Applied Mathematics Letters 74 (2017) 33–45

Contents lists available at ScienceDirect

Applied Mathematics Letters

www.elsevier.com/locate/aml

AIC under the framework of least squares estimation

H.T. Banks[a,*], Michele L. Joyner[b]

[a] *Center for Research in Scientific Computation, North Carolina State University, Raleigh, NC, United States*
[b] *Department of Mathematics and Statistics, East Tennessee State University, Johnson City, TN 37614, United States*

ARTICLE INFO

Article history:
Received 8 May 2017
Accepted 8 May 2017
Available online 23 May 2017

Keywords:
Inverse problems
Akiake information content
Least squares estimation
Biological applications

ABSTRACT

In this note we explain the use of the Akiake Information Criterion and its related model comparison indices (usually derived for maximum likelihood estimator inverse problem formulations) in the context of least squares (ordinary, weighted, iterative weighted or "generalized", etc.) based inverse problem formulations. The ideas are illustrated with several examples of interest in biology.

1. Introduction and overview of AIC

The Akaike Information Criterion (AIC) is one of the most widely used methods for choosing a "best approximating" model from several competing models given a particular data set [1,2]. It was first developed by Akaike in 1973 [3] and expanded upon in several subsequent papers [4–9]. The basis of the Akaike Information Criterion relies on several assumptions. It is assumed that the given data or set of observations is a realization of a random variable which has some unknown probability distribution; however, one can draw inferences about the "true" distribution using the distribution of the data. Using this assumption, the best approximating model would be the model in which the "distance" between the estimated distribution and "true" distribution is as small as possible. Kullback–Leibler (K–L) information is a well-known measure of the "distance" between two probability distribution models. Suppose $\mathbf{Y}$ is a random variable characterized by a probability density function $p(\mathbf{y}|\boldsymbol{\theta})$ where $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$ is a k-dimensional parameter vector, $\boldsymbol{\theta} \in \mathbb{R}^k$, for the distribution. We assume there exists a true parameter $\boldsymbol{\theta}_0$ such that $p_0 = p(\cdot|\boldsymbol{\theta}_0)$ is the true probability density function of observations $\mathbf{Y}$. Then the K–L information between the estimated model and

* Corresponding author.
E-mail address: htbanks@ncsu.edu (H.T. Banks).

http://dx.doi.org/10.1016/j.aml.2017.05.005
0893-9659/

"true" model is given by

$$\begin{aligned}\mathcal{I}(p_0, p(\cdot,\boldsymbol{\theta})) &= \int_{\Omega_y} p_0(\mathbf{y}) \ln\left(\frac{p_0(\mathbf{y})}{p(\mathbf{y}|\boldsymbol{\theta})}\right)\, d\mathbf{y} \\ &= \int_{\Omega_y} p_0(\mathbf{y}) \ln(p_0(\mathbf{y}))\, d\mathbf{y} - \int_{\Omega_y} p_0(\mathbf{y}) \ln(p(\mathbf{y}|\boldsymbol{\theta}))\, d\mathbf{y}\end{aligned} \tag{1}$$

where Ω_y is the set of all possible values for $\mathbf{y}$. We know that $\mathcal{I}(p_0, p(\cdot,\boldsymbol{\theta})) = 0$ if and only if $p_0 = p(\cdot|\boldsymbol{\theta})$; therefore, a good approximation model is one in which K–L information is small. However, the K–L information quantity cannot be calculated directly as the true model p_0 is generally unknown.

Yet, the maximum likelihood estimate $\boldsymbol{\theta}_{MLE}(\mathbf{Y})$ is shown to be a natural estimator for $\boldsymbol{\theta}_0$ [1,6,10]. In the misspecified case (i.e., when there does not exist a "true" value $\boldsymbol{\theta}_0$ for $\boldsymbol{\theta}$ such that $p(\cdot|\boldsymbol{\theta}) \equiv p_0$), the asymptotic normality property of the maximum likelihood estimator gives that $\boldsymbol{\theta}_{MLE}(\mathbf{Y})$ is normally distributed with

$$\mathbb{E}(\boldsymbol{\theta}_{MLE}(\mathbf{Y})) = \arg\min_{\boldsymbol{\theta}\in\Omega_\theta} \mathcal{I}(p_0, p(\cdot|\boldsymbol{\theta})).$$

Furthermore, $\mathbb{E}_{\mathbf{Y}}(\mathcal{I}(p_0, p(\cdot|\boldsymbol{\theta}_{MLE}(\mathbf{Y}))) > \mathcal{I}(p_0, p(\cdot|\theta)))$ [2]; therefore, $\mathbb{E}_{\mathbf{Y}}(\mathcal{I}(p_0, p(\cdot|\boldsymbol{\theta}_{MLE}(\mathbf{Y}))))$ can be used to estimate the "distance" between p and p_0. Thus the best approximating model would be the one that solves

$$\min_{p\in\mathbf{P}} \mathbb{E}_{\mathbf{Y}}(\mathcal{I}(p_0, p(\cdot|\boldsymbol{\theta}_{MLE}(\mathbf{Y}))))$$

where $\mathbb{P}$ is a set of candidate models. Following the derivation in [2], we can write

$$\begin{aligned}\mathbb{E}_{\mathbf{Y}}(\mathcal{I}(p_0, p(\cdot|\boldsymbol{\theta}_{MLE}(\mathbf{Y})))) &= \int_{\Omega_y} p_0(\mathbf{x}) \ln(p_0(\mathbf{x}))\, d\mathbf{x} - \mathbb{E}_{\mathbf{Y}}\left(\int_{\Omega_y} p_0(\mathbf{x}) \ln(p(\mathbf{x}|\boldsymbol{\theta}_{MLE}(\mathbf{y})))\, d\mathbf{x}\right) \\ &= \int_{\Omega_y} p_0(\mathbf{x}) \ln(p_0(\mathbf{x}))\, d\mathbf{x} - \mathbb{E}_{\mathbf{Y}}\mathbb{E}_{\mathbf{X}}\left(\ln(p(\mathbf{X}|\boldsymbol{\theta}_{MLE}(\mathbf{Y})))\right).\end{aligned}$$

Therefore,

$$\min_{p\in\mathbf{P}} \mathbb{E}_{\mathbf{Y}}(\mathcal{I}(p_0, p(\cdot|\boldsymbol{\theta}_{MLE}(\mathbf{Y})))) = \max_{p\in\mathbf{P}} \mathbb{E}_{\mathbf{Y}}\mathbb{E}_{\mathbf{X}}\left(\ln(p(\mathbf{X}|\boldsymbol{\theta}_{MLE}(\mathbf{Y})))\right).$$

Furthermore, for a large sample and "good" model, it can be shown (see [2] for details) that

$$\max_{p\in\mathbf{P}} \mathbb{E}_{\mathbf{Y}}\mathbb{E}_{\mathbf{X}}\left(\ln(p(\mathbf{X}|\boldsymbol{\theta}_{MLE}(\mathbf{Y})))\right) \approx \ln\left(\mathcal{L}(\hat{\boldsymbol{\theta}}_{MLE}|\mathbf{y})\right) - \kappa_\theta$$

where $\mathcal{L}(\hat{\boldsymbol{\theta}}_{MLE}|\mathbf{y}) = p(\mathbf{y}|\hat{\boldsymbol{\theta}}_{MLE})$ represents the likelihood of $\hat{\boldsymbol{\theta}}_{MLE}$ given sample outcomes $\mathbf{y}$ and κ_θ is the total number of estimated parameters. For historical reasons, Akaike multiplied by -2 yielding the well-known Akaike information criterion (AIC):

$$AIC = -2\ln\left(\mathcal{L}(\hat{\boldsymbol{\theta}}_{MLE}|\mathbf{y})\right) + 2\kappa_\theta. \tag{2}$$

Note that the complexity of the model, as given by the total number of parameters in the model, is considered in the AIC. Given the same level of accuracy, the simpler model is preferable to the more complex one.

In this paper, we focus on models which are n-dimensional vector dynamical systems or mathematical models of the form

$$\begin{aligned}\frac{d\mathbf{x}}{dt}(t) &= \mathbf{g}(t, \mathbf{x}(t), \mathbf{q}), \\ \mathbf{x}(t_0) &= \mathbf{x}_0\end{aligned}$$