A KULLBACK'S SYMMETRIC DIVERGENCE CRITERION WITH APPLICATION TO LINEAR REGRESSION AND TIME SERIES MODEL

Hocine BELKACEMI

Laboratoire des signaux et systèmes
CNRS/Supélec,
91192 Gif sur Yvette, FRANCE
belkacemi@lss.supelec.fr

Abed-Krim SEGHOUANE

National ICT Australia SEACS Program
Locked Bag 8001
Canberra ACT 2601 Australia
Abd-krim.seghouane@nicta.com.au

ABSTRACT

The Kullback information criterion (KIC) is a recently developed tool for statistical model selection. KIC serves as an asymptotically unbiased estimator of the Kullback symmetric divergence, known as J-divergence. A corrected version for KIC denoted by KIC_C have been also proposed to correct the bias of KIC. This version tends to overfit when the sample size increases.In this paper we propose an alternative to KIC_C, the KIC_U criterion which is unbiased estimator of the Kullback's symmetric divergence. It provides better model choice than KIC_C for moderate to large sample size.

1. INTRODUCTION

In statistical modelling, one of the main objectives is to select a suitable model from a candidate class to characterize the underlying data. Model selection criteria provide a useful tool in this regard. A selection criterion assesses whether a fitted model offers an optimal balance between goodness-of-fit and parsimony.The first model selection criterion to gain widespread acceptance was the Akaike information criterion (AIC) [1]. Many other criteria have been subsequently introduced and studied such as Bayesian Information Criterion (BIC) [2] and the Minimum Description Length (MDL) [3]. AIC serves as an estimator of Kullbacks directed divergence between the generating or true model (i.e., the model which presumably gave rise to the data) and a fitted candidate model. The corrected AIC, denoted by $AICc$, is an adjusted version of AIC originally proposed for linear regression towards a bias reduction. [4]. Hurvich and Tsai justified its application to nonlinear regression models and investigated its small sample superiority over AIC [5]. Another possible criterion is the Kullback Information Criterion KIC [6] based on the Kullback's symmetric divergence khown also as J-divergence as a measure of model dissimilarity. As for AIC, when the number of candidate model k increases compared to the sample size n, KIC becomes strongly negatively biased estimate of the Kullback's symmetric divergence and leads to the choice of over parameterized models. A bias corrected version of KIC, denoted KIC_C has been proposed recently for linear regression and univariate autoregressive models [7]. However, the simulation studies indicates that KIC_C tends to overfit when the sample size increases. The aim of this paper is to propose an alternative to KIC_C, the KIC_U criterion which is unbiased estimator of the Kullback's symmetric divergence. It provides better model choice than KIC_C for moderate to large sample size. The remainder of this paper is organized as follows. In section 2, we briefly review KIC and its corrected version KIC_C. Section 3, is devoted to a development of the new proposed criterion, KIC_U. Simulation examples of comparaison are given in section 4 and a conclusion is given in section 5.

2. REVIEW OF KIC AND KIC_C

Suppose a collection of data $\mathbf{y}_n = (y_1, \cdots, y_n)^T$ has been generated according to an unknown parametric model $p(\mathbf{y}|\theta_0)$. We try to find a parametric model which provides a suitable approximation for $p(\mathbf{y}|\theta_0)$. Let $\mathbf{M}_k = \{\ p(\mathbf{y}|\theta_k)|\theta_k \in \Theta_k\ \}$ denote a k-dimensional parametric family and let $\hat{\theta_k}$ denote the vector of parameters estimate obtained by maximizing the likelihood function $p(\mathbf{y}_n|\theta_k)$ over Θ_k. For simplicity, we will assume $k = 1, 2, \ldots, k_{\max}$, so the collection $\mathbf{M}_k$s consists of nested families, i.e, $\Theta_1 \subset \Theta_2 \subset \cdots \subset \Theta_{k_{\max}}$ of dimension 1 through $(kmax)$ [1]. To determine which candidate model best approximates the generating unknown model $p(\mathbf{y}|\theta_0)$, we need a measure which provides a suitable reflection of the disparity between $p(\mathbf{y}|\theta_0)$ and an approximating model $p(\mathbf{y}|\theta_k)$. The Kullbacks symmetric divergence is one of such measure. Kullbacks symmetric divergence between

two parametric densities $p(\mathbf{y}|\theta_0)$ and$p(\mathbf{y}|\theta_k)$ is defined as

$$\begin{aligned} 2J_n(\theta_0,\theta_k) &= 2I_n(\theta_0,\theta_k)+2I_n(\theta_k,\theta_0) \quad (1)\\ &= E_{\theta_0}\{-2\ln p(\mathbf{y}|\theta_k)\}-E_{\theta_0}\{-2\ln p(\mathbf{y}|\theta_0)\}\\ &\quad +E_{\theta_k}\{-2\ln p(\mathbf{y}|\theta_0)\}\\ &\quad -E_{\theta_k}\{-2\ln p(\mathbf{y}|\theta_k)\}\\ &= d_n(\theta_0,\theta_k)-d_n(\theta_0,\theta_0)+d_n(\theta_k,\theta_0)\\ &\quad -d_n(\theta_k,\theta_k) \end{aligned}$$

where $I_n(\theta_i,\theta_j)$ is the directed Kullback divergence and $E_{\theta_j}\{.\}$ represents the expectation with respect to the density $p(\mathbf{y}|\theta_j)$. Since $d_n(\theta_0,\theta_0)$ does not depend on θ_k, any ranking of the candidate models according to $2J_n(\theta_0,\theta_k)$ would be identical to ranking them according to $K_n(\theta_0,\theta_k)$ defined by

$$K_n(\theta_0,\theta_k)=d_n(\theta_0,\theta_k)+d_n(\theta_k,\theta_0)-d_n(\theta_k,\theta_k) \quad (2)$$

In [6] Cavanaugh proposed the Kullback Information Criterion KIC as a bias correction to $-2\ln p(\mathbf{y}|\hat{\theta}_k)$ using the Kullbacks symmetric divergence variant as defined in equation(2) and given by

$$KIC=-2\ln p(\mathbf{y}|\hat{\theta}_k)+3(k+1) \quad (3)$$

Motivated by the fact that KIC is only asymptotically unbiased, the authors in [7] proposed a bias corrected version KIC_C given by

$$\begin{aligned} KIC_C &\simeq -2\ln p(\mathbf{y}|\hat{\theta}_k)+\frac{2(k+1)n}{n-k-2}+n\ln\left(\frac{n}{n-k}\right)\\ &\quad +\frac{n}{n-k} \quad (4) \end{aligned}$$

KIC_C is found to provide a better model order choice than KIC for small sample linear regression and univariate autoregressive model selection.

3. DERIVATION OF KIC_U

Suppose that the linear regression generating model for the data and the k^{th} candidate model are respectively given by

$$\mathbf{y}=X\beta_0+\varepsilon_0, \qquad \varepsilon_0\sim N(0,\sigma_0^2\mathbf{I}_n) \quad (5)$$

$$\mathbf{y}=X\beta_k+\varepsilon_k, \qquad \varepsilon_k\sim N(0,\sigma_k^2\mathbf{I}_n) \quad (6)$$

Here, $\mathbf{y}$ is an $n\times 1$ observation vector, ε_0 and ε_k are $n\times 1$ noise vectors, β_0 and β_k are $k\times 1$ parameter vectors, and X is an $n\times k$ design matrix of full column rank. The vector of parameters for the candidate model is $\theta_k=\left[\beta_k^t \quad \hat{\sigma^k}^2\right]$. Considering the candidate model as described in equation(6), we have

$$\begin{aligned} d_n(\theta_i,\theta_j) &= E_{\theta_i}\{\ -2\ln p(\mathbf{y}|\theta_j)\ \}\\ &= n\ln 2\pi+n\ln\sigma_j^2+n\frac{\sigma_i^2}{\sigma_j^2} \quad (7)\\ &\quad +\frac{1}{\sigma_j^2}(\beta_i-\beta_j)^TX^TX(\beta_i-\beta_j) \end{aligned}$$

Using the exact Kullback's symmetric divergence as defined in equation(1) and ignoring constant terms, we get

$$\begin{aligned} KIC &\simeq n(\ln(\hat{\sigma}_k^2)-\ln(\sigma_0^2))+\frac{2(k+1)n}{n-k-2} \quad (8)\\ &\quad +n\ln\left(\frac{n}{n-k}\right)+\frac{n}{n-k} \end{aligned}$$

where $\hat{\sigma}_k^2=\mathbf{y}^T(\mathbf{I}_n-(X(X^TX)^{-1}X)\mathbf{y}/n$ is the maximum likelihood estimate of σ_0^2 We can define a general family of KIC_C by using the penalty function in equation(4)

$$\begin{aligned} KIC_C(\sigma^2) &\simeq n(\ln(\sigma^2)-\ln(\sigma_0^2))+\frac{2(k+1)n}{n-k-2}\\ &\quad +n\ln\left(\frac{n}{n-k}\right)+\frac{n}{n-k} \quad (9) \end{aligned}$$

The optimal value occurs when $\ln(\sigma^2)=\ln(\sigma_0^2)$. In practice $\ln(\sigma_0^2)$ is not known. A good unbiased estimator of $\ln(\sigma_0^2)$ have been proposed in [8] defined by

$$n\ln\tilde{\sigma}^2=n\ln\hat{\sigma}^2+n\ln\frac{n}{n-k-1} \quad (10)$$

Substituting this estimate in equation 9 and ignoring the constant term $n\ln(\sigma_0^2)$, results in the new criterion denoted by KIC_U

$$\begin{aligned} KIC_U &\simeq n\ln(\hat{\sigma}_k^2)+\frac{2(k+1)n}{n-k-2}+n\ln\left(\frac{n}{n-k}\right)\\ &\quad +\frac{n}{n-k}+n\ln\frac{n}{n-k-1} \quad (11) \end{aligned}$$

Since $KIC_U=KIC_C+n\ln\frac{n}{n-k-1}$, KIC_U has a greater penalty for overfitting, especially as the sample size increases.

4. SIMULATION RESULTS

4.1. Linear regression

We consider three simulations sets, based on three different sample size ($n=25$, $n=50$ and $n=100$) generated using the linear regression model as described in equation 5 with parameters $\sigma_0^2=1$ and $\beta_0=(1,1,1,1)^T$. Ten candidate models were stored in $\mathbf{X}$, an $n\times 10$ matrix of independent identically normal random variables. The candidate models include the columns of $\mathbf{X}$ in a sequentially

nested fashion; that is, columns 1 to k define the design matrix for the candidate model with dimension k. The first four columns of $\mathbf{X}$ define the true model. For purpose of comparison, we have included the consistent criterion $BIC = n\ln\hat{\sigma}^2+(k+1)\ln(n)$, the $AIC = n\ln\hat{\sigma}^2+2(k+1)$ and the corrected $AIC_C = n\ln\hat{\sigma}^2+\frac{2(k+1)n}{n-k-2}$. Table 1 gives a comparison of KIC_U to KIC_C, KIC, AIC, AIC_C and BIC using 1000 runs of Monte Carlo simulation. KIC_U exhibits the best frequency of selection of the true order k_O for low sample following slightly the consistent criterion BIC for large sample size.

4.2. Autoregressive model

The problem of regression and autoregressive model selection are closely related. Indeed of the proposed solutions can be applied to both problems. A univariate autoregressive process of order k, AR(k), can be represented as

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_k y_{t-k} + \varepsilon_t \qquad (12)$$

where ε_t are i.i.d. $\sim N(0,\sigma_0^2)$. Given a set of observations $\mathbf{y}_n = (y_1, y_2, \ldots, y_n)^T$, our objective is to determine the appropriate order k that fits well the data $\mathbf{y}_n$. Equation 12 can be expressed in a linear form

$$\mathbf{y} = \mathbf{X}\phi + \mathbf{z} \qquad (13)$$

where $\phi = (\phi_1, \phi_2, \ldots, \phi_n)^T$, $\mathbf{z} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^T$ and $\mathbf{X}$ is the $n \times k$ design matrix

$$\mathbf{X} = \begin{bmatrix} y_0 & y_{-1} & \cdots & y_{1-k} \\ y_1 & y_0 & \cdots & y_{2-k} \\ \vdots & \vdots & & \vdots \\ y_{n-1} & y_{n-2} & \cdots & y_{n-k} \end{bmatrix}$$

Based on the equivalence between equation 13 and linear model, we can justify the use of or criterion KIC_U for the class of AR models. As in the previous example, three sample size sets are used ($n = 25$, $n = 50$ and $n = 100$) with $\sigma_0^2 = 1$. A Monte Carlo simulation that consist of 1000 data set realizations of size n were generated from the following second order AR model

$$y_t = 0.99y_{t-1} - 0.8y_{t-2} + \varepsilon_t, \qquad t = 1, \ldots, n.$$

For each data set, Levinson-Durbin method was used to fit candidate AR models of order 1 to 10. Other selection criteria are used in the simulation for comparaison purpose.The frequency of order selected by the various criteria is given in table 2. Here again KIC_U shows a good performance compared to KIC_C in small sample followed slightly by BIC for large sample with low tendency of overfitting compared to the other criteria. The tendency of the criteria to underestimate the correct dimension decreases as the sample size increases and is equal to zero.

5. CONCLUSION

In this paper, we have derived and investigated the model selection criterion, KIC_U, based on Kullback's symmetric divergence, fro model selection in univariate AR models and linear regression model. Our results of simulations show that KIC_U outperforms the KIC and the corrected KIC version for small and large sample size and it outperform the consistent criterion BIC except when the sample size is large with the lower tendency of overfitting compared to KIC and KIC_C.

6. REFERENCES

[1] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Automatic Control*, vol. AC-19, pp. 716–723, 1974.

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

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

[4] N. Sugiura, "Further analysis of the data by akaikes information criterion and the finite corrections," *Communication in Statistics*, vol. A7, pp. 13–26, 1987.

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

[6] 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.

[7] A. K. Seghouane and M. Bekara, "A small sample model selection criterion based on kullback's symetric divergence," *IEEE Trans. on Signal Processing*, vol. 52, no. 12, pp. 3314–3323, Dec. 2004.

[8] R. Shumway A. McQuarrie and C.L. Tsai, "The model selection criterion aicu," *Statistics and Probabilityy letters*, vol. 34, pp. 285–292, 1997.

Set	n	Order	AIC	AIC_c	BIC	KIC	KIC_c	KIC_u
1	25	$< k_0$	6	11	11	10	19	40
1	25	$= k_0$	548	853	785	763	913	**926**
1	25	$> k_0$	446	136	204	227	68	34
2	50	$< k_0$	0	0	0	0	0	0
2	50	$= k_0$	647	858	924	858	912	**961**
2	50	$> k_0$	353	192	76	142	88	39
3	100	$< k_0$	0	0	0	0	0	0
3	100	$= k_0$	701	762	962	852	883	**949**
3	100	$> k_0$	299	238	38	148	117	51

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

Set	n	Order	AIC	AIC_c	BIC	KIC	KIC_c	KIC_u
1	25	$< k_0$	12	18	36	29	45	56
1	25	$= k_0$	858	923	921	921	936	**938**
1	25	$> k_0$	130	59	43	50	19	6
2	50	$< k_0$	0	0	0	0	0	0
2	50	$= k_0$	799	872	963	919	944	**978**
2	50	$> k_0$	201	128	37	81	56	22
3	100	$< k_0$	0	0	0	0	0	0
3	100	$= k_0$	746	786	966	891	912	**958**
3	100	$> k_0$	254	214	34	109	88	42

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