



Asymptotic bootstrap corrections of AIC for linear regression models

Abd-Krim Seghouane ^{*,1}

National ICT Australia, Canberra Research Laboratory and The Australian National University, College of Engineering and Computer Science, Canberra, Australia

ARTICLE INFO

Article history:

Received 10 October 2008

Received in revised form

19 May 2009

Accepted 9 June 2009

Available online 16 June 2009

Keywords:

Model selection

Kullback–Leibler information

Bootstrap

AIC

AIC_c

ABSTRACT

The Akaike information criterion, AIC, and its corrected version, AIC_c are two methods for selecting normal linear regression models. Both criteria were designed as estimators of the expected Kullback–Leibler information between the model generating the data and the approximating candidate model. In this paper, two new corrected variants of AIC are derived for the purpose of small sample linear regression model selection. The proposed variants of AIC are based on asymptotic approximation of bootstrap type estimates of Kullback–Leibler information. These new variants are of particular interest when the use of bootstrap is not really justified in terms of the required calculations. As its the case for AIC_c , these new variants are asymptotically equivalent to AIC. Simulation results which illustrate better performance of the proposed AIC corrections when applied to polynomial regression in comparison to AIC, AIC_c and other criteria are presented. Asymptotic justifications for the proposed criteria are provided in the Appendix.

© 2009 Elsevier B.V. All rights reserved.

1. Introduction

In a variety of scientific and engineering modelling problems, an investigator is generally confronted with the problem of determining a suitable model from a class of candidate models to describe or characterize an experimental data set. For example, this is the case in image segmentation [1], signal denoising [2], channel order estimation [3], estimating the number of signals arriving at an array [4], estimating the number of principal components in PCA [5] and in the determination of filter order in adaptive estimation [6]. A companion to the problem of model determination or selection in data modelling is the problem of parameter estimation which is generally solved by maximum likelihood or least squares.

The selection of a model is often facilitated by the use of a model selection criterion where one only has to evaluate two simple terms. The underlying idea of model selection criteria is the parsimonious principle which says that there should be a trade-off between data fitting and complexity. Thus all criteria have one term defining a measure of fit, typically a deviance statistic and one term characterizing the complexity, a multiple of the number of free parameters in the model also called penalty term.

Largely stimulated by the ground-breaking work of Akaike [7], different strategies have been used to derive different model selection criteria. The minimum description length (MDL) suggested in [8] implement the parsimony principle of economy to code a data set using the idea of universal coding introduced by Kolmogorov. Therefore, the model chosen with the MDL can be considered as providing the best explanation of the data in terms of code length. In [9,10], based on Bayesian arguments and maximum *a posteriori* probability, the Bayesian information criterion (BIC) was introduced. A model selection criterion can also be designed to estimate an expected overall discrepancy, a quantity which reflects the degree of disparity between a fitted approximating model and the generating or “true” model. Estimation of

* Tel.: +61 2 62 67 62 29; fax: +61 62 67 62 20.

E-mail address: Abd-krim.seghouane@nicta.com.au

¹ NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program.

Kullback–Leibler information [11] is the key to derive the Akaike information criterion, AIC [12]. The FPE [13] and C_p [14] criteria follow from estimating the L_2 distance between the fitted candidate model and the true model. From the estimation of Kullback–Leibler symmetric divergence [15] follows the Kullback information criterion, KIC [16]. In [17], a thorough investigation on the role of cross-validation was made and the use of leave-many-out cross-validation, which requires intensive computing, was suggested for model selection.

These criteria are both computationally and heuristically appealing, which partly explains their enduring popularity among practitioners. Except the techniques based on cross-validation, these criteria suffer from one commonly observed drawback: their penalty terms are simple minded bias corrections and there is no assurance that such penalty terms yield a good model order estimate. Indeed, these criteria have the tendency to produce a wrong model order estimation when the sample size is small relative to the larger model order represented within the class of approximating models. Many attempts have been made to improve these criteria and AIC particularly (on which we focus in this paper) by reducing their finite sample bias (or correcting their penalty terms). One such approach is to asymptotically evaluate the penalty terms as precisely as possible to provide better estimates of the model order. In practice such precise asymptotic approximation can be quite effective [18].

As an alternative to using asymptotic approximation techniques to evaluate penalty terms one can use bootstrap techniques to estimate these terms. The idea of using bootstrap to improve the performance of a model selection rule has been suggested and investigated in [19,20]. There are various advantages of using bootstrap estimate instead of asymptotic approximations. Firstly, the obtained criteria are more general than those obtained using asymptotic approximations. Secondly, bootstrapping may produce better finite-sample accuracy than asymptotic approximations. However, the computational burden required to evaluate these bootstrap based penalty terms is not always justifiable in practice, which makes the availability of different asymptotic approximations still interesting.

In this paper, two new corrected variants of AIC are derived for the purpose of small sample linear regression model selection. These proposed AIC variants can be considered as alternative to AIC_c [18] when the bootstrap based penalty terms are not justifiable in terms of the required calculations. Since AIC is a well established and widely used criterion, any criterion which can have a better performance will be an attractive substitute.

The rest of the paper is organized as follows. In the next section a review of AIC is provided. Section 3 describes the different existing bootstrap based penalty terms that can be used instead of the simple minded bias correction used in AIC to estimate the Kullback–Leibler information. The two new variants of AIC are derived in Section 4. In Section 5, the performance of the proposed criteria are compared to that of other criteria in a simulation example based on polynomial regression.

Section 6 examines the probabilities of overfitting in small sample size of one of the proposed criteria. Concluding remarks are given in Section 7 and the theoretical justifications of the purposed criteria are presented in the Appendices.

2. Review of the Akaike information criterion

Suppose a collection of observed data $\mathbf{y} = (y_1, \dots, y_n)$ has been sampled from an unknown distribution $G(\mathbf{y})$ having density function $g(\mathbf{y})$. Estimation of $g(\mathbf{y})$ is done within a set of candidate models M_1, \dots, M_K characterized by probability densities $f(\mathbf{y}|\theta_k)$, $k = 1, \dots, K$, where $\theta_k \in \Theta_k \subset \mathbb{R}^k$. Let $\hat{\theta}(\mathbf{y})$ denote the vector of estimated parameters obtained by maximizing the likelihood $f(\mathbf{y}|\theta_k)$ over Θ_k and let $f(\mathbf{y}|\hat{\theta}_k)$ denote the corresponding fitted model.

To determine which candidate density model best approximates the true unknown model $g(\mathbf{y})$, we require a measure which provide a suitable reflection of the separation between $g(\mathbf{y})$ and an approximating model $f(\mathbf{y}|\hat{\theta}_k)$. The Kullback–Leibler divergence also known as the cross-entropy or discrepancy is one of such measure.

For the two probability densities $g(\mathbf{y})$ and $f(\mathbf{y}|\hat{\theta}_k)$, the Kullback–Leibler divergence between $g(\mathbf{y})$ and $f(\mathbf{y}|\hat{\theta}_k)$ with respect to $g(\mathbf{y})$ is defined as

$$\begin{aligned} I_n(g(\cdot), f(\cdot|\hat{\theta}_k)) &= E_g \left\{ 2 \ln \frac{g(\mathbf{y})}{f(\mathbf{y}|\hat{\theta}_k)} \right\} \\ &= E_g \{-2 \ln f(\mathbf{y}|\hat{\theta}_k)\} - E_g \{-2 \ln g(\mathbf{y})\} \\ &= d_n(g, f_k) - d_n(g, g), \end{aligned}$$

where

$$d_n(g, f_k) = E_g \{-2 \ln f(\mathbf{y}|\hat{\theta}_k)\}, \quad (1)$$

and $E_g\{\cdot\}$ represents the expectation with respect to $g(\mathbf{y})$. The selection of the best family of candidate models can be carried out according to

$$\hat{k} = \arg \min_{k \in \{1, \dots, K\}} I_n(g(\cdot), f(\cdot|\hat{\theta}_k)).$$

Since $d_n(g, g)$ does not depend on θ_k , any ranking of the candidate models according to $I_n(g(\cdot), f(\cdot|\hat{\theta}_k))$ would be identical to ranking them according to $d_n(g, f_k)$.

The above discussion suggests that

$$\begin{aligned} d_n(g, f_k) &= E_g \{-2 \ln f(\mathbf{y}|\hat{\theta}_k)\} \\ &= -2 \ln f(\mathbf{y}|\hat{\theta}_k) + E_g \{-2 \ln f(\mathbf{y}|\hat{\theta}_k)\} \\ &\quad - \{-2 \ln f(\mathbf{y}|\hat{\theta}_k)\} \end{aligned} \quad (2)$$

would provide a suitable estimate of the Kullback–Leibler divergence up to the order of a constant between the generating model $g(\mathbf{y})$ and the candidate model $f(\mathbf{y}|\hat{\theta}_k)$. Yet evaluating $d_n(g, f_k)$ is not possible, since doing so requires the knowledge of $g(\mathbf{y})$.

However, as noted in [21] $-2 \ln f(\mathbf{y}|\hat{\theta}_k)$ serves as a biased estimator of (1) and that, under proper regularity conditions [22], the bias adjustment

$$E_g \{E_g \{-2 \ln f(\mathbf{y}|\hat{\theta}_k)\}\} - E_g \{-2 \ln f(\mathbf{y}|\hat{\theta}_k)\} \quad (3)$$

متن کامل مقاله

دریافت فوری ←

ISIArticles

مرجع مقالات تخصصی ایران

- ✓ امکان دانلود نسخه تمام متن مقالات انگلیسی
- ✓ امکان دانلود نسخه ترجمه شده مقالات
- ✓ پذیرش سفارش ترجمه تخصصی
- ✓ امکان جستجو در آرشیو جامعی از صدها موضوع و هزاران مقاله
- ✓ امکان دانلود رایگان ۲ صفحه اول هر مقاله
- ✓ امکان پرداخت اینترنتی با کلیه کارت های عضو شتاب
- ✓ دانلود فوری مقاله پس از پرداخت آنلاین
- ✓ پشتیبانی کامل خرید با بهره مندی از سیستم هوشمند رهگیری سفارشات