BIAS OF THE CORRECTED KIC FOR UNDERFITTED REGRESSION MODELS

Maiza Bekara * *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 [1] and the bias corrected version, $KICc$ [2] are two methods for statistical model selection of regression variables and autoregressive models. Both criteria may be viewed as estimators of the Kullback symmetric divergence between the true model and the fitted approximating model. The bias of KIC and $KICc$ is studied in the underfitting case, where none of the candidate models includes the true model. Here, only normal linear regression models are considered, where exact expression of the bias is obtained for KIC and $KICc$. The bias of $KICc$ is often smaller, in most case drastically smaller than KIC. A simulation study in which the true model is of inf inite order polynomial expansion shows that in small and moderate sample size $KICc$ provides a better model selection than KIC. Furthermore $KICc$ outperforms the two well-known criteria AIC and MDL.

1. INTRODUCTION

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 (within a constant) of the Kullback symmetric divergence, known also as the J- divergence between the generating model and the candidate fitted model. Since the logic behind this criterion is sound, one may hope that minimization of an *exactly* unbiased estimate of this divergence rather than an asymptotically unbiased estimate, will provide a good model selection. This idea has been put in a general framework by Linhart & Zucchini [3], who viewed model selection as the construction of an approximately unbiased estimator of the underlying target criterion function.

Simulation studies show that KIC produces a good model selection in large samples [1]. Nevertheless, KIC becomes strongly negatively biased for small samples of data or when the number of fitted parameters (of the candidate model) to the number of data gets large. Therefore, bias itself seems to be a property of KIC, that merits further investigation. In addition, one may hope that by improving the bias property, one will also improve the quality of the selected models. This indeed was the motivation behind the development of the corrected $KICc$ criterion proposed in [2] . $KICc$ is an *exactly* unbiased estimator of the Kullback symmetric divergence and not only produces drastic bias reduction, but also greatly improves the model selection in small samples.

For normal linear regression, or autoregressive models of k parameters (excluding the innovation variance σ^2), the KIC and $KICc$ are respectively defined as

$$KIC = n\left(\log(2\pi\hat{\sigma}^2) + 1\right) + 3(k+1), \tag{1}$$

$$\begin{aligned} KICc &= n\left(\log(2\pi\hat{\sigma}^2) + 1\right) + 2\frac{(k+1)n}{n-k-2} \\ &\quad -n\psi\left(\frac{n-k}{2}\right) + n\ln\frac{n}{2}, \end{aligned} \tag{2}$$

where $\hat{\sigma}^2$ is the estimate of the innovation variance for the fitted k-th model and n is the number of data. $\psi(.)$ is the *digamma* or the *psi* function [4].

In the derivation of $KICc$ as well as for KIC, the unbiasedness property (asymptotically or at finite sample) is restricted to the case where the true model is of finite dimension and is either *correctly specified* or *overfitted*. We say that a true model is correctly specified or overfitted if some configuration of parameter values in the candidate model, perhaps including some zero values, yields the true model. Otherwise the model is said to be *underfitted* and the candidate model is referred to as an approximating model [5]. In practice, however, since a variety of candidate models will be considered, it will often happens that the model is underfitted. Moreover, if the true model is of infinite dimension, which is the case of many typical example in practice, then none of the candidate models will be able to exactly reproduce the true model. Here, the true model will always be underfitted and we only hope to find a good approximating model.

In this paper, we study the bias properties and the model selection performance of KIC and $KICc$ in the underfitted case. We only consider the case of normal linear regression, where exact expression of the expected value of KIC, $KICc$ and the Kullback symmetric divergence are derived. The bias of KIC and $KICc$ depends on the true regression function, on the form and dimension of the candidate model and on the amount of noise. Numerical evaluation of the bias for the class of *polynomial* functions is presented. We find that although $KICc$ is not uniformly less biased than KIC, for many cases the expected $KICc$ and the expected Kullback symmetric divergence are minimized at the same model order, while that of the expected KIC is usually minimized at larger order. Furthermore as the ratio of the model order to the number of data increases, KIC becomes strongly negatively biased, while the bias of $KICc$ is much smaller. Finally we assess the quality of the models selected by $KICc$ and KIC in a polynomial regression when the true model is of infinite order polynomial expansion. For comparison purpose, we have included in the study the well-known model selection criterion: Akaike Information Criterion (AIC) [6] and the Minimum Description Length

* Corresponding author

(MDL) [7]. We find that $KICc$ significantly outperforms the other methods in terms of the quality of selected models as measured by the mean square error of approximation. These results strengthen the case for using $KICc$ instead of KIC as originally recommended in [2].

2. THEORETICAL DERIVATION

Given a set of data $\mathbf{y}_n = (y_1, y_2, \ldots, y_n)^\top$ generated from the operating model

$$\mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon}, \tag{3}$$

where $\boldsymbol{\mu}$ is the true mean of $\mathbf{y}$ and $\boldsymbol{\varepsilon} \sim N(0, \sigma_0^2 I_n)$, I_n is the $n \times n$ identity matrix. We consider the approximating model

$$\mathbf{y} = X_k \boldsymbol{\beta}_k + \boldsymbol{\epsilon}, \tag{4}$$

where X_k is a nonstochastic $n \times k$ matrix, $\boldsymbol{\beta}_k$ is a $k \times 1$ parameter vector and $\boldsymbol{\epsilon} \sim N(0, \sigma_k^2 I_n)$. The parameter $\hat{\theta}_k = (\hat{\boldsymbol{\beta}}_k^\top, \hat{\sigma}_k^2)^\top$ is estimated using the maximum likelihood principle

$$\hat{\boldsymbol{\beta}}_k = \left(X_k^\top X_k\right)^{-1} X_k^\top \mathbf{y}_n, \quad \hat{\sigma}_k^2 = ||\mathbf{y}_n - X_k \hat{\boldsymbol{\beta}}_k||^2 / n.$$

Let us define the vector of parameters $\boldsymbol{\beta}_k^*$ such that

$$\boldsymbol{\beta}_k^* = \arg\min_{\boldsymbol{\beta}_k} ||\boldsymbol{\mu} - X_k \boldsymbol{\beta}_k||^2, \quad \text{then}$$

$$X_k^\top (\boldsymbol{\mu} - X_k \boldsymbol{\beta}_k^*) = 0, \quad \text{and} \quad \boldsymbol{\beta}_k^* = \left(X_k^\top X_k\right)^{-1} X_k^\top \boldsymbol{\mu}$$

This leads to

$$\begin{aligned} ||\boldsymbol{\mu} - X_k \hat{\boldsymbol{\beta}}_k||^2 &= ||\boldsymbol{\mu} - X_k \boldsymbol{\beta}_k^* + X_k \left(\boldsymbol{\beta}_k^* - \hat{\boldsymbol{\beta}}_k\right)||^2 \\ &= ||\boldsymbol{\mu} - X_k \boldsymbol{\beta}_k^*||^2 + ||X_k \left(\boldsymbol{\beta}_k^* - \hat{\boldsymbol{\beta}}_k\right)||^2 \\ &= ||\boldsymbol{\mu} - X_k \left(X_k^\top X_k\right)^{-1} X_k^\top \boldsymbol{\mu}||^2 + ||X_k \left(\boldsymbol{\beta}_k^* - \hat{\boldsymbol{\beta}}_k\right)||^2 \\ &= \boldsymbol{\mu}^\top (I_n - H_k)\, \boldsymbol{\mu} + ||X_k \left(\boldsymbol{\beta}_k^* - \hat{\boldsymbol{\beta}}_k\right)||^2, \end{aligned}$$

where $H_k = X_k \left(X_k^\top X_k\right)^{-1} X_k^\top$. The nonnegative quantity $\lambda = \boldsymbol{\mu}^\top (I_n - H_k)\, \boldsymbol{\mu} / \sigma_0^2$, determines how much the true model is underfitted by the k-th candidate model. The larger the value of λ, the larger the undefit is.
Let $\chi_p^2(\lambda)$ denote the noncentral chi-square random variable with p degrees of freedom and noncentrality parameter λ, we have the following proposition.

Proposition: The random variables $||\boldsymbol{\mu} - X_k \hat{\boldsymbol{\beta}}||^2$ and $\hat{\sigma}_k$ are independent. Further

$$\left(||\boldsymbol{\mu} - X_k \hat{\boldsymbol{\beta}}_k||^2 / \sigma_0^2 - \lambda\right) \sim \chi_k^2 \quad \text{and} \quad \left(n\hat{\sigma}_k^2 / \sigma_0^2\right) \sim \chi_{n-k}^2(\lambda).$$

The proof follows from the arguments of Rao [8] (pp. 186,187,209). From Rao [8] (pp. 182), if a random variable $X \sim \chi_p^2(\lambda)$, then X has a probability density function

$$g(x) = e^{-\frac{1}{2}\lambda} \sum_{r=0}^{\infty} \frac{1}{r!} \left(\frac{1}{2}\lambda\right)^r f_{2r+p}(x),$$

where $f_{2r+p}(x)$ is the probability density function of a central χ_{2r+p}^2 random variable.

Let $p_0(\mathbf{y})$ be the sampling distribution of the true unknown model and by $p(\mathbf{y}|\hat{\theta}_k)$ the sampling distribution of the kth fitted candidate model, which we denote for simplicity by $p_k(\mathbf{y})$. Recall that the Kullback symmetric divergence between the two models is def ined as [9]

$$J_n(p_0, p_k) = 2 \int p_0(\mathbf{y}) \log\left(\frac{p_0(\mathbf{y})}{p_k(\mathbf{y})}\right) + p_k(\mathbf{y}) \log\left(\frac{p_k(\mathbf{y})}{p_0(\mathbf{y})}\right) d\mathbf{y}.$$

Let's define the quantity

$$d_n(i, j) = -2 \int p_i(\mathbf{y}) \log p_j(\mathbf{y}) d\mathbf{y}$$

Neglecting the term $d_n(0, 0)$, since it does not depend on k, the quantity

$$K_n(p_0, p_k) = d_n(0, k) + d_n(k, 0) - d_n(k, k),$$

would provide a suitable measure of the Kullback symmetric divergence without affecting its discrimination ability. Under the modelling framework of (3) and (4), the above quantity will be equal to

$$\begin{aligned} K_n(p_0, p_k) = n \log(\sigma_0^2) + n \log(2\pi) + ||\boldsymbol{\mu} - X_k \hat{\boldsymbol{\beta}}_k||^2 / \sigma_0^2 \\ + ||\boldsymbol{\mu} - X_k \hat{\boldsymbol{\beta}}_k||^2 / \hat{\sigma}_k^2 + n \frac{\sigma_0^2}{\hat{\sigma}_k^2} + n \frac{\hat{\sigma}_k^2}{\sigma_0^2} - n. \end{aligned} \tag{5}$$

Now, let us find the expected value of $K_n(p_0, p_k)$ denoted by

$$\Omega_n(k) = \mathbb{E}_0 \left\{K_n(p_0, p_k)\right\},$$

where the expectation is with respect to the true unknown model $p_0(\mathbf{y})$.
Since the inverse of χ_{2r+p}^2 has an expected value of $(2r+p-2)^{-1}$, it follows that

$$\mathbb{E}_0 \left\{\frac{\sigma_0^2}{n\hat{\sigma}_k^2}\right\} = e^{-\frac{1}{2}\lambda} \sum_{r=0}^{\infty} \frac{1}{r!} \left(\frac{1}{2}\lambda\right)^r \frac{1}{2r + n - k - 2}.$$

and

$$\begin{aligned} \mathbb{E}_0 \left\{||\boldsymbol{\mu} - X_k \hat{\boldsymbol{\beta}}_k||^2 / \hat{\sigma}_k^2\right\} &= n \mathbb{E}_0 \left\{\frac{||\boldsymbol{\mu} - X_k \hat{\boldsymbol{\beta}}_k||^2}{\sigma_0^2}\right\} \mathbb{E}_0 \left\{\frac{\sigma_0^2}{n\hat{\sigma}_k^2}\right\} \\ &= n(\lambda + k) e^{-\frac{1}{2}\lambda} \sum_{r=0}^{\infty} \frac{1}{r!} \left(\frac{1}{2}\lambda\right)^r \frac{1}{2r + n - k - 2} \end{aligned}$$

after simplification we, get

$$\begin{aligned} \Omega_n(k) \quad = \quad & n \log(\sigma_0^2) + 2\lambda + n \log(2\pi) \\ & + n e^{-\frac{1}{2}\lambda} \sum_{r=0}^{\infty} \frac{1}{r!} \left(\frac{1}{2}\lambda\right)^r \frac{n + k + \lambda}{2r + n - k - 2}. \end{aligned} \tag{6}$$

Now to compute the expected value of KIC and $KICc$, all that we need is to compute the quantity

$$\mathbb{E}_0 \left\{\log\left(\frac{n\hat{\sigma}_k^2}{\sigma_0^2}.\right)\right\}$$

Since the logarithm of a χ_{2r+p}^2 has expected value of $\psi\left(r + \frac{p}{2}\right) + \log 2$, it follows that

$$\mathbb{E}_0 \left\{\log\left(\frac{n\hat{\sigma}_k^2}{\sigma_0^2}\right)\right\} = e^{-\frac{1}{2}\lambda} \sum_{r=0}^{\infty} \frac{1}{r!} \left(\frac{1}{2}\lambda\right)^r \psi\left(r + \frac{n-k}{2}\right) + \log 2$$

Thus, using the above result with the definition of KIC and $KICc$ as in (1) and (2), one can get

$$\mathbb{E}_0\{KIC\} = n\log(\sigma_0^2) + ne^{-\frac{1}{2}\lambda}\sum_{r=0}^{\infty}\frac{1}{r!}\left(\frac{1}{2}\lambda\right)^r\psi\left(r+\frac{n-k}{2}\right) + n\left\{\log\left(\frac{4\pi}{n}\right)+1\right\} + 3(k+1) \quad (7)$$

and

$$\begin{aligned}\mathbb{E}_0\{KICc\} &= n\log(\sigma_0^2) + 2\frac{(k+1)n}{n-k-2} - n\psi\left(\frac{n-k}{2}\right) \\ &\quad + ne^{-\frac{1}{2}\lambda}\sum_{r=0}^{\infty}\frac{1}{r!}\left(\frac{1}{2}\lambda\right)^r\psi\left(r+\frac{n-k}{2}\right) \\ &\quad + n(1+\log(2\pi)) \end{aligned} \quad (8)$$

Equations (6), (7) and (8) will be used later in the simulation to compute the expected value of the Kullback symmetric divergence KIC and $KICc$ respectively.

3. SIMULATION

Let us consider the operating model

$$y_i = f(x_i) + \varepsilon_i, \qquad i = 1, 2, \ldots, n, \quad (9)$$

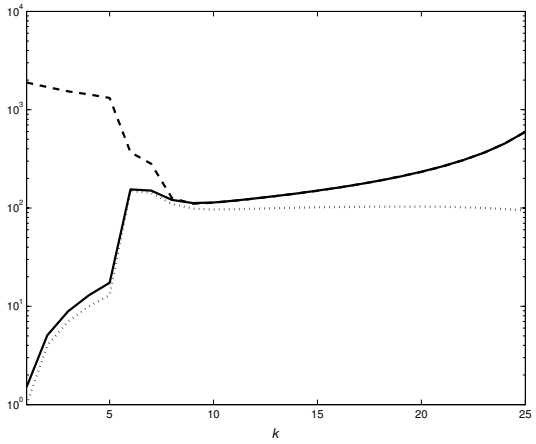
where $f(x) = x\sin(4\pi x)$, x_i are an equally spaced grid over the interval $[0, 1]$ and ε_i are i.i.d Gaussian random variable with zero mean and variance σ_0^2. We also consider the approximating models

$$y_i = \sum_{p=1}^{k} a_p x_i^{p-1} + \epsilon_i, \quad (10)$$

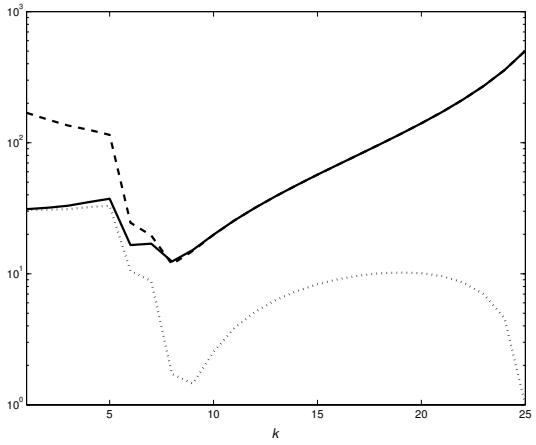
where $a_1, a_2, \ldots, a_k$ are real valued parameters and ϵ_i are independent and identically distributed normal random variables with zero mean and variance σ_k^2. We denote the k-th candidate approximating model by

$$f_k(x) = a_k x^{k-1} + \ldots + a_2 x + a_1.$$

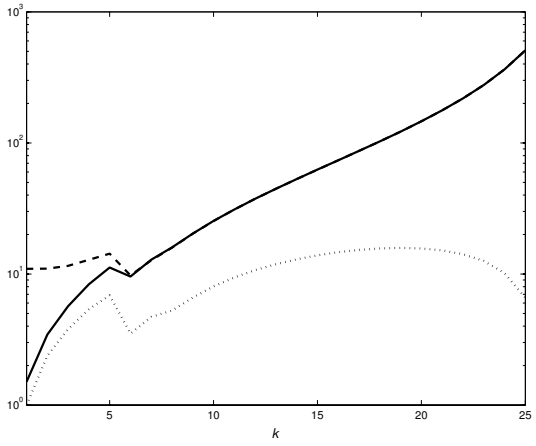
The motivation for studying this example is that polynomials create a difficult model selection problem that has a strong tendency to produce catastrophic overfitting effects. An other benefit is that polynomials are an interesting class of linear models, for which there are efficient techniques for computing the best fit.
Figure 1 gives plots of $\Omega_n(k)$, together with the expectation of KIC and $KICc$ as a function of k, where $k = 1, 2, \ldots, 25$, for sample size of $n = 30$, under three different values of σ_0^2. Although $KICc$ is not uniformly less biased than KIC, the expected value of $KICc$ outperforms that of KIC in capturing the over all shape of $\Omega_n(k)$. In particular, $\mathbb{E}_0\{KIC\}$ is often minimized at a large value of k, and clearly suboptimal, whereas the minimizer of $\mathbb{E}_0\{KICc\}$ and $\Omega_n(k)$ are similar. This may happen when a large order cutoff is imposed on the class of candidate models as a consequence of a lack of prior information about the nature of the true model. Furthermore as the ratio of the model order to the number of data increases, KIC becomes strongly negatively biased, while the bias of $KICc$ is much smaller. Finally we note that both KIC and $KICc$ are biased for low dimensional approximating models and that their biases increase when the noise variance decreases.



(a) $\sigma_0^2 = 0.005$



(b) $\sigma_0^2 = 0.05$



(c) $\sigma_0^2 = 0.5$

Fig. 1. $\mathbb{E}_0\{KIC\}$ (. . .), $\mathbb{E}_0\{KICc\}$(—) and Ω_n(--) as function of k for Polynomial regression candidates.

σ_0^2	$\overline{D_{KICc}}$	$\overline{D_{KIC}}$	$\overline{D_{MDL}}$	$\overline{D_{AIC}}$
0.005	**41.18**	119.23	102.02	163.84
0.05	**38.60**	115.08	94.44	172.05
0.5	**25.52**	88.08	64.91	164.51

Table 1. Averages of Kullback symmetric divergence, $n = 30$

n	$\overline{D_{KICc}}$	$\overline{D_{KIC}}$	$\overline{D_{MDL}}$	$\overline{D_{AIC}}$
40	**33.10**	48.55	39.07	77.40
50	**30.97**	37.21	32.30	55.25
60	**29.58**	32.71	29.99	44.76
100	**28.23**	29.04	28.34	35.36
200	**27.63**	27.86	28.96	31.32

Table 2. Averages of Kullback symmetric divergence, $\sigma_0^2 = 0.05$.

Next we explore the quality of models selected by KIC and $KICc$ for polynomial model selection. Since there is no true finite polynomial model order in the current study, we will measure the quality using the average Kullback symmetric divergence, instead of simply examining the selected model orders. For comparison purpose we have considered two other well-known criteria, the Akaike Information Criterion (AIC) [6] and the Minimum Description Length (MDL) [7]

$$AIC(k) = n \log \hat{\sigma}_k^2 + 2k, \quad MDL(k) = n \log \hat{\sigma}_k^2 + k \log n.$$

For each data realization obtained using the model in (9), the criterion A leads to selected model order $\hat{k}_A$, and the discrepancy $K_n(p_0, p_{\hat{k}_A})$ as defined in (5). In order to allow these discrepancy value to be viewed as relative to an absolute zero, the constant $d_n(0,0)$ was subtracted, yielding to $D_A = K_A - d_n(0,0)$. The average values of D_{KIC}, D_{KICc}, D_{MDL}, D_{AIC} are obtained using Monte Carlo simulation over 10000 data realizations with data size $n = 30$. The results are shown in Table 1. For all noise levels considered the average value of D_{KICc} is less than those of D_{KIC}, D_{AIC}, D_{MDL}, suggesting that $KICc$ provides the best model selections on the average. We have repeated the same set of simulations with different sample size n and kept the noise variance constant at $\sigma_0^2 = 0.05$. The same results can be noticed as shown in Table 2. Clearly the performance of $KICc$ is outstanding for small sample cases and as the number of data increases all the different competing criteria became equivalent.
Another criterion for assessing the quality of the fitted polynomial model using a model selection criterion is the mean square error of approximation, defined as

$$M_A = \int_0^1 \left(f(x) - f_{\hat{k}_A}(x) \right)^2 dx \qquad (11)$$

Table 3 and Table 4 give averages of M_{KIC}, M_{AIC},M_{MDL} and M_{KICc} using the same Monte Carlo simulation reported earlier. The results are reasonably similar to those found in Table 1 and Table 2 for the Kullback symmetric divergence, with $KICc$ uniformly performing best, especially at small and moderate data sample.

σ_0^2	$\overline{M_{KICc}}$	$\overline{M_{KIC}}$	$\overline{M_{MDL}}$	$\overline{M_{AIC}}$
0.005	**.0019**	.0522	.0337	.1315
0.05	**.0199**	.6212	.3676	1.7792
0.5	**.1776**	3.2904	1.5881	1.6227

Table 3. Averages of Mean Square Approximation error, $n = 30$

n	$\overline{M_{KICc}}$	$\overline{M_{KIC}}$	$\overline{M_{MDL}}$	$\overline{M_{AIC}}$
40	**.0135**	.0185	.0398	.0127
50	**.0104**	.0113	.0105	.0162
60	**.0087**	.0090	.0087	.0114
100	**.0054**	.0055	.0056	.0064
200	**.0029**	.0029	.0031	.0032

Table 4. Averages of Mean Square Approximation error, $\sigma_0^2 = 0.05$

4. 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] A.-K. Seghouane, M. Bekara, and G. Fleury, "A small model selection criterion based on Kullback's symmetric divergence," *Proceeding of ICASSP, (Hong-Kong)*, pp. 145–148, 2003.

[3] H. Linhart and W. Zucchini, *Model Selection*, Wiley series in Probability and Mathematical Statistics. John Wiley & Sons, NY, 1986.

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

[5] R. Shibata, "Asymptotically efficient selection of the order of of the model for estimating parameters of a linear process," *Annals of Statistics*, vol. 8, pp. 147–164, 1980.

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

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

[8] C. R. Rao, *Linear statistical enference and its applications*, Wiley, New York, 2 edition, 1973.

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