FOUNDATIONS AND COMPARISON OF INFORMATION CRITERIA: AKAIKE AND BAYESIAN

Paulo César EMILIANO[1]
Mario Javier Ferrua VIVANCO[1]
Fortunato Silva de MENEZES[1]
Fabrício Goecking AVELAR[2]

▪ *ABSTRACT: The choice of the best model is an important stage on modelling data and, parsimony is one of the principles which should be taken into account. Despite of being widely used on this stage, the foundations of information criteria of Akaike (AIC) and Bayesian (BIC) have been little understood, in general. The AIC and BIC are information criteria that penalizes the likelihood, such that a parsimony model should be selected, and these concepts are based on concepts of information and entropy, which are of the key points for their complete understanding. Such concepts are explained such that the understanding these information criteria are complete. An application comparison between those two information criteria through Monte Carlo simulation were also made, and it was found those information criteria did not present a good performance for sample sizes of 100 and 150; nevertheless once the samples size increases, the information criteria enhance their performance, with BIC showing superior in comparison to AIC for large sample size (equal and greater than 5000). On application to real data, with samples of size 123, both information criteria provide the same results.*

KEYWORDS Akaike information criterion; entropy; Schwarz information criterion; Kullback-Leibler information; model selection.

1 Introduction

Many people have the gift of science, they are the scientists and pursue the understanding of phenomena that intrigues Mankind. However, the majority of

[1]Departamento de Ciências Exatas, Universidade Federal de Lavras – UFLA, Caixa Postal 3037, CEP: 37200-000, Lavras, MG, Brasil. E-mail: *pequenokaiser2002@yahoo.com.br / ferrua@dex.ufla.br / fmenezes@dex.ufla.br*
[2]Departamento de Ciências Exatas, Universidade Federal de Alfenas – UNIFAL, CEP: 37130-000, Alfenas, MG, Brasil. E-mail: *fabricio@unifal-mg.edu.br*

people do not study these phenomena, because they think the phenomena are complicated enough or because they do not have access of information to understand them. It relies, then, on scientists to provide the information to explain the phenomena to the layman on the simplest possible way.

In general, a phenomena in study may be explained through a model, which is the main tool used on statistics, providing a simplified version of a given problem (or situation) in real life which illustrates certain aspects of the problem, without binding to all its details.

However in practical situations, the total knowledge of the phenomena do not happen. Usually, the phenomena observed are rather complex and it is impractical to describe them in all their extension and exactness. Due to the difficulties on the exact description of the observed phenomena on the way of symbolisms and mathematical formulas, statistical models, that embodies a random part and a systematic part, are used.

In the representation of a phenomena by a probabilistic model there is lost of information. In order to not compromise the understanding of the phenomena in study this information loss, due the probabilistic model used, must be minimum.

Quite often, more than one model may describe the same phenomena, since there is no recipe to be followed; each research has the freedom to model the phenomena following the methodology he judges more adequate. In this way, facing with two (or more) models it is natural to ask: "Which model, among all used, are more adequate?". The concept of best model is controversy, but a good model shall balance the quality of adjustment and the complexity, which in general is measured by the number of parameters in the model; the more parameters the model has, more complex the model is, making it more difficult to explain. The selection of the "best" model then arises naturally.

Burnham and Anderson (2004), emphasized the importance to select models based on scientific principles. A variety of methodologies are used to this end, such as Mallows C_p, stepwise regression, Akaike information criterion (AIC), Bayesian information criterion (BIC), generalized information criterion (GIC), among others.

On the AIC and BIC information criteria, for each model it is obtained a value and the model which presents the smaller AIC (or BIC) value is regarded as the "best" model. A question which then arise naturally is:"Why the criterion with smaller AIC (or BIC) value is selected?".

The purpose of this work is to illustrate and to compare the AIC and BIC criteria. It is expected the methodology used on these criteria be understood, to allow its use with full security and correct interpretation of the result obtained.

2 Criteria to model selection

A model is a simplified version of some problem or situation in real life, and its purpose is to illustrate some aspects of the same problem without taking into account all and every detail. Before the model construction it is needed to bear in mind that true models do not exist. There is only approximate models of the

reality that causes, inevitably, information loss. It is then necessary to minimize such loss. George Box made a famous statement about this: "All models are wrong, but some are useful".[1]. In such way, it is necessary to make a selection of "best" model, among those that were adjusted, to explain the phenomena under study.

According Mazerolle (2004), model selection is the task to choose a statistical model from a set of plausible models. In its basic form, this is one of the fundamentals tasks of the scientific research. From the wide variety of plausible models that we could adjust with the data provided, how is it possible to choose a good model?. It would be naive to expect the best result embodies all the variables of the model. This would violate the scientific principle founded on parsimony, which require that among all models that explain well the data, the simplest one should be chosen. Then, it is necessary to find the simplest model that explain the phenomena in study.

In order to quantify the information loss when we adjust the model, there are many diverse proposals on literature. Such as:

1- The Statistics χ^2, given by:

$$\chi^2 = \sum_{i=1}^{k} \frac{g_i^2}{f_i} - 1 = \sum_{i=1}^{k} \frac{(f_i - g_i)^2}{f_i}.$$

2- The Hellinger distance, given by:

$$I_K(g; f) = \int \left\{ \sqrt{f(x)} - \sqrt{g(x)} \right\}^2 dx.$$

3- The generalized information, given by:

$$I_\lambda(g; f) = \frac{1}{\lambda} \int \left\{ \left(\frac{g(x)}{f(x)} \right)^\lambda - 1 \right\} g(x)\, dx.$$

4- The Deviance criterion, given by:

$$D(\psi) = -2 \left[\log L(\psi; x) - \log L\left(\widehat{\psi}; x\right) \right],$$

where ψ is the parametric space and $\widehat{\psi}$ is the restrict space.

5- The divergence, given by:

$$D(g; f) = \int u(t(x)) g(x)\, dx = \int u\left(\frac{g(x)}{f(x)} \right) g(x)\, dx,$$

where $t(x) = \dfrac{g(x)}{f(x)}$.

6- The $L^1 - norm$, given by:

$$L_1(g; f) = \int |g(x) - f(x)| dx.$$

[1] "All models are wrong but some are useful"(Draper e Smith, 1998)

7- The $L^2 - norm$, given by:

$$L_2(g; f) = \int \{g(x) - f(x)\}^2 dx.$$

8- The Kullback-Leibler information, given by:

$$I(g; f) = E_g\left[\log\left(\frac{g(X)}{f(X)}\right)\right] = \int_{-\infty}^{+\infty} g(x) \log\left(\frac{g(x)}{f(x|\theta)}\right) dx, \quad (1)$$

where f, g, f_i and g_i are any distribution functions, $\lambda \in \mathbb{R}_+^*$, $x \in \mathbb{R}$ e $u(x) \geq 0$ a function.

According Mazerolle (2004), Kullback and Leibler define this measure, called later Kullback-Leibler information (K-L), to represent the information loss due the approximation of model adjusted from the reality. Many authors show this measure is a natural measure to discriminate between two probability functions.

The equation (1) may be expressed as:

$$I(g; f) = \int_{-\infty}^{+\infty} g(x) \log[g(x)] dx - \int_{-\infty}^{+\infty} g(x) \log[f(x|\theta)] dx. \quad (2)$$

In this way, for two models $f_1(x|\theta)$ e $f_2(x|\theta)$, it is possible to obtain from (1) that

$$I(f_1, g) = \int g(x) \log(g(x)) dx - \int g(x) \log(f_1(x|\theta)) dx$$

and

$$I(f_2, g) = \int g(x) \log(g(x)) dx - \int g(x) \log(f_2(x|\theta)) dx.$$

Only the second term of equation (2) is important on evaluation of statistical models $f_1(x)$ and $f_2(x)$, since the first depends solely on the true models g. However, the second term also depends of the unknown distribution g, being represented by

$$E_g[\log f(X)] = \int g(x) \log(f(x|\theta)) dx. \quad (3)$$

According Akaike (1974), the K-L information is appropriate to test if a given model is adequate. However, its use is limited since it depends of the distribution g, which is unknown. If a good estimation for the expectation log likelihood can be obtained through the data, this estimation could be used as a criterion to compare the models.

Under certain circumstances of regularity the Maximum Likelihood Estimator (MLE) is asymptotically efficient, since the likelihood function tends to be more sensitive to small deviations of the model parameters from their true values. The maximum likelihood estimators constitute good estimators due its asymptotically properties.

Firstly, the maximum likelihood estimators are used to estimate the vector parameter θ in $f(x|\theta)$, obtained then $f(x|\widehat{\theta})$ that it will be used in (3) to find

$$E_g\left[\log f\left(X\right)\right] = \int g\left(x\right) \log\left(f\left(x|\widehat{\theta}\right)\right) dx. \tag{4}$$

Lately, with the maximum likelihood estimators as base estimators, the purpose is to find a good estimator to (4). According Konishi and Kitagawa (2008), one estimation of the expected support function may be obtained replacing the unknown probability distribution function G, on equation (4), by a empirical distribution function $\widehat{G}$ based on data.

In order to find such estimation, the following definitions are required:

Definition 2.1. Let $A = \{x_1, x_2, ..., x_n\}$ be the set of observed data from a distribution $G(x)$. The empirical distribution function $\widehat{G}$ is a accumulated density function which provides $\frac{1}{n}$ of probability for each x_i. Formally,

$$\widehat{G}_n\left(x\right) = \frac{1}{n}\sum_{i=1}^{n} I\left(x_i \leq x\right)$$

where

$$I\left(x_i \leq x\right) = \begin{cases} 1, \ se \ x_i \leq x \\ 0, \ se \ x_i > x. \end{cases}$$

Definition 2.2. A statistical functional $T(G)$ is any function of G, where G is a distribution and T an arbitrary function.

Definition 2.3. The estimator for $\theta = T\left(G\right)$ is defined by $\widehat{\theta}_n = \widehat{G}_n$.

If a functional can be written on the form $T\left(G\right) = \int u\left(x\right) dG\left(x\right)$, Konishi and Kitagawa (2008) showed that the correspondent estimator is given by

$$T\left(\widehat{G}\right) = \int u\left(x\right) d\widehat{G}\left(x\right) = \sum_{i=1}^{n} \widehat{g}\left(x_i\right) u\left(x_i\right) = \frac{1}{n}\sum_{i=1}^{n} u\left(x_i\right) \tag{5}$$

in other words, a replacement of the accumulated probability density function G by the empirical accumulated distribution $\widehat{G}$ is made, and the density function $\widehat{g}_n = \frac{1}{n}$ for each observation x_i.

From equation (5), it is possible estimate the expected support function by:

$$\begin{aligned} E_{\widehat{G}}\left[\log f\left(x|\widehat{\boldsymbol{\theta}}\right)\right] &= \int \log f\left(x|\widehat{\boldsymbol{\theta}}\right) d\widehat{G}\left(x\right) \\ &= \sum_{i=1}^{n} \widehat{g}\left(x_i|\widehat{\boldsymbol{\theta}}\right) \log f\left(x_i\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} \log f\left(x_i|\widehat{\boldsymbol{\theta}}\right). \end{aligned}$$

2.1 Information criteria

One way to compare n models, $g_1(x|\boldsymbol{\theta_1})$, $g_2(x|\boldsymbol{\theta_2})$, ..., $g_n(x|\boldsymbol{\theta_n})$, is simply to compare the magnitude of the maximized support function $L\left(\widehat{\boldsymbol{\theta}}_{\boldsymbol{i}}\right)$. However, such method does not provide a true comparison, since as the true model $g(x)$ is not known, firstly the method of maximum likelihood estimates the parameters $\boldsymbol{\theta_i}$ of each model $g_i(x), i = 1, 2, ...n$, and lately the same data are used to estimate $E_G\left[logf(x|\widehat{\boldsymbol{\theta}})\right]$, what introduce a bias in $L\left(\widehat{\boldsymbol{\theta}}_{\boldsymbol{i}}\right)$, where the magnitude of this bias varies according the dimension of the vector of parameters.

The bias is given by

$$b(G) = E_{G(\boldsymbol{x_n})}\left[\log f\left(\boldsymbol{X_n}|\widehat{\boldsymbol{\theta}}(X_n)\right) - n\, E_{G(Z)}\left[\log f\left(Z|\widehat{\boldsymbol{\theta}}(X_n)\right)\right]\right], \qquad (6)$$

where the expectation is taken with respect the joint distribution.

In this way, the information criteria are built to evaluate and correct the bias of the support function. According Koniski and Kitagawa (2008), the information criterion has the following form:

$$\begin{aligned} CI\left(\boldsymbol{X_n}, \widehat{G}\right) &= -2\sum_{i=1}^{n} \log f\left(X_i|\widehat{\boldsymbol{\theta}}(X_n)\right) + 2\,(\text{bias}) \\ &= -2\sum_{i=1}^{n} \log f\left(X_i|\widehat{\boldsymbol{\theta}}(X_n)\right) + 2\,(b(G)). \end{aligned} \qquad (7)$$

It may also be used others criteria common in the literature to model selection. These criteria consider the complexity of the model in the selection criterion and, essentially, penalize the likelihood, using the number of variables of the model and, eventually the sample size. The penalization is made subtracting from the value of the likelihood a given quantity, which measures how complex the model is (the model is more complex as much parameters it presents).

Akaike (1974), proposed to use the Kullback-Leibler information to select models, establishing a relation between the maximum likelihood and Kullback-Leibler information, and developing then a criterion to estimate the Kullback-Leibler information, which was called lately, the Akaike information criterion (AIC).

Criteria for models selection, as the Akaike information criterion (AIC) and the Bayesian information criterion (BIC), are used often to select models in a variety of areas. According to these criteria, the best model is the one which present the smaller value of AIC or BIC.

As these results are asymptotic, the results on this work are valid for "large" sample sizes, where the concept of "large" samples being difficult to define, as such concept depends on the area of study, the availability of resources for a larger sample, among others factors.

2.2 Akaike information criterion - AIC

The Akaike information criterion (AIC, developed by Hurotugu Akaike with the name of "an information criterion" in 1971 and proposed by Akaike (1974), is a relative measure of quality of adjustment of a estimate statistical model. It is founded by the concept of information and it offers a relative measure of information loss, when a given model is used to describe the reality. Akaike found a relation between the relative expectation of K-L information and the maximized support function, allowing more interaction between the theory and practice, on models selection and analyses of complex data set.

Akaike (1974), showed the bias is given, asymptotically, by:

$$b\left(G\right) = tr\left\{I\left(\boldsymbol{\theta}_0\right) J\left(\boldsymbol{\theta}_0\right)^{-1}\right\}, \tag{8}$$

where $J\left(\boldsymbol{\theta_0}\right)$ the Fisher information matrix and, $I\left(\boldsymbol{\theta}_0\right)$, given by

$$I\left(\boldsymbol{\theta_0}\right) = \int g\left(x\right) \frac{\partial f\left(x|\theta\right)}{\partial \boldsymbol{\theta}} \frac{\partial f\left(x|\boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta^T}}\, dx.$$

The AIC is a criterion which evaluates the quality of adjustment of the parametric model, estimated by the method of maximum likelihood, and it has its foundations with the result the bias (8) tends to the number of parameters to be estimate on the model, since under the assumption it exists $\boldsymbol{\theta}_0 \in \Theta$ such that $g(x) = f(x|\boldsymbol{\theta}_0)$, it occurs the equality of expressions $I(\boldsymbol{\theta}_0) = J(\boldsymbol{\theta}_0)$ and then it is obtained from (8) that:

$$\begin{aligned} b\left(G\right) &= E_{G(\boldsymbol{x_n})}\left[\log f\left(\boldsymbol{X_n}|\widehat{\boldsymbol{\theta}}\left(X_n\right)\right) - n\, E_{G(Z)}\left[\log f\left(Z|\widehat{\boldsymbol{\theta}}\left(X_n\right)\right)\right]\right] \\ &= tr\left\{I\left(\boldsymbol{\theta_0}\right) J\left(\boldsymbol{\theta_0}^{-1}\right)\right\} = tr\left(I_p\right) = p, \end{aligned} \tag{9}$$

where p is the number of parameters to be estimated on the model.

With this results, Akaike (1974) defined his information criterion as:

$$AIC = -2\left(\text{Maximized support function}\right) + 2\left(\text{number of parameters}\right),$$

$$AIC = -2\log L\left(\widehat{\boldsymbol{\theta}}\right) + 2\left(p\right) \tag{10}$$

2.3 Bayesian information criterion - BIC

The Bayesian Information Criterion (BIC), also called Schwarz criterion, was proposed by Schwarz (1978), and it is a criterion to evaluate models defined in terms of a posterior probability, being so called because Schwarz gave an Bayesian argument to prove it. In the following, some concepts will be presented that it will end up with the proof given by Schwarz.

Definition 2.4. Two sets A and B, which are subsets of Ω, are defined as mutually exclusive (disjoined) if $A \cap B = \emptyset$. Subsets $A_1, A_2, \ldots$ are said mutually exclusive if $A_i \cap A_j = \emptyset$ for all $i \neq j$, $i, j \in \mathbb{N}$.

Theorem 2.5. *If $(\Omega, \mathcal{A}, P[\cdot])$ is a probability space and $B_1, B_2, \ldots, B_n$ is a collection of events mutually exclusive in $\mathcal{A}$, which satisfy $\Omega = \bigcup_{j=1}^{n} B_j$ and $P[B_j] > 0$, for $j = 1, 2, \ldots, n$, then for all $A \in \mathcal{A}$, such that $P[A] > 0$, it results:*

$$P[B_k|A] = \frac{P[A|B_k]\, P[B_k]}{\sum_{j=1}^{n} P[A|B_k]\, P[B_k]}, \tag{11}$$

with Ω the amostral space and $\mathcal{A}$ the parametrical space.

According Konishi and Kitagawa (2008), let $M_1, M_2, \ldots, M_k$, be k candidate models, each one of models M_i with one probability distribution $f_i(x|\theta_i)$ and one priori, $\pi_i(\boldsymbol{\theta_i})$ for the k_ith vector θ_i. If it exists n observations $\boldsymbol{x_n} = \{x_1, x_2, \ldots, x_n\}$, then for the ith model M_i, the marginal distribution $\boldsymbol{x_n}$ is given by:

$$p_i(x_n) = \int f_i(\boldsymbol{x_n}|\boldsymbol{\theta_i})\, \pi_i(\boldsymbol{\theta_i})\, d\boldsymbol{\theta_i}. \tag{12}$$

This quantity may be regard as the likelihood for the ith model and it will be refer as the marginal likelihood of data.

With $P(M_i)$ the priori distribution of ith model, from equation (11) the posterior distribution will be:

$$P(M_i|\boldsymbol{x_n}) = \frac{p_i(\boldsymbol{x_n})\, P(M_j)}{\sum_{j=1}^{n} p_j(\boldsymbol{x_n})\, P(M_j)} \tag{13}$$

The posterior probability indicates the probability of data to be generated from ith model when data $\boldsymbol{x_n}$ are observed. If a model is selected from r models, it is natural to adopt the model which has the highest posterior probability. This principle shows that the model which maximizes the numerator $p_j(\boldsymbol{x_n})\, P(M_j)$ must be selected, since all models share the same denominator in equation (13).

If the priori distributions $P(M_i)$ are equals in all models, then the model which maximize the marginal probability of data $p_i(\boldsymbol{x_n})$, must be selected. Therefore, if an approximation for the marginal probability express in terms of the integral in equation (13) could be obtained, then the basic need of finding the integral in each problem disappears, what makes the BIC criterion satisfactory for models selection.

The BIC is defined as:

$$\begin{aligned} -2\, \log(p_i(\boldsymbol{x_n})) &= -2\, \log \int f_i(\boldsymbol{x_n}|\boldsymbol{\theta_i})\, \pi_i(\boldsymbol{\theta_i})\, d\boldsymbol{\theta_i} \\ &\approx -2\, \log\left(f_i\left(\boldsymbol{x_n}|\widehat{\boldsymbol{\theta_i}}\right)\right) + k_i\, \log(n) \end{aligned} \tag{14}$$

where $\widehat{\boldsymbol{\theta}}_{\boldsymbol{i}}$ is the maximum likelihood estimator for the k_ith parametric vector θ_i of model $f_i\left(\boldsymbol{x_n}|\boldsymbol{\theta_i}\right)$.

Consequently, from the r models evaluated using the maximum likelihood method, the model which minimized the BIC is the best model for the data.

Therefore, under the assumption that all models have equal prior probability distributions, the posterior probability, obtained using the information from the data, it is used to contrast different models and helps to identify the model which generates the data.

Consider M_1 and M_2 two models we wish to compare. For each model it exists the marginal likelihoods $p_i(\boldsymbol{x_n})$, the priors $P(M_i)$ and the posteriors $P(M_i|\boldsymbol{x_n})$ with $i = \{1, 2\}$. Then, the ratio of posterior in favor of model M_1 versus the model M_2 is:

$$\frac{P\left(M_1|\boldsymbol{x_n}\right)}{P\left(M_2|\boldsymbol{x_n}\right)} = \frac{\frac{p_1(\boldsymbol{x_n})P(M_1)}{\sum_{j=1}^{n} p_j(\boldsymbol{x_n})P(M_j)}}{\frac{p_2(\boldsymbol{x_n})P(M_2)}{\sum_{j=1}^{n} p_j(\boldsymbol{x_n})P(M_j)}} = \frac{p_1\left(\boldsymbol{x_n}\right)P\left(M_1\right)}{p_2\left(\boldsymbol{x_n}\right)P\left(M_2\right)}.$$

The ratio

$$\frac{p_1\left(\boldsymbol{x_n}\right)}{p_2\left(\boldsymbol{x_n}\right)} \tag{15}$$

is called the *Bayes Factor*.

The problem of finding the value of BIC lies on evaluate the integral in equation (12). This is done using the Laplace approximation for integrals.

- **Laplace approximation for integrals**

  It is need a Laplace approximation for the integral

$$\int \exp\left\{nq\left(\boldsymbol{\theta}\right)\right\} d\boldsymbol{\theta}, \tag{16}$$

where $\boldsymbol{\theta}$ is a $p-$dimensional vector of parameters and $q(\boldsymbol{\theta})$ is a real function $p-$dimensional.

The great advantage of Laplace approximation lies on the fact that when the number of observations is large, the integrant concentrate in a neighborhood $\widehat{\boldsymbol{\theta}}$ of $q(\boldsymbol{\theta})$ and, consequently, the value of integral depends only on the behavior of the integrand on the neighborhood of $\widehat{\boldsymbol{\theta}}$.

Therefore, $\left.\frac{\partial q(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} = 0$ and the expansion of $q\left(\boldsymbol{\theta}\right)$ around of $\widehat{\boldsymbol{\theta}}$ is:

$$q\left(\boldsymbol{\theta}\right) = q\left(\widehat{\boldsymbol{\theta}}\right) - \frac{1}{2}\left(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\right)^T J_q\left(\widehat{\boldsymbol{\theta}}\right)\left(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\right) + \ldots, \tag{17}$$

where

$$J_q\left(\widehat{\boldsymbol{\theta}}\right) = -\left.\frac{\partial^2 q\left(\boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} \tag{18}$$

Definition 2.6. If $q(\boldsymbol{\theta})$ is a real value function evaluated in the neighborhood of $\widehat{\boldsymbol{\theta}}$, and $\boldsymbol{\theta}$ is a vector of parameters, then the **Laplace approximation** for the integral is given by:

$$\int \exp\{nq(\boldsymbol{\theta})\, d\boldsymbol{\theta}\} \approx \frac{(2\pi)^{p/2}}{(n)^{p/2}\left|J_q\left(\widehat{\boldsymbol{\theta}}\right)\right|^{p/2}} \exp\left(nq\left(\widehat{\boldsymbol{\theta}}\right)\right) \tag{19}$$

where $J_q\left(\widehat{\boldsymbol{\theta}}\right)$ is defined in equation (18).

Using the Laplace approximation to approximate equation (12), it can be rewritten as

$$\begin{aligned} p(x_n) &= \int f_i(\boldsymbol{x_n}|\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})\, d\boldsymbol{\theta} \\ &= \int \exp\{\log f(\boldsymbol{x_n}|\boldsymbol{\theta})\}\pi(\boldsymbol{\theta})\, d\boldsymbol{\theta} \\ &= \int \exp\{\ell(\boldsymbol{\theta})\}\pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}, \end{aligned} \tag{20}$$

where $\ell(\boldsymbol{\theta})$ is the support function $\ell(\boldsymbol{\theta}) = \log(f(x_n|\boldsymbol{\theta}))$.

Therefore, performing the Taylor series expansion of $\ell(\boldsymbol{\theta})$ and $\pi(\boldsymbol{\theta})$ around $\widehat{\boldsymbol{\theta}}$ we obtain, respectively:

$$\ell(\boldsymbol{\theta}) = \ell\left(\widehat{\boldsymbol{\theta}}\right) - \frac{1}{2}\left(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\right)^T J\left(\widehat{\boldsymbol{\theta}}\right)\left(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\right) + \ldots, \tag{21}$$

$$\pi(\boldsymbol{\theta}) = \pi\left(\widehat{\boldsymbol{\theta}}\right) + \left(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\right)^T \left.\frac{\partial \pi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} + \ldots, \tag{22}$$

and replacing equations (21) and (22) in equation (20) we obtain:

$$\begin{aligned} p(x_n) = & \int \exp\left\{\pi\left(\widehat{\boldsymbol{\theta}}\right) + \left(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\right)^T \left.\frac{\partial \pi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} + \ldots\right\} d\boldsymbol{\theta} \\ & \times \left\{\pi\left(\widehat{\boldsymbol{\theta}}\right) + \left(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\right)^T \left.\frac{\partial \pi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} + \ldots\right\} d\boldsymbol{\theta} \\ & \approx \exp\left\{\ell\left(\widehat{\boldsymbol{\theta}}\right)\right\}\pi\left(\widehat{\boldsymbol{\theta}}\right) \int \exp\left\{-\frac{n}{2}\left(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\right)^T J(\boldsymbol{\theta})\left(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\right)\right\} d\boldsymbol{\theta} \end{aligned} \tag{23}$$

The integral in equation (23) satisfy the equation (19), consequently it can be approximated using the Laplace integral, and it is obtained:

$$\int \exp\left\{-\frac{n}{2}\left(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\right)^T J(\boldsymbol{\theta})\left(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\right)\right\} d\boldsymbol{\theta} = (2\pi)^{p/2}\, n^{-p/2}\left|J\left(\widehat{\boldsymbol{\theta}}\right)\right|^{-1/2}, \tag{24}$$

where the integrand is a $p-$dimensional normal density function with average vector $\widehat{\boldsymbol{\theta}}$ and covariance matrix $J^{-1}\left(\widehat{\theta}\right)/n$.

For large values of n,

$$p(x_n) \approx \exp\left\{\ell\left(\widehat{\boldsymbol{\theta}}\right)\right\} \pi\left(\widehat{\boldsymbol{\theta}}\right) (2\pi)^{p/2}\, n^{-p/2} \left|J\left(\widehat{\boldsymbol{\theta}}\right)\right|^{-1/2} \tag{25}$$

Taking the logaritm of equation (25) and multipling by the factor -2, we obtain

$$\begin{aligned} -2\ \log p\left(x_n\right) &= -2\ \log\left\{\int f\left(x_n|\boldsymbol{\theta}\right) \pi\left(\boldsymbol{\theta}\right) d\boldsymbol{\theta}\right\} \qquad (26)\\ &= -2\,\ell\left(\widehat{\boldsymbol{\theta}}\right) + p\ \log(n) + \log\left|J\left(\widehat{\boldsymbol{\theta}}\right)\right| - p\ \log\left(2\pi\right) - 2\ \log\pi\left(\widehat{\boldsymbol{\theta}}\right) \end{aligned}$$

Therefore, the Bayesian Information Criterion can be obtained on the following form (negleting the constant terms in the equation):

Definition 2.7. Let $F\left(\boldsymbol{x_n}|\widehat{\boldsymbol{\theta}}\right)$ be a statistical model estimated through the maximum likelihood method. Then, the Bayesian Information Criterion (BIC) is given by:

$$BIC = -2\ \log\left(f\left(x_n|\boldsymbol{\theta}\right)\right) + p\ \log(n), \tag{27}$$

where $f\left(x_n|\boldsymbol{\theta}\right)$ is the chosen model, p is the number of parameters to be estimated and n is the number of sample observations.

2.4 Simulation for equality of averages and/or deviations of normal distributions.

Let be two set of data $\{y_1, y_2, \ldots, y_n\}$ and $\{y_{n+1}, y_{n+2}, \ldots, y_{n+m}\}$, where $y_1, y_2, \ldots, y_n \sim N\left(\mu_1, \sigma_1^2\right)$ and $y_{n+1}, y_{n+2}, \ldots, y_{n+m} \sim N\left(\mu_2, \sigma_2^2\right)$. The purpose is to verify whether:

$$\mu_1 = \mu_2 = \mu \quad \text{and} \quad \sigma_1^2 = \sigma_2^2 = \sigma^2 \ \text{ or} \tag{28}$$

$$\mu_1 \neq \mu_2 \quad \text{and} \quad \sigma_1^2 \neq \sigma_2^2 \ \text{ or} \tag{29}$$

$$\mu_1 \neq \mu_2 \quad \text{and} \quad \sigma_1^2 = \sigma_2^2 = \sigma^2 \ \text{ or} \tag{30}$$

$$\mu_1 = \mu_2 = \mu \quad \text{and} \quad \sigma_1^2 \neq \sigma_2^2. \tag{31}$$

It was evaluated through 1000 Monte Carlo simulations the performance of AIC and BIC criteria for normal distributions on the conditions described on equations (28), (29), (30) and (31), for sample sizes of n = 100, 150, 1000, 5000, 10000, 15000 and 20000. The performance here means, the percentage the criterion in question selected the model correctly.

On the simulations were considered the following values of parameters:

For the situation on (28) $\mu_1 = \mu_2 = 1$ and $\sigma_1^2 = \sigma_2^2 = 1$;

For the situation on (29) $\mu_1 = 1$, $\mu_2 = 0.9$, $\sigma_1^2 = 1$ and $\sigma_2^2 = 0.9$;

For the situation on (30) $\mu_1 = 1$, $\mu_2 = 0.9$ and $\sigma_1^2 = 1 = \sigma_2^2 = 1$;

For the situation on (31) $\mu_1 = \mu_2 = 1$, $\sigma_1^2 = 0.9$ and $\sigma_2^2 = 1$.

All the results are summarized on Tables 1 to 4 below:

Table 1 - Percentage of correctness of AIC and BIC for the case described on (28)

Criterion	n=100	n=150	n=1000	n=5000	n=10000	n=15000	n=20000
AIC	67.88	68.16	68.64	67.02	68.08	68.84	67.96
BIC	95.06	95.98	98.60	99.50	99.74	99.70	99.70

Table 2 - Percentage of correctness of AIC and BIC for the case described on (29)

Criterion	n=100	n=150	n=1000	n=5000	n=10000	n=15000	n=20000
AIC	34.82	44.82	97.08	100	100	100	100
BIC	9.02	11.92	71.8	100	100	100	100

Table 3 - Percentage of correctness of AIC and BIC for the case described on (30)

Criterion	n=100	n=150	n=1000	n=5000	n=10000	n=15000	n=20000
AIC	20.28	20.58	64.52	84.92	83.8	83.6	84.16
BIC	5.30	5.56	29.40	97.4	99.72	99.92	99.88

Table 4 - Percentage of correctness of AIC and BIC for the case described on (31)

Critério	n=100	n=150	n=1000	N=5000	n=10000	n=15000	n=20000
AIC	38.2	46.28	91.24	94.7	93.86	94.34	94.34
BIC	12.8	15.08	74.34	100	100	100	100

From Tables 1 to 4 we can notice that for the samples of size 100 and 150, the performance of both criteria was poor, selecting a low percentage of the true model from where it was simulated the data. However, the performance increases, as the samples size increases, where the BIC presented a better performance in comparison with AIC for large samples size ($n \geq 5000$).

3 Application

In this work, it was analyzed a data set of 123 scores from 6^{th} year students of Itália Cautieiro Franco School (CAIC) at city of Lavras, Minas Gerais, Brazil on the disciplines of geography and portuguese, on the years 2002 (geography) and 2006 (portuguese). After the application of Shapiro-Wilks test, the purpose is, through the AIC and BIC criteria, to verify whether the scores follow a normal distribution with the same average and same deviation, same average and different deviations, different average and same deviation and both average and deviation different.

3.1 Equality of averages and/or deviations of normal distributions.

An application to real data of samples size $n_1 = 123$ and $n_2 = 123$ is given in the following. It shall be stress that samples size similar to this application are used on a variety of researches.

One utility of Akaike and Schwartz criteria is to test whether the data originates from a normal distribution with same average and same deviation; or same average and different deviations, or different averages and same deviation or a normal distribution with different averages and different deviations.

For the case described on (28) the AIC is given by:

$$AIC_1 = (n+m)\left(\log\left(\widehat{\sigma}^2\right) + \log(2\pi) + 1\right) + 4, \tag{32}$$

and

$$BIC_1 = (n+m)\left(\log\left(\widehat{\sigma}^2\right) + \log(2\pi) + 1\right) + 2\log(n+m), \tag{33}$$

where $\widehat{\mu}$ and $\widehat{\sigma}^2$ are given by:

$$\widehat{\mu} = \frac{1}{n+m}\sum_{i=1}^{n+m} y_i \tag{34}$$

$$\widehat{\sigma}^2 = \frac{1}{n+m}\sum_{i=1}^{n+m} (y_i - \widehat{\mu})^2. \tag{35}$$

For the case described on (29) we have

$$AIC_2 = (n+m)\left(\log(2\pi) + 1\right) + n\log\left(\widehat{\sigma_1^2}\right) + m\log\left(\widehat{\sigma_2^2}\right) + 8, \tag{36}$$

e

$$BIC_2 = (n+m)\left(\log(2\pi) + 1\right) + n\log\left(\widehat{\sigma_1^2}\right) + m\log\left(\widehat{\sigma_2^2}\right) + 4\log(n), \tag{37}$$

Where $\widehat{\mu_1}$, $\widehat{\mu_2}$, $\widehat{\sigma_1^2}$ and $\widehat{\sigma_2^2}$ are given, respectively by (38), (39), (40) and (41).

$$\widehat{\mu_1} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{38}$$

$$\widehat{\mu_2} = \frac{1}{m}\sum_{i=n+1}^{n+m} y_i \tag{39}$$

$$\widehat{\sigma_1^2} = \frac{1}{n}\sum_{i=1}^{n} (y_i - \widehat{\mu_1})^2 \tag{40}$$

$$\widehat{\sigma_2^2} = \frac{1}{m}\sum_{i=1}^{n} (y_i - \widehat{\mu_2})^2. \tag{41}$$

For the case described on (30) we have

$$AIC_3 = (n+m)\log\left(\widehat{\sigma^2}\right) + (n+m)\left(\log(2\pi)+1\right) + 6 \tag{42}$$

and

$$BIC_3 = (n+m)\log\left(\widehat{\sigma^2}\right) + (n+m)\left(\log(2\pi)+1\right) + 3\log(n). \tag{43}$$

Where the estimators of μ_1, μ_2, and σ^2 are given, respectively, by:

$$\widehat{\mu_1} = \frac{\sum_{i=1}^{n} y_i}{n} \tag{44}$$

$$\widehat{\mu_2} = \frac{\sum_{i=n+1}^{n+m} y_i}{m} \tag{45}$$

$$\widehat{\sigma^2} = \frac{1}{(n+m)}\left[\sum_{i=1}^{n} (y_i - \widehat{\mu_1})^2 + \sum_{i=n+1}^{n+m} (y_i - \widehat{\mu_2})^2\right]. \tag{46}$$

Finally, for the case described on (30) it was obtained:

$$AIC_4 = (n+m)\left(\log(2\pi)+1\right) + n\ \log\left(\widehat{\sigma_1^2}\right) + m\ \log\left(\widehat{\sigma_2^2}\right) + 6 \tag{47}$$

and

$$BIC_4 = (n+m)\left(\log(2\pi)+1\right) + n\log\left(\widehat{\sigma_1^2}\right) + m\ \log\left(\widehat{\sigma_2^2}\right) + 3\log(n) \tag{48}$$

where

$$\widehat{\sigma_1^2} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{\mu})^2 \tag{49}$$

$$\widehat{\sigma_2^2} = \frac{1}{m}\sum_{i=n+1}^{n+m}(y_i - \widehat{\mu})^2 \tag{50}$$

and the estimator of μ is found solving the $3rd$ degree equation, given by

$$\widehat{\mu^3} + A\widehat{\mu^2} + B\widehat{\mu} + C = 0 \tag{51}$$

where A, B and C are given, respectively, by:

$$A = -\left(\mu_1\left(1+v\right) + \left(1+w\right)\mu_2\right) \tag{52}$$

$$B = \left(2\mu_1\mu_2 + vs_1^2 + ws_2^2\right) \tag{53}$$

$$C = -\left(\mu_1 w s_2^2 + v\mu_2 s_1^2\right) \tag{54}$$

and

$$w = \frac{n}{m+n} \quad , \quad v = \frac{m}{m+n},$$
$$\mu_1 = \frac{\sum_{i=1}^{n} y_i}{n}, \mu_2 = \frac{\sum_{i=n+1}^{n+m} y_i}{m} \quad , \quad s_1^2 = \frac{\sum_{i=1}^{n} y_i^2}{n}, s_2^2 = \frac{\sum_{i=n+1}^{n+m} y_i^2}{m}. \tag{55}$$

For all data analized it was obtained the Tables 5 and 6 showed below:

Table 5 - Values of AIC obtained from scholar data

AIC	Value
AIC_1	2091.52
AIC_2	2093.72
AIC_3	2093.34
AIC_4	2091.79

The analysis of Table 5 shows that through the Akaike criterion the scores of portuguese and geography were considered as originating from normal distributions with same average and same deviation.

Using the BIC to select the best normal distribution which fits the data, can be seen, from Table 6 that as the AIC, this criterion selected the normal distribution with the same average and same deviation as the one which represent the data.

Table 6 - Values of BIC obtained from the scholar data

BIC	Value
BIC_1	2098.53
BIC_2	2107.74
BIC_3	2103.86
BIC_4	2102.31

Conclusions

The purpose of this work was to explain the information criteria of Akaike and Schwarz, to present two explanations currently used and to evaluate the their performance through Monte Carlo simulation for different samples sizes.

In the study of simulation to select the models of normal distribution both AIC and BIC presented a poor performance. For large samples size, both criteria had an increase in the percentage of selection of the correct model. For large samples size, BIC obtained better results in comparison to AIC.

In the application, the results obtained were the same results using both criteria, and selecting the scores as originating from the normal distribution with same average and same deviation.

Considering the results obtained through Monte Carlo simulation it can be seen that, both criteria had poor performance to samples size 100 and 150; in this way, researches that use samples size near these, shall have their results questioned.

Due the limitation in the use of both criteria in small samples size, the corrections of these criteria, already existent, may be a way out to this problem, as well as other measures of model selection, such as the L measure. The comparison among these corrections as well as the L measure will be object of a future work.

Acknowledgment

This work was developed at the Department of Exact Sciences, Universidade Federal de Lavras, and it was performed with funding of Fundação de Amparo de Pesquisa do Estado de Minas Gerais - FAPEMIG.

EMILIANO, P. C.; VIVANCO, M. J. F; MENEZES, F. S. M.; AVELAR, F. G. Fundamentos e comparacao de criterios de informacao: Akaike and Bayesian. *Rev. Bras. Biom.*, São Paulo, v.27, n.3, p.394-411, 2009.

▪ RESUMO: A escolha do melhor modelo é uma etapa importante na modelagem dos dados e a parcimônia é um dos princípios que devem ser levados em consideração. Apesar de serem amplamente utilizados nesta etapa, os critérios de informação de Akaike (AIC) e Bayesiano (BIC) têm seus fundamentos pouco entendidos, em geral. O AIC e o BIC são critérios que penalizam a verossimilhança, para que um modelo mais parcimonioso seja selecionado, e estes conceitos estão baseados nos conceitos de informação e entropia, que são de fundamental importância para o completo entendimento dos mesmos. Tais conceitos foram explicados para que o entendimento dos mesmos fosse completo. Foi feita também, uma aplicação comparando os dois critérios via simulação Monte Carlo, sendo que eles não apresentaram bom desempenho para amostras de tamanho 100 e 150, e à medida que o tamanho das amostras aumentou, os critérios melhoraram seu desempenho, sendo que o BIC foi superior ao AIC para amostras de tamanho grandes (maiores ou iguais que 5000). Na aplicação a dados reais, com amostras de tamanho 123 os critérios proporcionaram os mesmos resultados.

▪ PALAVRAS-CHAVE: Critério de informação de Akaike; entropia; critério de informação de Schwarz; informação de Kullback-Leibler; seleção de modelos.

References

AKAIKE, H. A new look at the statistical model identification. *IEEE Transactions on Automat. Contr.*, Boston, v.19, n.6, p.716–723, 1974.

BURNHAM, K. P.; ANDERSON, D. R. Multimodel inference: understanding aic and bic in model selection. *Sociol. Methods Res.*. Beverly Hills, v.33, n.2, p.261–304, 2004.

CHAKRABARTI, C. G.; CHAKRABARTY, I. Boltzmann entropy : probability and information. *Rom. J. Phys.*. Bucharest, v.52, n.5-6, p.525–528, Jan. 2007.

DRAPER, N. R.; SMITH, H. *Applied regression analysis*. 3. ed. New York: J. Wiley, 1998. 706p.

EMILIANO, P.C. *Fundamentos e aplicações dos critérios de informação: Akaike e Bayesiano*, 2009. 92f. Dissertação (Mestrado em Estatística e Experimentação Agronômica) - Universidade Federal de Lavras, Lavras, 2009.

GHOSH, J. K.; SAMANTA, T. Model selection - an overview. *Curr. Sci.*, Bangalore, v.80, n.9, p.1135–1144, 2001.

KONISHI, S.; KITAGAWA, G. *Information criteria and statistical modeling*. New York: Springer, 2008. 321p.

MARTINS, R. C. Sobre a atualidade de proposições de Ludwig Boltzmann. *Rev. SBHC*, São Paulo, n.13, p.81–94, 1995.

MAZEROLLE, M. J. *Mouvements et reproduction des amphibiens en tourbières perturbées* 2004. 78f. Thesis (Ph.D.) - Université Laval, Québec, 2004.

PEREIRA, T.L.; SOUZA, T.C.; CRIBARI-NETO, F. *Critérios de seleção de modelos: uma comparação*. In: SIMPÓSIO NACIONAL DE PROBABILIDADE E ESTATÍSTICA, 18., 2008, São Pedro. Anais ... São Pedro: ABE, 2008.

SCHWARZ, G. Estimating the dimensional of a model. *Ann. Stat.*, Hayward, v.6, n.2, p.461–464, 1978.

SHANNON, C. E. A mathematical theory of communication. *Bell System Tech. J.*, New York, v.27, n.3, p.623–656, 1948.

Received in 31.03.2009.

Approved after revised in 26.10.2009.