DIGITALCOMMONS @WAYNESTATE

Journal of Modern Applied Statistical Methods

Volume 13 | Issue 2 Article 25

11-2014

The Information Criterion

Masume Ghahramani
Department of Statistics, Payam Noor University, Iran, mghdatagilan@gmail.com

Journal of Modern Applied Statistical Methods
November 2014, Vol. 13, No. 2, 444-454.

ISSN 1538 − 9472

The Information Criterion

M. Ghahramani
Payam Noor University
Tehran, Iran

The Akaike information criterion, AIC, is widely used for model selection. Using the AIC as the estimator of asymptotic unbias for the second term Kullbake-Leibler risk considers the divergence between the true model and offered models. However, it is an inconsistent estimator. A proposed approach the problem is the use of A'IC, a consistently offered information criterion. Model selection of classic and linear models are considered by a Monte Carlo simulation.

Keywords: Consistency, AIC, information criterion, Kullbake-Leibler risk, model selection

Introduction

Statistical modeling is used for investigating a random phenomenon that isn't completely predictable. One of the criteria frequently used in model selection is the Kullbake-Leibler (KL) information criterion (Kullback and Leibler, 1951). This information criterion was introduced as one risk in model selection. Akaike (1973) introduced information criterion, AIC, as an estimator of asymptotic unbias for the second term KL risk and to form a penalty likelihood function. Akaike stated modeling isn't only finding a model which describes the behavior of the observed data, but its main aim is predicated as a possible good, and the future of the process is under investigation. Hall (1987) used the Kullbake-Leibler risk considered bias and variance in the approximate density function. Bozdogan (2000), with the error distinction in the model selection, considered two errors from bias and variance in the estimation of model selection. Choi and Kiffer (2006), and Cawley and Talbot (2010) have considered the over fitting in model selection, and they showed over fitting results from the bias when modeling phenomena have been considered. Over the years, corrections have been made on penalty term, and criteria such as AIC (Akaike, 1973), TIC (Takeuchi, 1976), and KIC (Cavanaugh, 1994) have been

M. Ghahramani is in the Department of Mathematical Statistics. Email him at estimatormgh@gmail.com.

introduced. In section 2, we state the Kullbake-Liebler risk, and the necessity of definitions. In section 3, a consistent information criterion is proposed instead of the AIC. In section 4, we present the results of our simulation studies.

Kullbake-Leibler Risk

Let $X = (X_{1,} X_{2,} ..., X_n)$ be a (i.i.d) random sample from true model and unknown, $h(.)$, and the family $F_{\theta_k} = \{ f(.;\theta_k) = f_{\theta_k}; \theta_k \varepsilon \Theta \subseteq R^k \}$ from offered models has been considered for approximate true model.

Definition 1

The family F_{θ_k} is well specified if there is a $\theta_0 \varepsilon \theta$ such that $h(.) = f(.;\theta_0)$; otherwise it is misspecified.

Definition 2

The KL risk defines for generate model and unknown $h(.)$, and offered model f_{θ_k} as

$$KL(h, f_{\theta_k}) = E_h \left[\log(\frac{h(.)}{f(.;\theta_k)} \right] = E_h[\log\ h(.)] - E_h[\log\ f(.;\theta_k)] \tag{1}$$

where the expectation is taken with respect to the unknown model $h(.)$. The first term in the right hand side of (1) is called irrelevant part, because it doesn't depend on θ_k, and the second term is called relevant part. Based on the properties of the KL risk, the smaller value showed the closeness of the offered model to the unknown and true model. Therefore the problem is reduced to obtain a good estimate of the expected log-likelihood. Since the expectation is with respect to the model with unknown parameters, one estimator is

$$E_h\{\log\ f(.;\hat{\theta}_n)\} = \frac{1}{n} \Sigma_{i=1}^n \log\ f(x_i\ ;\hat{\theta}_n).$$

Thus, $\hat{\theta}_n$ is the maximum likelihood estimator of θ_k and $f(.;\hat{\theta}_n)$ is the maximum likelihood function. The bias of maximum log-likelihood is as

Bias estimator = $E_h\{\log f(.;\hat{\theta}_n) - nE_h\{\log f(Z\ ;\ \hat{\theta}_n)\}$

where Z is a random variable (i.i.d) with $X_i s$ .The general form of the information criterion that has been shown by IC, as

$$IC = -2\Sigma_{i=1}^{n} \log f(X_i\ ;\hat{\theta}_n) + 2\{\text{bias estimator}\} = -2l_f(\hat{\theta}_n) + 2\{\text{bias estimator}\}.$$

Akaike, when offered family is well specified, size of bias is estimated with dimensional parameter $\hat{\theta}_n$, means k, and the Akaike information criterion is stated as

$$AIC = -2l_f(\hat{\theta}_n) + 2k.$$

With attention to form the AIC by increasing the number of parameters in the offered model the penalty term, 2k will be increased and the term $-2\Sigma_{i=1}^{n} \log f(X_i\ ;\hat{\theta}_n)$ will be decrease. Penalty term is constant to chance of size sample in the information criterion AIC, and by increasing the size sample, AIC cannot distinguish the true model with the probability one. Therefore this problem is the same concept of inconsistency for an information criterion. Following the inconsistency of information criterion AIC, based on the definition similar to the definition of AIC, a consistent of information criterion, which called A'IC has presented. Akaike information criterion, by Akaike for model selection is introduced, but this useful criterion is inconsistent (Akaike, 1973).

The information criterion is obtained as follows. The basis of the log-likelihood function is

$$b = E_h\ \{\log f(.;\ \hat{\theta}_n) - nE_h \log f(Z\ ;\hat{\theta}_n)\}$$

where in the second term of the right hand side the inner expectation is calculated with respect to h(z) and the outer expectation is calculated with respect to h(x). By evaluating the bias it is decomposed as follows:

$$\begin{aligned} b = E_h\ \{gf(.;\ \hat{\theta}_n) - \log f(.;\ \theta_0)\} + E_h\{\log f(.;\ \theta_0) - nE_h\{\log f(Z\ ;\theta_0)\}\} \\ + nE_h\{E_h\{\log f(Z\ ;\theta_0) - E_h\{\log f\left(Z;\hat{\theta}_n\right)\}\} = b_1 + b_2 + b_3. \end{aligned}$$

The three expectations are calculated separately b_1, b_2, and b_3.

a) For calculation of b_1 by writing $l_f(\theta_0) = \log f(.;\theta_0)$ and by applying a Taylor series expansion around the maximum likelihood estimator $\hat{\theta}_n$, results in

$$l_f(\theta_0) = l_f(\hat{\theta}_n) + (\theta_0 - \hat{\theta}_n)^T \frac{\partial l_f(\theta)}{\partial \theta}|_{\theta=\hat{\theta}_n} + \frac{1}{2}(\theta_0 - \hat{\theta}_n)^T \frac{\partial^2 l_f(\theta)}{\partial \theta}|_{\theta=\hat{\theta}_n} (\theta_0 - \hat{\theta}_n) + o_p(1), \quad (1)$$

$O_p(1)$ is an expression of quantity that in the probability tends to zero.

With attention, the $\frac{\partial l_f(\theta)}{\partial \theta} | \theta = \hat{\theta}_n = 0$ and $\frac{1}{n}\frac{\partial^2 l_f(\theta)}{\partial \theta \partial \theta^T}|_{\theta=\hat{\theta}_n}$ is converged to $J(\theta_0)$. (Akaike, 1973).Thus,

$$J(\theta_0) = -E_h[\frac{\partial^2 l_f(\theta)}{\partial \theta \partial \theta^T}]|_{\theta=\theta_0}$$

Thus, the relation above can be approximated, as

$$l_f(\hat{\theta}_n) - l_f(\theta_0) \approx \frac{n}{2}(\theta_0 - \hat{\theta}_n)^T J(\theta_0)(\theta_0 - \hat{\theta}_n) + o_p(1),$$

This based on the b_1 can be written as

$$b_1 = E_h\{l_f(\hat{\theta}_n) - l_f(\theta_0)\} \approx E_h\{\frac{n}{2}(\theta_0 - \hat{\theta}_n)^T J(\theta_0)(\theta_0 - \hat{\theta}_n)\} \quad (2)$$

b) The b_2 doesn't contain an estimator and it can easily be written as

$$b_2 = E_h\{gf(.;\theta_0) - nE_h\{\log f(Z;\theta_0)\}\} = 0 \quad (3)$$

c) For calculation of value the b_3, first, the phrase $E_h\{\log f(Z;\theta_0)\}$ be defined equally of $\eta(\hat{\theta}_n)$. By using from Taylor expectation $\eta(\hat{\theta}_n)$ around θ_0,

$$\eta(\hat{\theta}_n) = \eta(\theta_0) + (\hat{\theta}_n - \theta_0)^T \frac{\partial \eta(\theta)}{\partial \theta}|_{\theta=\theta_0} + \frac{1}{2}(\hat{\theta}_n - \theta_0)^T \frac{\partial^2 \eta(\theta)}{\partial \theta \partial \theta^T}|_{\theta=\theta_0} (\hat{\theta}_n - \theta_0) + o_p(1)$$

with attention to the $\frac{\partial \eta(\theta)}{\partial \eta}|_{\theta=\theta_0}=0$. Thus when n tends to infinity, the relation above can be approximated as

$$\eta(\hat{\theta}_n) \approx \eta(\theta_0) + \frac{1}{2}(\hat{\theta}_n - \theta_0)^T J(\theta_0)(\hat{\theta}_n - \theta_0) + o_p(1).$$

Thus the b_3 can be written as

$$\begin{aligned} b_3 &= nE_h\{E_h\{\log f(Z\ ;\theta_0)\} - E_h\{\log f\left(Z\ ;\hat{\theta}_n\right)\}\} \\ &\approx \frac{n}{2}E_h\{(\theta_0 - \hat{\theta}_n)^T J(\theta_0)(\hat{\theta}_n - \theta_0)\} \end{aligned} \tag{4}$$

If the family of F_{θ_k} is well specified, with attention to quadratic forms in relations (2) and (4), that converge to centrally distributed chi-square with k degrees of freedom, then b_1 and b_3 can be written as

$$b_1 = b_3 = \frac{n}{2}k \tag{5}$$

By combining of b_1 and b_3, in relation (5) and b_2, in relation (3), bias the b is $b = b_1 + b_2 + b_3 = nk$.

By replacing the value of b in the general form of the information criterion, the offered information criterion called, A'IC is obtained as

$$A'IC = -2\Sigma_{i=1}^{n} \log f(X_i\ ;\hat{\theta}_n) + 2nk \tag{6}$$

In the offered information criterion A'IC, penalty term 2nk changes will change with sample size. So, if sample size will be very large, information criterion A'IC, with the probability of one, find the true model data. In other words, information criterion A'IC is the only consistent information criterion that has been obtained based on the Kullback-Leibler risk. To show consistency of information criterion A'IC, let the maximum likelihood function estimator for the offered model $(f(.;\theta_k) = f(\theta_k))$ and optimal model $(f(.;\theta_{k_0}) = f(\theta_{k_0}))$ with respectively

$l_f(\hat{\theta}_{k(n)})$ and $l_f(\hat{\theta}_{k_0(n)})$. With regard to relation (6) information criterion A'IC, for the model $f(\theta_k)$ and $f(\theta_{k_0})$, we have

$$A'IC(f(\theta_k)) = -2l_f(\hat{\theta}_{k(n)}) + 2nk,$$

$$A'IC(f(\theta_{k_0})) = -2l_f(\hat{\theta}_{k_0(n)}) + 2nk_0$$

If there is $k > k_0$, consistency for information criterion A'IC is given by

$$P(A'ICf(\theta_k) - A'ICf(\theta_{k_0}) > 0)$$

$$= P(-2l_f(\hat{\theta}_{k(n)}) + 2nk - (-2l_f(\theta_{k_0(n)}) + 2nk_0) > 0)$$

$$= P(2l_f(\hat{\theta}_{k(n)}) - 2l_f(\theta_{k_0(n)}) < 2nk - 2nk_0)$$

$$= P(U_n < 2n(k-k_0)) = F(2n(k-k_0))^p \rightarrow F(\infty) = 1 \qquad (7)$$

In relation (7), U_n is $2l_f(\hat{\theta}_{k(n)}) - 2l_f(\hat{\theta}_{k_0(n)})$ and the distribution function of chi-square has been shown by F. Therefore it tends in of the probability to one. Thus A'IC is a consistent information criterion. (For further study about the consistency of an information criterion, see Hu and Shao 2008).

Simulation

A simulation was conducted for usage and comparison of the offered information criterion, A'IC, with the information criterion AIC, by using Monte-Carlo simulation, for linear regression and classic models. This simulation of linear regression model is supposed that well specified family $F_{\theta_k} = \{f(.;\theta_k) = f_{\theta_k}; \theta_k \varepsilon \Theta \subseteq R^k\}$, and misspecified family $G_{\beta_d} = \{g(.;\beta_d) = g_{\beta_d}; \beta_d \varepsilon B \subseteq R^d\}$ are given for estimating the true model. Let $f: y_i = 0.3 + 0.5x_{i1} + x_{i2} + 0.7x_{i3} + \varepsilon_{i1} i = 1,...,n$ as the true model so that ε_{i1}, has been generated as random from distribution N(0,2). Models $f_1: y_i = \hat{\theta}_0 + \hat{\theta}_1 x_{i1} + \hat{\theta}_2 x_{i2} + \hat{\theta}_3 x_{i3}$ $\text{i} = 1,...,n$ and,

$f_2 : y_i = \hat{\theta}_0 + \hat{\theta}_1 x_{i1} + \hat{\theta}_2 x_{i2} + \hat{\theta}_3 x_{i3} + \hat{\theta}_4 x_{i4}$ $\mathrm{i} = 1,...,n$ offered models, which have been generated from F_{θ_k}. Also we have $g : y_i = 0.3 + 0.5 z_{i1} + 3 z_{i2} + 1.1 z_{i3} + \varepsilon_{i2} i = 1,...,n$

Table 1. Comparison of AIC with A'IC by using from Monte-Carlo simulation for linear regression models f_1, f_2, g_1, and g_2.

Size	Model	AIC	A'IC	Δ AIC	Δ A'IC
n=50	f_1	-2990	-2598	-	-
	f_2	-2700	-2210	290	388
	g_1	200	592	3190	2006
	g_2	248	738	3238	1860
n=100	f_1	-3500	-2708	-	-
	f_2	-3200	-2210	300	498
	g_1	430	1222	3930	3930
	g_2	455	1445	3955	4153
n=200	f_1	-5400	-3808	-	-
	f_2	-4360	-2370	1040	1438
	g_1	210	1802	5610	5610
	g_2	240	2230	5640	6038
n=350	f_1	-7230	-4438	-	-
	f_2	-6400	-2910	30	1528
	g_1	325	3117	7555	7555
	g_2	360	3850	7590	8288
n-500	f_1	-9730	-5738	-	-
	f_2	-9300	-4310	430	1428
	g_1	400	4392	10130	10130
	g_2	425	5415	10155	11153

Thus, ε_{i_2} , has been generated as random from distribution N(0,1), and Models $g_1 : y_i = \hat{\beta}_0 + \hat{\beta}_1 z_{i1} + \hat{\beta}_2 z_{i2} + \hat{\beta}_3 z_{i3} \ i = 1,...,n$ and $g_2 : y_i = \hat{\beta}_0 + \hat{\beta}_1 z_{i1} + \hat{\beta}_2 z_{i2} + \hat{\beta}_3 z_{i3} + \hat{\beta}_4 z_{i4} \ i = 1,...,n$. The models are generated from G_{β_d}. This simulation is achieved by using from software R, and the number of repetitions are 10^3, and samples n = 50, 100, 200, 350, 600, have been considered. The results of simulation are presented in the Table 1.

In the third and fourth columns of Table 1, the value of AIC and A'IC are presented in order to various values of n and for offered models f_1, f_2, g_1, and g_2. Therefore the relation between values AIC for offering models is obvious as $AIC(f_1) < AIC(f_2) < AIC(g_1) < AIC(g_2)$.

The family F_{θ_k} is well specified, but the family G_{β_d} is misspecified. Thus, this relation is logical. With attention to the fourth column of Table 1 recent relation also is confirmed for A'IC. In other words $A'IC(f_1) < A'IC(f_2) < A'IC(g_1) < A'IC(g_2)$.

With increasing n, the value of A'IC has been increased for the offered models, but the direction is confirmed unequally. The absolute magnitude difference of the value AIC and A'IC between the model of f_1 and other models is presented in the fifth and sixth columns of table. The absolute magnitude differences have been shown by the symbols of ΔAIC and Δ A'IC. If there are symbols, as

$$\Delta AIC_{|f_1 - f_2|} = |AIC(f_1) - AIC(f_2)| \text{ and } \Delta AIC_{|f_1 - g_j|} = |AIC(f_1) - AIC(g_j)|, \ \ \mathrm{j} = 1, 2$$

$$\Delta A'IC_{|f_1 - f_2|} = |A'IC(f_1) - A'IC(f_2)| \text{ and } \Delta A'IC_{|f_1 - g_j|} = |A'IC(f_1) - A'IC(g_j)|, \ \ \mathrm{j} = 1, 2$$

For n=50, 100, 150, 200, 350, 500, and models f_1, f_2, g_1, and g_2 will be confirmed the relation as

$$\Delta AIC_{|f_1 - f_2|} < \Delta AIC_{|f_1 - g_1|} < \Delta AIC_{|f_1 - g_2|} \text{ and} < \Delta A'IC_{|f_1 - f_2|} < \Delta A'IC_{|f_1 - g_1|} < \Delta A'IC_{|f_1 - g_2|}.$$

With attention to these relations the direction of similarity the model selection for information criteria AIC and A'IC for various n have been shown with this the quality that the criterion A'IC is a consistent information criterion.

Table 2. Comparison of AIC with A'IC by using Monte-Carlo simulation, for the state that generates model data is Normal standard and offered models are from a Laplace family with different parameters.

Size	Model	AIC	A'IC	Δ AIC	Δ A'IC
n=50	$f_1 = lap(0,1.3)$	-90	106	-	-
	$f_2 = lap(0,1)$	-70	126	20	20
	$f_3 = lap(2,1)$	-56	140	34	34
	$f_4 = lap(-2,0.9)$	-50	146	40	40
n=100	$f_1 = lap(0,1.3)$	-200	196	-	-
	$f_2 = lap(0,1)$	-160	236	40	40
	$f_3 = lap(2,1)$	-143	253	57	57
	$f_4 = lap(-2,0.9)$	-130	266	70	70
n=200	$f_1 = lap(0,1.3)$	-345	451	-	-
	$f_2 = lap(0,1)$	-295	501	50	50
	$f_3 = lap(2,1)$	-255	541	90	90
	$f_4 = lap(-2,0.9)$	-240	556	105	105
n=350	$f_1 = lap(0,1.3)$	-610	786	-	-
	$f_2 = lap(0,1)$	-525	871	85	85
	$f_3 = lap(2,1)$	-487	909	123	123
	$f_4 = lap(-2,0.9)$	-441	955	169	169
n=500	$f_1 = lap(0,1.3)$	-986	1010	-	-
	$f_2 = lap(0,1)$	-865	1131	121	121
	$f_3 = lap(2,1)$	-777	1219	209	209
	$f_4 = lap(-2,0.9)$	-670	11326	316	316

In the third and fourth column Table 2 values of AIC and A'IC for n=50, 100, 200, 350 and 500, have been respectively considered Laplace offered models f_1, f_2, f_3, and f_4. Therefore the relation between values AIC for offered models of Laplace family is obvious as $A'IC(f_1) < A'IC(f_2) < A'IC(f_3) < A'IC(f_4)$.

With attention to the fourth column in the Table 2, the recent relationis also confirmed for A'IC. In other words, $A'IC(f_1) < A'IC(f_2) < A'IC(f_3) < A'IC(f_4)$. In the fifth and sixth columns the absolute magnitude difference have been presented respectively for the value AIC and A'IC between the model of f_1 and any which from other models to confirm with any n, symbols of ΔAIC and ΔA'IC has been shown. With attention to these two columns for n's different have ΔAIC = ΔA'IC. If we have these symbols as

$$\Delta AIC_{|f_i - f_j|} = | AIC(f_i) - AIC(f_j) | \ \ \text{i} \neq \text{j and } \Delta A'IC_{|f_i - f_j|} = | A'IC(f_i) - A'IC(f_j) | \ \ i \neq \text{j}$$

for any n= 50,100, 200, 350, 500, models f_1, f_2, f_3, and f_4, confirms the relation as

$$\Delta AIC_{|f_1 - f_2|} < \Delta AIC_{|f_1 - f_3|} < \Delta AIC_{|f_1 - f_4|} \text{ and } \Delta A'IC_{|f_1 - f_2|} < \Delta A'IC_{|f_1 - f_3|} < \Delta A'IC_{|f_1 - f_4|}.$$

With attention to these relations, the direction of similarity model selection for information criteria AIC and A'IC for various n has been shown. But the information criterion A'IC is the consistent information criterion.

Conclusion

In this article investigating the inconsistent information criterion AIC, and by eliminating the inconsistency problem, a method for achieving an information criterion has been presented based on Kullback-Leibler risk and the consistent information criterion A'IC has been obtained. Therefore this information criterion is the only consistent information criterion and asymptotically unbiased. It is obtained based on Kullback-Leibler risk. Via simulation for linear regression and classic model, the quality of model selection was shown throughout the two information criterion, AIC and A'IC. According to the consistent information criterion of A'IC, it is possible for further discussion and to refine the other information criteria, which are based on Kullback-Leibler risk (as AICc and KICc) and add the consistency feature to the criteria.

References

Akaike, H. (1973). Information theory and an extension of maximum likelihood principle. *Second Intention Symposium on Information Theory*, 267-281.

Bozdogan, H. (2000). Akaike's information criterion and recent developments in information complexity. *Journal of Mathematical Psychology, 44*(1), 62-91.

Cavanagh, J. E. (1994). A large sample model selection criterion based on kulback's symmetric divergence. *Statistics and Probability Letters, 44*, 333-344.

Cawley, G., & Talbot, N. (2010). On over-fitting model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, *11*, 2079-2107.

Choi, H., & Kiffer, N. M. (2006). Differential geometry and bias correction in nonnested hypothesis testing. Unpublished manuscript, Department of Economics, Cornell University, Ithaca, NY.

Hall, P. 1987. On Kullbake-Leibler loss and density estimation. *The Annals of Statistics*, *15*(4), 1491-1519.

Hu, B., & Shao, J. (2008). Generalized linear model selection using r. *Journal of Statistical Planning and Information, 138*(12), 3705-3712.

Konishi, S., & Kitagawa, G. (1996). Generalized information criteria in model selection. *Biometrika, 83*(4), 875-890.

Kullback, S., & Leibler, R.A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics, 22*(1), 76-86.

Takeuchi, K. (1976). Distribution of information statistics and criteria for adequacy of models. *Mathematical Sciences*, *153*, 12-18.