large variance may be a poor criterion for model selection.