***The First Conference on Computational Group Theory, Computational Number Theory and Applications*,**
University of Kashan, 26-28 Azar, 1393 (December 17-19 2014), pp: 63-67.

Poster Presentation

Considering the Information Criteria

Masume Ghahramani
Department of Mathematics, Gilan-E-Gharb Branch, Islamic Azad University, Gilan-E- Gharb, Iran
mghdatagilan@gmail.com

Mehdi Shams
Department of Statistics, School of Mathematical Sciences, University of Kashan, Kashan, Iran
mehdishams@kashanu.ac.ir

Abstract

Statistical modeling is a crucial issue in scientific data analysis. Models are used to represent stochastic structures, predict future behaviour, and extract usful information from data. Many researchers in medicine, engineering, social sciences and economics, planning, management, geography, physics, mathematics, statistics and other sciences to conduct an investigation, according to the data available for the test under consideration requires notice to the selection criteria models are suitable. When the correct model is unknown, the researchers proposed a family of models to find the closest model is the correct model, it is necessary to define a criterion for unbiased information. This paper examines the information criteria AIC and AICc deals.

Keywords: Bias, information criterion, Kullback-Leibler risk.

MSC(2010): Primary: 94A15; Secondary: 65F05, 46L05, 11Y50.

1 Introduction

Statistical modeling is used for investigating a random phenomenon that is not completely predictable. One of the criteria that have usage of the frequency in model selection is Kullback-Leibler (KL) information criterion (see Kullback and Leibler 1951). This information criterion was introduced as one risk in model selection. Akaike (1973) introduced information criterion, AIC as asymptotically the unbiased of an estimator for the second term the KL risk and to form penalty likelihood function. Akaike stated modeling is not only finding a model which describes the behavior of the observed data, but it is main aim is predicted as a possible good, the future of the process under investigation. The during these years has been made the corrections on penalty term, and criteria such as AIC (Akaike 1973), TIC (Takeuchi 1976), are introduced. In section 2, is stated the Kullbake-Liebler risk and a consistent information criterion is proposed instead of the AIC. In section 3, information criterion $A'ICc$ is proposed instead of the AICc. In section 4, we present the main results.

2 Kullback-Leibler (KL)

Let $X = (X_1, ..., X_n)$ is a i.i.d random sample from true model and unknown, $h(.)$ and the family $F_{\theta_k} = \{f(.;\theta_k) = f_{\theta_k}; \theta_k \in \Theta \subseteq R^k\}$ from offered models has been considered for approximate true model. The family F_{θ_k} is well specified, if there is a $\theta_0 \in \Theta$ such that $h(.) = f(.;\theta_0)$; otherwise it is mis specified. The KL risk definds for generate model and unknown $h(.)$, and offered model, f_{θ_k} as:

$$KL(h, f_{\theta_k}) = E_h[log(\frac{h(.)}{f(.;\theta_k)})] = E_h[logh(.)] - E_h[logf(.;\theta_k)] \quad (1)$$

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 does not 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 reduces 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 $E_h\{logf(.;\hat{\theta}_n)\} = \frac{1}{n}\sum_{i=1}^{n} logf(X_i;\hat{\theta}_n)$. So that $\hat{\theta}_n$ is the maximum likelihood estimator of θ_k and $f(.;\hat{\theta}_n)$ is the maximum likelihood function.
The general form of the information criterion that has been shown by IC, as:

IC = -2 (log-likelihood of statistical model - bias estimator)=$-2l_f(\hat{\theta}_n)$ +2 bias estimator.

Akaike, when offered family is well specified, size of bias is estimated with dimensional parameter $\hat{\theta}_n$, means k, and 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$\sum_{i=1}^{n} logf(X_i;\hat{\theta}_n)$ will be decrease. Penalty term is constant to chance of size sample in the information criterion AIC, and by increasing the size sample, AIC can not 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 (see Akaike 1973). In this selection the bias term has used in the general form information criterion is considered from another perspective. We obtain the information criterion that furthermore has nice specials the information criterion AIC, it is also consistent. In the beginning the bais of the log-likelihood function as follows:

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

so that Z is a random variable i.i.d with $X_i's$. 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 composed as follows:

$$b = E_h\{logf(.;\hat{\theta}_n) - logf(.;\theta_0)\} + E_h\{logf(.;\theta_0) - nE_h\{logf(Z;\theta_0)\}\}$$
$$+nE_h\{E_h\{logf(Z;\theta_0) - E_h\{logf(Z;\hat{\theta}_n)\}\} = b_1 + b_2 + b_3.$$

We calculate the three expectations separately b_1, b_2 and b_3.
a) For calculation of b_1 by writing $l_f(\theta_0) = logf(.;\theta_0)$ and by applying a Taylor series expansion around the maximum likelihood estimator $\hat{\theta}_n$, we have

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

$o_p(1)$ is expression of quantity that in the probability tends to zero. With attention to, the $\frac{\partial l_f(\theta)}{\partial \theta} \mid_{\theta=\hat{\theta}_n} = 0$ and $\frac{1}{n}\frac{\partial^2 l_f(\theta)}{\partial \theta \partial \theta^T} \mid_{\theta=\hat{\theta}_n}$ is converge to $J(\theta_0)$. (for more study see Akaike 1973).
So, $J(\theta_0) = -E_h[\frac{\partial^2 l_f(\theta)}{\partial \theta \partial \theta^T}] \mid_{\theta=\theta_0}$ 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 follow:

$$b_1 \approx E_h\{\frac{n}{2}(\theta_0 - \hat{\theta}_n)^T J(\theta_0)(\theta_0 - \hat{\theta}_n)\} \quad (3)$$

b) The b_2 does not contain an estimator and it can easily be written as;

$$b_2 = E_h\{logf(.;\theta_0) - nE_h\{logf(Z;\theta_0)\}\} = 0 \quad (4)$$

c) For calculation of value the b_3 first, the phrase $E_h\{logf(Z;\theta_0)\}$ be definded equally of $\Omega(\hat{\theta}_n)$. By using from Taylor expectation $\Omega(\hat{\theta}_n)$ around θ_0 we have:

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

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

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

Thus the b_3 can be written as:

$$b_3 \approx \frac{n}{2}E_h\{(\hat{\theta}_n - \theta_0)^T J(\theta_0)(\hat{\theta}_n - \theta_0)\} \quad (5)$$

If the family of F_{θ_k} is well specified, with attention to quadratic forms in relations (3) and (5), that converge to centrally distributed chi-square with k degrees of freedom. Therefore b_1 and b_3 can be written as:

$$b_1 = b_3 = \frac{n}{2}k \quad (6)$$

So by combining of b_1 and b_3 , in relation (6) and , in relation (4), bias the b is as follow: $b = b_1 + b_2 + b_3 = nk$. With replacing the value of b in the general form of the information criterion, the

offered information criterion called, $A'IC$ is obtained as: $A'IC = -2l_f(\hat{\theta}_n)$+2nk.
In the offered information criterion $A'IC$, penalty term 2nk changes will change with sample size changes. 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 Kullback-Leibler risk. (For further study about the consistency of an information criterion, see Hu and Shao 2008).

3 Information criterion AICc

The AIC penalizes for the addition of parameters, and thus selects a model that fits well but has a minimum number of parameters (i.e., simplicity and parsimony). For small sample sizes (i.e. $\frac{n}{k} < 40$), the second-order Akaike Information Criterion($AICc$) should be used instead:

$$AICc = -2(log - likelihood) + 2k + \frac{2k(k+1)}{n-k-1} = AIC + \frac{2k(k+1)}{n-k-1}$$

where n is the sample size. As sample size increases, the last term of the AICc approaches zero, and the AICc tends to yield the same conclusions as the AIC (Burnham and Anderson 2002). By attention to information criterion $A'IC$, information criterion $A'ICc$ is proposed instead of the AICc.

$$A'ICc = -2(log - likelihood) + 2nk + \frac{2k(k+1)}{n-k-1} + 2k - 2k = A'IC + \frac{2k(k+1)}{n-k-1}.$$

4 Main Results

The Akaike information criterion (AIC) is a measure of the relative quality of a statistical model, for a given set of data. As such, AIC provides a means for model selection. The AIC can not distinguish the true model with the probability one. Therefore this paper a consistent information criterion is proposed instead of the AIC, and also information criterion $A'ICc$ is proposed instead of the AICc. for future research to further explore the information criteria AIC, $A'IC$, AICc and $A'ICc$ in different models to be compared, and the principle of parsimony and goodness of fit can be examined.

References

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

[2] Akaike, H, *A new look at the statistical model identification*, ,IEEE Transactions on Automatic Control (1974), 19 (6): 716 723.

[3] Burnham, K. P.; Anderson, D. R. , *Multimodel inference: understanding AIC and BIC in Model Selection* , Sociological Methods and Research (2002), 33: 261304.

[4] Hu, B. Shao, J, *Generalized linear model selection using r*, journal of statistical planning and information (2008), 138, 3705-3712.

[5] Konishi, S. and Kitagawa, G, *Generalized information criteria in model selection*, Biometrika (1996), 83, 875-890.

[6] Kullback, S., Leibler, R.A., *On information and sufficiency*. Ann. Math. Statist, (1951), 22, 1,76-86.

[7] Takeuchi, K., *Distribution of information statistics and criteria for adequacy of models*, mathematical sciences in japan (1976), 153, 12-18 .