Information Criteria for Statistical Modeling in Data-Rich Era

Genshiro Kitagawa(✉)

Meiji Institute for Advanced Study of Mathematical Sciences,
Meiji University, Tokyo 164-8525, Japan
kitagawa@ism.ac.jp

Abstract. Due to the dramatic development of measuring instruments in recent years, a huge amount of large-scale data has been acquired in all research areas. Along with this, research method has changed, and data-driven methods are becoming important as the fourth scientific methodology. In the data-driven approach, the model is built according to the theory, knowledge, data, and further the purpose of the analysis. Once a model is built, useful information can be extracted from the data through the fitted model. In this data-driven method, it is crucial to use a good model and thud the evaluation of the model is essential in the success of the data-driven approach. This paper outlines the model evaluation criteria such as AIC, GIC, EIC, and so on, focusing on information criteria for evaluating prediction accuracy based on statistical models. Since L_1 regularization is important in recent data analysis, the evaluation of the regularized model is also outlined.

1 Introduction

Due to recent development of information and communication technologies, human society is changing very rapidly. Actually, by the development of sensor devices, huge amount of data are now accumulating in various fields of scientific research, such as in life science, marketing, finance, environmental science, seismology, meteorology, astronomy and high-energy physics, etc.

Various changes occurred in this background. Firstly, the objects of scientific research were expanded (Fig. 1). Until the 19th century, the main target of the research was the static physical world. However, by the impact of Darwinism, evolutionary and changing world such as the life, economy becomes important objects in the 20th century. Further in this 21st century, owing to the development of ICT, we are facing to the so-called cyber-physical world. Secondly, objective of the research changed from the "quest for the truth" to the "prediction, simulation, knowledge creation or decision making." Thirdly, model itself was changed from physical model derived from the first principle to the modeling to achieve the objective of the research.

In parallel to the academic area, big data also appear in various aspects of our society. Actually it emerged from internet communications, sensor, drone, transaction, multi-media and various logs. And the emergence of the big data

© Springer International Publishing AG 2018
L. H. Anh et al. (eds.), *Econometrics for Financial Applications*, Studies in Computational Intelligence 760, https://doi.org/10.1007/978-3-319-73150-6_2

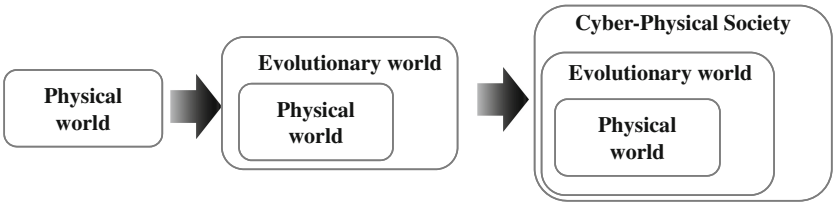

Fig. 1. Expansion of the objects of scientific research

is quickly changing our society. As examples, we can consider personalized medicine, marketing, recommendation system, data-driven industry and smartification of social infrastructure, and more recently, brilliant achievements of artificial intelligence in games, image analysis, automatic driving and so on.

In the book entitled "post-capitalist society," Drucker (1993) wrote

> *Every few hundred years in Western history there occurs a sharp transformation. We cross what I called a "divide." Within a few short decades, society rearranges itself, its worldview, its basic values, its social and political structure, its arts; its key institutions. Fifty years later, there is a new world. And the people born then cannot even imagine the world in which their grandparents lived and into which their own parents were born. We are currently living through just such a transformation.*

In the past history, the science has changed the society by expanding its fields of applications and many area that used to be treated by the intuition and experience of experts at one time became the objects of scientific approach. As such examples, we may imagine the astrology, navigation, alchemy, production process, management, marketing, finance, risk management. Further, in recent years, service and policy making, even the scientific discovery became the object of scientific research.

One typical transition is the emergence of data-driven society. In the book entitled "Super Crunchers," Ayres (2007) asserts that the "big data analysis" surpasses the "experience and intuition" of experts in many area of decision making, and showed many examples such as the evaluation of wine quality, recruiting baseball players, airline customer service, individual pricing of premium and online sales and so on.

This shows that cyber intelligence comes close to a human being in the intellectual labor and it reminds us of the historic moment of the match between horsecar and steam locomotive held at Baltimore & Ohio Railroad in 1830, when the machine has caught up with an animal's physical labor. We may say that a data-centric society will appear before long and also that all research will become data science.

From the viewpoint of the inductive inference, in the 20th century, the main objective used be the exact reasoning based on well designed small number of experimental data. Now, by the advent of the big data, an important problem is the knowledge discovery or information extraction based on big data.

However, although the big data may contain enormous knowledge and value, it is usually difficult to extract them by the current methods and technologies because it is mostly unstructured, has low value density, large scale, sparse and further it is heterogeneous in terms of precision, form, observation frequency.

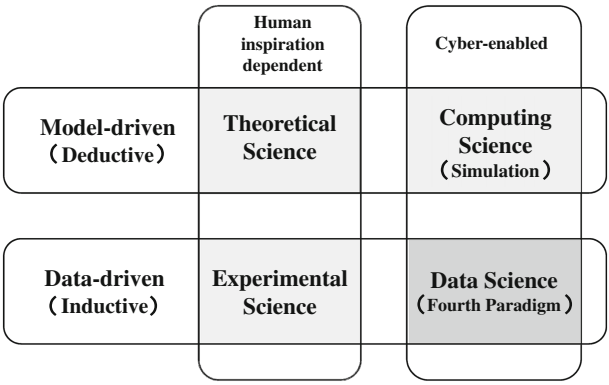

Fig. 2. Fourth Science: Data Science

To fully utilize the information contained in the big data, it is necessary to develop the fourth scientific methodology (Fig. 2). Until the 20th century, science was driven by two scientific methodologies, namely, the experimental science and the theoretical science. However, in the latter half of the 20th century, the computing science was developed for understanding or prediction of complex nonlinear systems. Now by the advent of big data, it is necessary to develop the fourth scientific methodology, namely the data science.

The basic technologies for the data science are big data processing, visualization and data analysis (Manyika et al. 2011). Big data processing is the techniques to handle scattered big data and consists of various information processing technologies such as distributed processing, parallel computation, etc. Visualization is the technologies to grasp high-dimensional data and computing results such as dimension reduction, feature extraction, pattern recognition, image processing. Data analysis is the method for obtaining deep knowledge from big data and is related to statistical modeling, Bayes inference, machine leaning, data mining, web information analysis, natural language processing and optimization.

In the data-driven approach, the model is built according to the theory, knowledge, data, and further the purpose of the analysis. Once a model is built, useful information can be extracted from the data through the fitted model. In this data-driven method, it is crucial to use a good model. Therefore, the problem of developing good model evaluation criteria is a very important.

This paper is organized as follows. In Sect. 2, we will consider the role of statistical modeling and viewpoint of predictive ability. Section 3 outlines the

information criteria AIC, TIC and AIC_c which are obtained as the approximately unbiased estimates of the expected log-likelihood of the model whose parameters are estimated by the maximum likelihood method. Section 4 outlines the GIC for the evaluation of any types of estimators defined by statistical functional, such as M-estimator and Bayes model. In Sect. 5, the bootstrap information criterion EIC is outlined which can be applied to wide class of models and situation. In Sect. 6, evaluation criteria for the models obtained by regularization methods are considered. Finally, Sect. 7 summarized the paper.

2 Statistical Modeling and Predictive Model Evaluation

In statistical modeling, model is built by properly combining the information from the theory, empirical knowledge and data and even the objective of the problem (Fig. 3). In general context, it can be formulated by using Bayes model. Once the model is obtained, we can extract useful information from data, do prediction and simulation, and decision making based on the model. So the knowledge is provided through the model and the knowledge improves the model. And thus it constitutes the spiral of knowledge development.

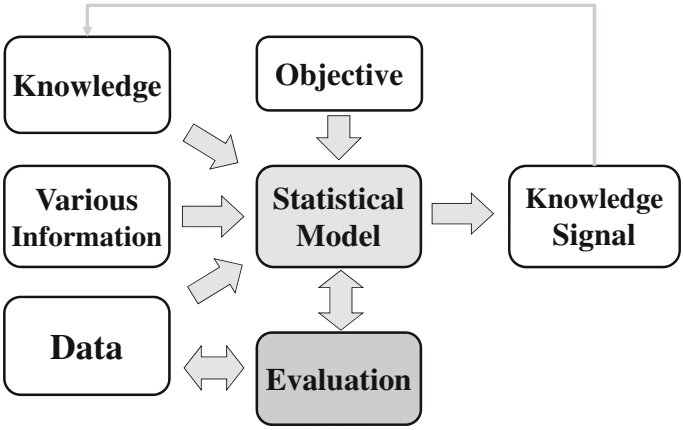

Fig. 3. Statistical modeling.

In statistical modeling, it is not necessarily assumed that the model is true or a close replica of the truth and we rather use it as a tool to extract useful information from data. Therefore, it is important to build a model by properly combining the information from the data and the prior information and knowledge on the subject and objective of the problem (Fig. 3).

In this situation, it is obvious that if we use a good model, then we can get good results but if we use a poor model, we will not be able to get meaningful results. Therefore, the use of good model is essential in statistical modeling and

statistical knowledge extraction, and the evaluation of the estimated model is one of the most important problems in the data-driven approach. To achieve this, development of criteria for evaluating the goodness of statistical model is indispensable.

In developing a model evaluation criterion, Akaike advocated the predictive point of view. In the conventional statistical procedure, the objective of model fitting and parameter estimation is to obtain a good model that can reasonably reproduce the true model as precise as possible (Fig. 4). In contrast to this, in the predictive point of view, the estimated model is evaluated by the prediction ability. Akaike (1973, 1974) measured this ability by the Kullback-Leibler information between the predictive distribution and future data distribution. The AIC is obtained as an estimate of (the essential part of) the Kullback-Leibler information.

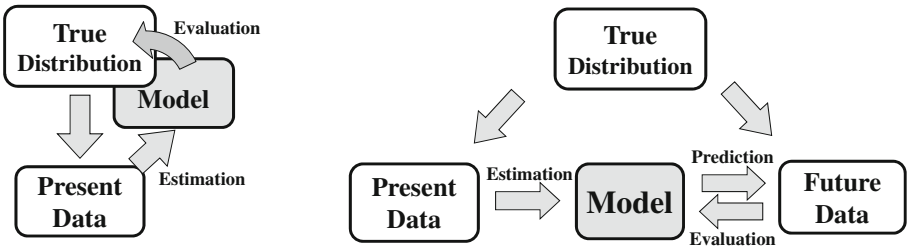

Fig. 4. Conventional statistical modeling (left) and predictive modeling (right).

Akaike's (1973, 1974) information criterion provides a useful tool for evaluating models estimated by the method of maximum likelihood and a number of successful applications of AIC in statistical modeling and data analysis have been reported (Bozdogan 1994; Kitagawa and Gersch 1996; Akaike and Kitagawa 1998). By extending Akaike's basic idea, several attempts have been made to relax the assumptions imposed in the derivation of AIC and obtained information theoretic criteria which may be applied to the various types of statistical models.

In recent years advances in the performance of computers enables us to construct models for analyzing data with complex structure, and consequently more flexible criteria are required for model evaluation and selection problems. The purpose of the present paper is to overview information criteria which yield more refined results than previously proposed criteria and may be applied to a variety of statistical models. The use of the bootstrap in model evaluation problems is also investigated from theoretical and practical points of view.

3 Information Criteria for ML Models

3.1 Estimation of Kullback-Leibler Information

Assume that the observations are generated from an unknown "true" distribution function $G(x)$ and the model is characterized by a density function $f(x)$. In the derivation of AIC (Akaike 1973, 1974; Konishi and Kitagawa 2008), the expected log-likelihood $E_Y \log f(Y) = \int \log f(y) dG(y)$ is used as the basic criterion to evaluate the closeness of a model to the true model, which is equivalent to the Kullback-Leibler information (1951). Here E_Y denotes the expectation with respect to the true distribution $G(y)$.

In actual statistical problems, the true distribution $G(x)$ is unknown and only a sample $\boldsymbol{X} = \{X_1, \ldots, X_n\}$ drawn from $G(x)$ is given. We then use the log-likelihood $n^{-1}\ell = \int \log f(x) d\hat{G}_n(x) = n^{-1} \sum_{i=1}^{n} \log f(X_i)$ as a natural estimator of the expected log-likelihood. Here $\hat{G}_n(x)$ is the empirical distribution function, having mass $1/n$ on each observation.

For a parametric model $f(x|\theta)$ with a parameter $\theta = (\theta_1, \ldots, \theta_m)^T$, it naturally leads to the maximum likelihood estimator, $\hat{\theta} = \hat{\theta}(\boldsymbol{X})$, which is the maximizor of the log-likelihood function

$$\ell(\theta) = \sum_{i=1}^{n} \log f(X_i|\theta) \equiv \log f(\boldsymbol{X}|\theta). \tag{1}$$

Interestingly, although the log-likelihood is a good estimate of the expected log-likelihood, $E_Y \log f(Y|\theta)$, the maximum log-liklihood $\log f(X|\hat{\theta})$ is NOT an unbiased estimate of $E_Y \log f(Y|\hat{\theta})$. Namely, ($n^{-1}$ times of) the maximum log-likelihood, $n^{-1}\ell(\hat{\theta}(\boldsymbol{X}))$, has a positive bias as an estimator of the expected log-likelihood, $E_Y \log f(Y|\hat{\theta}(\boldsymbol{X}))$, and it cannot be directly used for model selection.

This bias occurs because the same data set $\boldsymbol{X}$ was used twice for the estimation of the parameter and the expected log-likelihood. By correcting the bias

$$b(G) = nE_X \left\{ \frac{1}{n} \log f(\boldsymbol{X}|\hat{\theta}(\boldsymbol{X})) - E_Y \log f(Y|\hat{\theta}(\boldsymbol{X})) \right\}, \tag{2}$$

an unbiased estimator of the expected log-likelihood is obtained by $n^{-1}\{\log f(\boldsymbol{X}|\hat{\theta}(\boldsymbol{X})) - b(G)\}$. Therefore, considering the definition of AIC, generic information criteria is defined by

$$-2 \log f(\boldsymbol{X}|\hat{\theta}(\boldsymbol{X})) + 2\hat{b}(G), \tag{3}$$

where $\hat{b}(G)$ is a properly defined approximation to $b(G)$.

3.2 Information Criteria: AIC, TIC and AIC$_c$

In a general setting, it is difficult to obtain the bias $b(G)$ in a closed form. Under some setting, Akaike evaluated an asymptotic bias as $b(G) = m$, and advocated the information criterion

$$\text{AIC} = -2 \log f(\boldsymbol{X}|\hat{\theta}_{ML}) + 2m, \tag{4}$$

where m is the number of estimated parameters (Akaike 1973, 1974; Konishi and Kitagawa 2008). Numerous successful application of the statistical modeling based on AIC have been reported (Bozdogan 1994; Kitagawa and Gersch 1996; Akaike and Kitagawa 1998).

Using the properties of the maximum likelihood estimators $\hat{\theta}_{ML}$, for incorrectly specified models (Huber 1976), the asymptotic bias can be evaluated as (Takeuchi 1976)

$$b_T(G) = \mathrm{tr}\{I(G)J(G)^{-1}\}, \tag{5}$$

where $I(G)$ and $J(G)$ are respectively the Fisher information matrix and the expected Hessian defined by

$$I(G) = E_Y\left[\frac{\partial \log f(Y|\theta)}{\partial \theta}\frac{\partial \log f(Y|\theta)}{\partial \theta^T}\right],$$
$$J(G) = -E_Y\left[\frac{\partial^2 \log f(Y|\theta)}{\partial \theta \partial \theta^T}\right]. \tag{6}$$

By correcting the asymptotic bias of the log likelihood, TIC is defined by Takeuchi (1976)

$$\mathrm{TIC} = -2\log f(\boldsymbol{X}|\hat{\theta}_{ML}) + 2\mathrm{tr}\{\hat{J}(G)^{-1}\hat{I}(G)\}, \tag{7}$$

where $\hat{J}(G)$ and $\hat{I}(G)$ are consistent estimates of $J(G)$ and $I(G)$, respectively.

If the model contains the true distribution such that $g(x) = f(x|\theta)$ for some θ, it holds that $I(G) = J(G)$, and the asymptotic bias becomes $b_A(G) = m$, where m is the dimension of the parameter vector θ. Thus we obtain the Akaike information criterion, AIC (Fig. 5).

Further, for some specific models, it is possible to evaluate the bias directly and obtain a more precise bias correction term without resorting to asymptotic

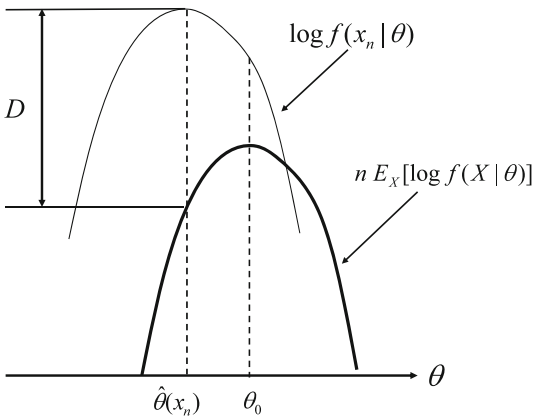

Fig. 5. Bias of the maximum log-likelihood as an estimator of the expected log-likelihood. (Konishi and Kitagawa 2008)

theory. As the simplest example, consider the normal distribution model, $y_n \sim N(\mu, \sigma^2)$. Then the log-likelihood of the model based on the data, $\{y_1, \ldots, y_n\}$, is given by

$$\ell(\mu, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{\alpha=1}^{n}(y_\alpha - \mu)^2.$$

By substituting the maximum likelihood estimators $\hat{a_j}$ and $\hat{\sigma^2}$ into this expression, we obtain the maximum log-likelihood $\ell(\hat{a_j}, \hat{\sigma}^2) = -\frac{n}{2}\log(2\pi\hat{\sigma}^2) - \frac{n}{2}$. If the data set is obtained from the same normal distribution $N(\mu, \sigma^2)$, then the expected log-likelihood is given by $E_G\left[\log f(Z|\hat{\mu}, \hat{\sigma}^2)\right] = -\frac{1}{2}\log(2\pi\hat{\sigma}^2) - \frac{1}{2\hat{\sigma}^2}\left\{\sigma^2 + (\mu - \hat{\mu})^2\right\}$, where $G(z)$ is the distribution function of the normal distribution $N(\mu, \sigma^2)$. Therefore, the difference between the two quantity is $\ell(\hat{\mu}, \hat{\sigma^2}) - nE_G\left[\log f(Z|\hat{\mu}, \hat{\sigma^2})\right] = \frac{n}{2\hat{\sigma}^2}\left\{\sigma^2 + (\mu - \hat{\mu})^2\right\} - \frac{n}{2}$. By taking the expectation with respect to the joint distribution of n observations distributed as the normal distribution $N(\mu, \sigma^2)$, and using $E_G\left[\frac{\sigma^2}{\hat{\sigma}^2(\boldsymbol{x}_n)}\right] = \frac{n}{n-3}$, $E_G\left[\{\mu - \hat{\mu}(\boldsymbol{x}_n)\}^2\right] = \frac{\sigma^2}{n}$, we obtain the bias correction term for the finite sample as

$$b_{cA}(G) = \frac{n}{2}\frac{n}{(n-3)\sigma^2}\left(\sigma^2 + \frac{\sigma^2}{n}\right) - \frac{n}{2} = \frac{2n}{n-3}. \tag{8}$$

Here, we used the fact that for a χ^2 random variable with degrees of freedom r, χ_r^2, we have $E[1/\chi_r^2] = 1/(r-2)$. Therefore, an information criterion (corrected AIC) for the normal distribution model is given by

$$\mathrm{AIC}_c = -2\ell(\hat{\mu}, \hat{\sigma}^2) + \frac{4n}{n-3}. \tag{9}$$

Similarly, for a linear regression model $y_n = \sum_{j=1}^{m} a_j x_{nj} + \varepsilon_n$, $\varepsilon \sim N(0, \sigma^2)$, where y_n and $x_{nj}, j = 1, \ldots, m$ are the objective variable and the regressors, respectively, the bias is evaluated as

$$b_{cA}(G) = \frac{(m+1)n}{n-m-2}. \tag{10}$$

4 Information Criteria for Wider Class of Models

4.1 Generalized Information Criterion GIC

This method of bias correction for the log-likelihood can be extended to a more general estimator defined by a statistical functional such as $\hat{\theta} = \boldsymbol{T}(\hat{G}_n)$, where $\boldsymbol{T}(\cdot) = (T_1(\cdot), \ldots, T_m(\cdot))^T$ is a functional on the space of distribution functions. For such a general estimator, the asymptotic bias is given by Konishi and Kitagawa (1996, 2008)

$$b_1(G) = \mathrm{tr}\left\{\int T^{(1)}(y; G)\frac{\partial \log f(y|\boldsymbol{T}(G))}{\partial \theta^T} dG(y)\right\}, \tag{11}$$

where $\boldsymbol{T}^{(1)}(\boldsymbol{Y};G) = (T_1^{(1)}(\boldsymbol{Y};G),\ldots,T_m^{(1)}(\boldsymbol{Y};G))^T$ and $T_i^{(1)}(\boldsymbol{Y};G)$ is the influence function defined by

$$T_i^{(1)}(\boldsymbol{X};G) = \lim_{\varepsilon\to\infty}\{T_i((1-\varepsilon)G+\varepsilon\delta_\alpha) - T_i(G)\}/\varepsilon \tag{12}$$

with δ_α being a point mass at X_α. By subtracting the asymptotic bias estimate from the log-likelihood, we have (Fig. 6)

$$\mathrm{GIC} = -2\log f(\boldsymbol{X}|\theta) + 2b_1(\hat{G}). \tag{13}$$

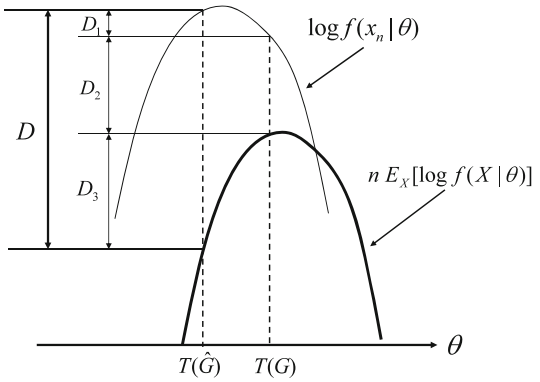

Fig. 6. Bias correction by GIC. (Konishi and Kitagawa 2008)

Example: GIC for the normal distribution model. Consider a simple normal distribution model with unknown mean μ and the variance σ^2

$$f(y|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\}. \tag{14}$$

The maximum likelihood estimators of μ and σ^2 are given by statistical functionals,

$$T_\mu(G) = \int x dG(x), \quad T_{\sigma^2}(G) = \int (x - T_\mu(G))^2 dG(x), \tag{15}$$

respectively. For these estimators, the derivatives of the functionals are given by

$$T_\mu^{(1)}(x;G) = x-\mu, \quad T_{\sigma^2}^{(1)}(x;G) = (x-\mu)^2 - \sigma^2. \tag{16}$$

Using these results, the bias correction term (11) is explicitly obtained by

$$b_1(G) = \frac{1}{2}\left(1+\frac{\mu_4}{\sigma^4}\right), \tag{17}$$

where μ_4 denotes the fourth central moments of the true distribution G. In particular, when the true distributions are standard normal distribution ($\mu_4 = 3$) and Laplace distribution ($\mu_4 = 6$), they are given by $b_1(G) = 2$ and 3.5, respectively.

4.2 Maximum Likelihood Method: Relationship Among AIC, TIC and GIC

Assume that the maximum likelihood method is used for the estimation of a model $f(x|\boldsymbol{\theta})$ based on the observed data from $G(x)$. The maximum likelihood estimator, $\hat{\boldsymbol{\theta}}_{ML}$, is defined as a solution of the equation

$$\sum_{\alpha=1}^{n} \frac{\partial \log f(x_\alpha|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{0}, \tag{18}$$

which can be expressed as $\hat{\boldsymbol{\theta}}_{ML} = \boldsymbol{T}_{ML}(\hat{G})$ using the p-dimensional functional $\boldsymbol{T}_{ML}(G)$ implicitly defined by

$$\int \left. \frac{\partial \log f(x|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right|_{\boldsymbol{\theta}=\boldsymbol{T}_{ML}(G)} dG(x) = \mathbf{0}. \tag{19}$$

The influence function for the maximum likelihood estimator can be obtain as follows: By replacing the distribution function G in (19) with $(1-\varepsilon)G + \varepsilon\delta_x$, we have

$$\int \frac{\partial \log f(y|\boldsymbol{T}_{ML}((1-\varepsilon)G + \varepsilon\delta_x))}{\partial \boldsymbol{\theta}} d\left\{(1-\varepsilon)G(y) + \varepsilon\delta_x(y)\right\} = \mathbf{0}. \tag{20}$$

Differentiating both sides with respect to ε and setting $\varepsilon = 0$ yield

$$\begin{aligned} &\int \frac{\partial \log f(y|\boldsymbol{T}_{ML}(G))}{\partial \boldsymbol{\theta}} d\left\{\delta_x(y) - G(y)\right\} \\ &\quad + \int \frac{\partial^2 \log f(y|\boldsymbol{T}_{ML}(G))}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} dG(y) \cdot \frac{\partial}{\partial \varepsilon} \left\{\boldsymbol{T}_{ML}((1-\varepsilon)G + \varepsilon\delta_x)\right\} \Big|_{\varepsilon=0} = \mathbf{0}, \end{aligned} \tag{21}$$

Noting that

$$\int \frac{\partial \log f(y|\boldsymbol{T}_{ML}(G))}{\partial \boldsymbol{\theta}} d\delta_x(y) = \frac{\partial \log f(x|\boldsymbol{T}_{ML}(G))}{\partial \boldsymbol{\theta}} \tag{22}$$

and using (19), from (21), we obtain the influence function for the maximum likelihood estimator $\hat{\boldsymbol{\theta}}_{ML} = \boldsymbol{T}_{ML}(\hat{G})$

$$\boldsymbol{T}^{(1)}_{ML}(x;G) \equiv \frac{\partial}{\partial \varepsilon} \left\{\boldsymbol{T}_{ML}((1-\varepsilon)G + \varepsilon\delta_x)\right\} \Big|_{\varepsilon=0} = J(G)^{-1} \left. \frac{\partial \log f(x|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right|_{\boldsymbol{\theta}=\boldsymbol{T}_{ML}(G)}, \tag{23}$$

where $J(G)$ is a $p \times p$ matrix given by

$$J(G) = -\int \left. \frac{\partial^2 \log f(x|\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \right|_{\boldsymbol{\theta}=\boldsymbol{T}_{ML}(G)} dG(x). \tag{24}$$

By replacing the influence function $\boldsymbol{T}^{(1)}(x;G)$ in (11) with (23), we obtain the asymptotic bias of the log-likelihood for the estimated model $f(x|\hat{\boldsymbol{\theta}}_{ML})$

$$\begin{aligned} b_{ML}(G) &= \mathrm{tr}\left\{ J(G)^{-1} \int \left. \frac{\partial \log f(x|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \log f(x|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \right|_{\boldsymbol{\theta}=\boldsymbol{T}_{ML}(G)} dG(x) \right\}, \\ &= \mathrm{tr}\left\{ J(G)^{-1} I(G) \right\} \end{aligned} \tag{25}$$

where the $p \times p$ matrix $I(G)$ is given by

$$I(G) = \int \frac{\partial \log f(x|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \log f(x|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}\Big|_{\boldsymbol{\theta}=\boldsymbol{T}_{ML}(G)} dG(x). \tag{26}$$

Therefore, for the model $f(x|\hat{\boldsymbol{\theta}}_{ML})$ estimated by the maximum likelihood method, GIC in (13) is reduced to

$$\text{TIC} = -2\sum_{\alpha=1}^{n} \log f(x_\alpha|\hat{\boldsymbol{\theta}}_{ML}) + 2\text{tr}\left\{J(\hat{G})^{-1}I(\hat{G})\right\}. \tag{27}$$

4.3 GIC for the Models Estimated by M-estimators

In this subsection we derive an information criterion for evaluating a statistical model estimated by M-estimators, using the generalized information criterion GIC in (13).

Suppose that $f(x|\hat{\boldsymbol{\theta}}_M)$ is the model of the true distribution $G(x)$, where $\hat{\boldsymbol{\theta}}_M$ is a p-dimensional M-estimator defined as the solution of the system of implicit equations

$$\sum_{\alpha=1}^{n} \boldsymbol{\psi}(x_\alpha, \hat{\boldsymbol{\theta}}_M) = \mathbf{0}. \tag{28}$$

Here, $\boldsymbol{\psi} = (\psi_1, \psi_2, \cdots, \psi_p)^T$ and $\psi_i(x, \boldsymbol{\theta})$ is a real-valued function defined on the product space of the sample and parameter spaces. The M-estimator $\hat{\boldsymbol{\theta}}_M$ is given by $\hat{\boldsymbol{\theta}}_M = \boldsymbol{T}_M(\hat{G})$ for the p-dimensional functional $\boldsymbol{T}_M(G)$ defined as the solution of the implicit equations

$$\int \boldsymbol{\psi}(y, \boldsymbol{T}_M(G))dG(y) = \mathbf{0}. \tag{29}$$

Then the influence function for the M-estimator is obtained by the same method as for the MLE as

$$\boldsymbol{T}_M^{(1)}(x; G) \equiv \frac{\partial}{\partial \varepsilon}\left\{\boldsymbol{T}_M((1-\varepsilon)G + \varepsilon\delta_x)\right\}_{\varepsilon=0} = R(\boldsymbol{\psi}, G)^{-1}\boldsymbol{\psi}(x, \boldsymbol{T}_M(G)), \tag{30}$$

where $R(\boldsymbol{\psi}, G)$ is a $p \times p$ matrix whose (i, j)-components is given by

$$R(\boldsymbol{\psi}, G)(i, j) = -\int \frac{\partial \psi_j(x, \boldsymbol{\theta})}{\partial \theta_i}\Big|_{\boldsymbol{\theta}=\boldsymbol{T}_M(G)} dG(x), \qquad (i, j = 1, \cdots, p). \tag{31}$$

Substituting this influence function $\boldsymbol{T}_M^{(1)}(x; G)$ into (11), we have the asymptotic bias of the log-likelihood of the model $f(x|\hat{\boldsymbol{\theta}}_M)$ in estimating the expected log-likelihood in the form

$$\begin{aligned} b_M(G) &= \text{tr}\left\{R(\boldsymbol{\psi}, G)^{-1}\int \boldsymbol{\psi}(x, \boldsymbol{T}_M(G))\frac{\partial \log f(x|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}\Big|_{\boldsymbol{\theta}=\boldsymbol{T}_M(G)} dG(x)\right\} \\ &= \text{tr}\left\{R(\boldsymbol{\psi}, G)^{-1}Q(\boldsymbol{\psi}, G)\right\}, \end{aligned} \tag{32}$$

where $Q(\boldsymbol{\psi}, G)$ is a $p \times p$ matrix defined by

$$Q(\boldsymbol{\psi}, G) = \int \boldsymbol{\psi}(x, \boldsymbol{T}_M(G)) \frac{\partial \log f(x|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}\bigg|_{\boldsymbol{\theta}=\boldsymbol{T}_M(G)} dG(x). \tag{33}$$

Then, GIC for evaluating the statistical model $f(x|\hat{\boldsymbol{\theta}}_M)$ with the M-estimator $\hat{\boldsymbol{\theta}}_M$ is given by

$$\mathrm{GIC_M} = -2 \sum_{\alpha=1}^{n} \log f(x_\alpha|\hat{\boldsymbol{\theta}}_M) + 2\mathrm{tr}\left\{ R(\boldsymbol{\psi}, \hat{G})^{-1} Q(\boldsymbol{\psi}, \hat{G}) \right\}, \tag{34}$$

where $R(\boldsymbol{\psi}, \hat{G})$ and $Q(\boldsymbol{\psi}, \hat{G})$ are $p \times p$ matrices given by

$$\begin{aligned} R(\boldsymbol{\psi}, \hat{G}) &= -\frac{1}{n} \sum_{\alpha=1}^{n} \frac{\partial \boldsymbol{\psi}(x_\alpha, \boldsymbol{\theta})^T}{\partial \boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}, \\ Q(\boldsymbol{\psi}, \hat{G}) &= \frac{1}{n} \sum_{\alpha=1}^{n} \boldsymbol{\psi}(x_\alpha, \hat{\boldsymbol{\theta}}) \frac{\partial \log f(x_\alpha|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}. \end{aligned} \tag{35}$$

4.4 GIC for Bayes Models

The basic predictive distribution model based on Bayesian approach is defined by the parametric model $\{f(x|\theta); \theta \in \Theta\}$ and the prior distribution $\pi(\theta)$ of the parameter θ as follows

$$h(z|\boldsymbol{X}_n) = \int f(z|\theta)\pi(\theta|\boldsymbol{X}_n)d\theta, \tag{36}$$

where $\boldsymbol{\pi(\theta|X_n)}$ is the posterior distribution of the θ based on the sample $\boldsymbol{X}_n$ and the prior distribution $\pi(\theta)$ and is given by

$$\pi(\theta|\boldsymbol{X}_n) = \prod_{\alpha=1}^{n} f(X_\alpha|\theta)\pi(\theta) \Big/ \int \prod_{\alpha=1}^{n} f(X_\alpha|\theta)\pi(\theta)d\theta. \tag{37}$$

By substituting the posterior distribution (37), the predictive distribution is obtained by

$$h(z|\boldsymbol{X}_n) = \int \exp\left[n \left\{ q(\theta|\boldsymbol{X}_n) + \frac{1}{n} \log f(z|\theta) \right\} \right] d\theta \Big/ \int \exp\left\{ nq(\theta|\boldsymbol{X}_n) \right\} d\theta. \tag{38}$$

Here, $\boldsymbol{q(\theta|X_n)}$ is given by

$$q(\theta|\boldsymbol{X}_n) = \frac{1}{n} \sum_{\alpha=1}^{n} \log f(X_\alpha|\theta) + \frac{1}{n} \log \pi(\theta). \tag{39}$$

For this density function, by obtaining the asymptotic expansion with respect to the sample size n based on the Laplace approximation of integrals (Tierney and Kadane 1986; Davison 1986), it becomes possible to apply information criterion GIC.

Assume that $\hat{\theta}_q$ and $\hat{\theta}_q(z)$ are the modes of $\boldsymbol{q(\theta|X_n)}$ and $\boldsymbol{q(\theta|X_n) + n^{-1} \log f(z|\theta)}$, respectively.

In the Laplace's method of integrals, the integrand is Taylor expansion around the mode, and obtain an approximation formula. For example, by applying the Laplace's approximation to the denominator of Eq. (38), we obtain

$$\int \exp\{nq(\theta|\boldsymbol{X}_n)\}\, d\theta = \frac{(2\pi)^{p/2}}{n^{p/2}\left|J_q(\hat{\theta}_q)\right|^{1/2}} \exp\left\{nq(\hat{\theta}_q|\boldsymbol{X}_n)\right\}\left\{1 + O_p(n^{-1})\right\}. \quad (40)$$

Here, $J_q(\hat{\theta}_q) = -\partial^2\{q(\hat{\theta}_q|\boldsymbol{X}_n)\}/\partial\theta\partial\theta^T$. Similarly, by the Laplace approximation of the integrals in the numerator of (38), we obtain the approximation of the predictive distribution

$$h(z|\boldsymbol{X}_n) = (|J_q(\hat{\theta}_q)|/|J_{q(z)}(\hat{\theta}_q(z))|)^{1/2} \exp\left[n\left\{q(\hat{\theta}_q(z)|\boldsymbol{X}_n) - q(\hat{\theta}_q|\boldsymbol{X}_n) + \frac{1}{n}\log f(z|\hat{\theta}_q(z))\right\}\right] \times \{1 + O_p(n^{-2})\},$$

where $J_{q(z)}(\hat{\theta}_q(z)) = -\partial^2\{q(\hat{\theta}_q(z)|\boldsymbol{X}_n) + n^{-1}\log f(z|\hat{\theta}_q(z))\}/\partial\theta\partial\theta^T$. From this Laplace approximation, we obtain the following asymptotic expansion of the Bayesian predictive distribution model

$$h(z|\boldsymbol{X}_n) = f(z|\hat{\theta})\left\{1 + \frac{1}{n}a(z|\hat{\theta}) + O_p(n^{-2})\right\}. \quad (41)$$

The estimator of the model $\hat{\theta}$ depends on whether the prior distribution $\pi(\theta)$ depends on the sample size n or not. Here, we consider the following two cases for the prior distribution (i) $\log\pi(\theta) = O(1)$, and (ii) $\log\pi(\theta) = O(n)$. In the case of (i), $\hat{\theta}$ becomes the maximum likelihood estimator $\hat{\theta}_{ML}$. On the other hand, for the case (ii), it becomes the mode of the posterior distribution $\hat{\theta}_B$. The statistical functionals corresponding to these estimators are respectively given as the solutions to

$$\int \frac{\partial}{\partial\theta}\log f(x|\boldsymbol{T}_{ML}(G))dG(x) = \boldsymbol{0}, \quad \int \frac{\partial}{\partial\theta}\left[\log\{f(x|\boldsymbol{T}_B(G))\pi(\boldsymbol{T}_B(G))\}\right]dG(x) = \boldsymbol{0}.$$

Therefore, in (34), by putting $\boldsymbol{\psi(x,\hat{\theta})} = \partial\log f(x|\hat{\theta}_{ML})\partial\theta$ and $\boldsymbol{\psi(x,\hat{\theta})} = \partial\left\{\log f(x|\boldsymbol{T}_B(\hat{G})) + \log\pi(\boldsymbol{T}_B(\hat{G}))\right\}\partial\theta$, we obtain the following evaluation criterion for the Bayes predictive distribution model $h(z|\boldsymbol{X}_n)$

$$\mathrm{GIC_B} = -2\sum_{\alpha=1}^{n}\log h(X_\alpha|\boldsymbol{X}_n) + 2\mathrm{tr}\left\{J(\psi,\hat{G})^{-1}I(\psi,\hat{G})\right\}. \quad (42)$$

4.5 Higher Order Bias Correction

The information criteria proposed previously are based on large-sample theory to obtain approximately unbiased estimators for the expected log-likelihood or equivalently the Kullback-Leibler information number.

We consider the statistical model $f(y|\hat{\theta})$, where $\hat{\theta}$ is defined by $\hat{\theta} = \boldsymbol{T}(\hat{G}_n)$ with $\boldsymbol{T}(\cdot)$ being a suitably defined m-dimensional functional. Hence by taking the expectation of $E_Y \log f(Y|\hat{\theta}(\boldsymbol{X}))$ over the sampling distribution G of $\boldsymbol{X}$, we have an expectation of the form

$$E_X E_Y \log f(Y|\hat{\theta}(\boldsymbol{X})) = \int g(y) \log f(y|\boldsymbol{T}(G))dy + \frac{1}{n}a_1(G) + \frac{1}{n^2}a_2(G) + O(n^{-3}). \quad (43)$$

Information criteria based on the asymptotic bias-corrected log-likelihood is second order correct for $E_Y \log f(Y|\hat{\theta})$ in the sense that the expectations of $n^{-1}\left\{\log f(\boldsymbol{X}|\hat{\theta}) \ -b_1(\hat{G})\right\}$ and $E_Y \log f(Y|\hat{\theta})$ are in agreement up to and including the term of order n^{-1}, while the expectations of $n^{-1}\log f(\boldsymbol{X}|\hat{\theta})$ and $E_Y \log f(Y|\hat{\theta})$ differ in term of order n^{-1}.

We consider the bias of $\log f(\boldsymbol{X}|\hat{\theta}) - b_1(\hat{\theta})$, as the estimator of the expected log-likelihood, defined by

$$\begin{aligned} &E_X\left[\log f(\boldsymbol{X}|\hat{\theta}) - b_1(\hat{G}) - nE_Y \log f(Y|\hat{\theta})\right] \\ &= E_X\left[\log f(\boldsymbol{X}|\hat{\theta}) - nE_Y \log f(Y|\hat{\theta})\right] - E_X\left[b_1(\hat{G})\right]. \end{aligned} \quad (44)$$

The first term in the right-hand side of the above equation can be expanded as

$$b(G) = E_X\left[\log f(\boldsymbol{X}|\hat{\theta}) - nE_Y \log f(Y|\hat{\theta})\right] = b_1(G) + \frac{1}{n}b_2(G) + O(n^{-2}), \quad (45)$$

where $b_1(G)$ is the first order bias correction term given in (11) and $b_2(G)$ is the second order bias correction term.

The expectation of the asymptotic bias estimate $b_1(\hat{G})$ is given by

$$E_X\left[b_1(\hat{G})\right] = b_1(G) + \frac{1}{n}\Delta b_1(G) + O(n^{-2}). \quad (46)$$

Hence noting that the bias of $\log f(\boldsymbol{X}|\hat{\theta}) - b_1(\hat{G})$ is

$$E_X\left[\log f(\boldsymbol{X}|\hat{G}) - b_1(\hat{G}) - nE_Y \log f(Y|\hat{\theta})\right] = \frac{1}{n}\left\{b_2(G) - \Delta b_1(G)\right\} + O(n^{-2}), \quad (47)$$

we have the second order bias corrected information criterion in the form

$$\mathrm{GIC}_2(\hat{G}_n) = -2\ell(\hat{G}) + 2\left\{b_1(\hat{G}_2) + \frac{1}{n}\left(b_2(\hat{G}_n) - \Delta b_1(\hat{G}_n)\right)\right\}. \quad (48)$$

It might be noted that GIC_2 is third-order correct for the expected log-likelihood. However, analytic expression of $b_2(G)$ and $\Delta b_1(G)$ are very complicated (Kitagawa and Konishi 2010).

Example: Second order bias correction for the normal distribution model. For normal distribution model, these correction terms are explicitly given by

$$b_2(G) - \Delta b_1(G) = \frac{1}{2}\left(\frac{\mu_4}{\sigma^4} + \frac{\mu_6}{\sigma^6}\right), \tag{49}$$

$$b_1(G) - \frac{1}{n}\Delta b_1(G) + \frac{1}{n}b_2(G) = \frac{1}{2}\left(1 + \frac{\mu_4}{\sigma^4}\right) + \frac{1}{2n}\left(\frac{\mu_4}{\sigma^4} + \frac{\mu_6}{\sigma^6}\right), \tag{50}$$

where μ_j is the j-th cumulant of the true distribution.

These show that the estimated bias correction term $b_1(\hat{G}_n)$ is biased as an estimator of $b_1(G)$, and the difference may not be negligible for small n. One of the merit of AIC is that the bias correction term does not depend on G and thus $\Delta b_A(\hat{G}_n) = 0$.

5 Bootstrap Information Criterion EIC

The bootstrap method provides an alternative method for the evaluation of the bias of the log-likelihood (Cavanaugh and Shumway 1997; Ishiguro et al. 1997; Konishi and Kitagawa 1996; Shibata 1997). The advantage of the method is that the calculation does not require the exact form of bias correction term. In the bootstrapping, the true distribution function $G(x)$ is replaced by the empirical distribution function $\hat{G}_n(x)$ defined from the observations. Therefore, in the bias term in (2), the samples $\boldsymbol{X}$ and Y from $G(x)$ are replaced by $\boldsymbol{X}^*$ and Y^* from bootstrap sample $\hat{G}_n(x)$, and the expectation $E_Y \log f(Y|\cdot)$ by $E_{Y^*} \log f(Y^*|\cdot)$. Here E_{Y^*} denotes the expectation with respect to the empirical distribution function $\hat{G}_n(y)$. The bootstrap estimate of the bias $b_B(\hat{G}_n)$ is obtained by (Fig. 7).

$$b_B(\hat{G}_n) = nE_{X^*}\left\{\frac{1}{n}\log f(\boldsymbol{X}^*|\tilde{\theta}(\boldsymbol{X}^*)) - E_{Y^*}\log f(Y^*|\tilde{\theta}(\boldsymbol{X}^*))\right\}, \tag{51}$$

where $\tilde{\theta}(\cdot)$ is an arbitrarily defined estimator of θ. In the simple i.i.d. case, we have

$$E_{Y^*}\log f(Y^*|\tilde{\theta}(\boldsymbol{X}^*)) = \int \log f(y^*|\tilde{\theta}(\boldsymbol{X}^*))d\hat{G}_n(y^*) = \frac{1}{n}\log f(\boldsymbol{X}|\tilde{\theta}(\boldsymbol{X}^*)), \tag{52}$$

and the bootstrap estimate of the bias becomes simply

$$b_B(\hat{G}_n) = E_{X^*}\left\{\log f(\boldsymbol{X}^*|\tilde{\theta}(\boldsymbol{X}^*)) - \log f(\boldsymbol{X}|\tilde{\theta}(\boldsymbol{X}^*))\right\}. \tag{53}$$

In actual computation, the bootstrap bias correction term $b_B(\hat{G}_n)$ is estimated by

$$b_B^*(\hat{G}_n) = \frac{1}{M}\sum_{i=1}^{M}\left\{\log f(\boldsymbol{X}^*_{(i)}|\tilde{\theta}(\boldsymbol{X}^*_{(i)})) - \log f(\boldsymbol{X}|\tilde{\theta}(\boldsymbol{X}^*_{(i)}))\right\}, \tag{54}$$

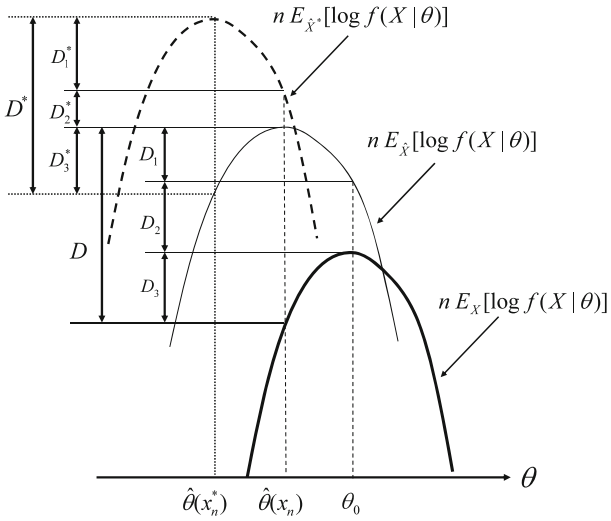

Fig. 7. Bias correction by EIC. $nE_X[\log f(X|\theta)$, $\log f(x_n|\theta)$ and $\log f(x_n^*|\theta\ *)$ are the expected log-likelihood, log-likelihood and the bootstrap log-likelihood, respectively. The expectation of D is the bias and that of D^* is the bootstrap bias. The expectation of D_2 is known to be 0 (Konishi and Kitagawa 2008).

where M is the number of bootstrap replication, $\boldsymbol{X}^*_{(1)}, \cdots, \boldsymbol{X}^*_{(M)}$ are M independent bootstrap resamples of size n from $\hat{G}_n(\boldsymbol{X})$. The bootstrap information criterion EIC then is defined by Ishiguro et al. (1997)

$$\text{EIC} = -2\log f(\boldsymbol{X}|\tilde{\theta}(\boldsymbol{X})) + 2b^*_B(\hat{G}_n). \tag{55}$$

This method of bootstrap bias correction can be easily extended to a predictive distribution of a Bayesian model defined by $\boldsymbol{p(y|X)} = \int \boldsymbol{p(y|\theta)\pi(\theta|X)d\theta}$ where $\boldsymbol{\pi(\theta|X)}$ is the posterior distribution of θ given data $\boldsymbol{X}$ (Konishi and Kitagawa 2008).

5.1 Decomposition of the Bias Term and the Reduction of the Variance in Bootstrapping

A practically important problem with the bootstrap method for the model selection is the reduction of the variance of the bias estimate. If the variance in the bootstrap simulation is large, a large M in (54) is necessary to obtain precise bootstrap estimate $b^*_B(\hat{G}_n)$ requiring long computing time especially when the model is very complicated. The variance of the bootstrap estimate of the bias defined in (54) can be reduced by the decomposition of the bias term $D(\boldsymbol{X};G)$ into three terms as follows (Fig. 7, Konishi and Kitagawa 1996, 2008; Ishiguro et al. 1997):

$$D(\boldsymbol{X};G) = D_1(\boldsymbol{X};G) + D_2(\boldsymbol{X};G) + D_3(\boldsymbol{X};G) \tag{56}$$

where

$$D_1(\boldsymbol{X};G) = \sum_{i=1}^{n} \log f(X_i|\boldsymbol{T}(\hat{G}_n)) - \sum_{i=1}^{n} \log f(X_i|\boldsymbol{T}(G))$$
$$D_2(\boldsymbol{X};G) = \sum_{i=1}^{n} \log f(X_i|\boldsymbol{T}(G)) - n\int \log f(y|\boldsymbol{T}(G))dG(y) \tag{57}$$
$$D_3(\boldsymbol{X};G) = n\int \log f(y|\boldsymbol{T}(G))dG(y) - n\int \log f(y|\boldsymbol{T}(\hat{G}_n))dG(y).$$

Note that if $\hat{\theta}$ is the MLE, then $\boldsymbol{T}(G)$ and $\boldsymbol{T}(\hat{G}_n)$ are the maximizer of $\int \log f(y|\boldsymbol{T}(G))dG(y)$ and $\sum_{i=1}^{n} \log f(X_i|\boldsymbol{T}(\hat{G}_n))$, respectively.

For a general estimator defined by a statistical functional $\hat{\theta} = \boldsymbol{T}(\hat{G}_n)$, each term can be evaluated. See Kitagawa and Konishi (2010) for details.

Further, it can be seen that $Var\{D\} = O(n)$ and $Var\{D_1 + D_3\} = O(1)$. Therefore by estimating the bias by

$$b^*(\hat{G}_n) = E_{X^*}[D_1 + D_3], \tag{58}$$

a significant reduction of the variance can be achieved for any estimators defined by statistical functional especially for large n.

5.2 Second Order Bootstrap Bias Correction

The bias of the log-likelihood shown in (2) can be expressed as

$$\frac{1}{n}b(G) = \frac{1}{n}b_1(G) + \frac{1}{n^2}b_2(G) + \frac{1}{n^3}b_3(G) + \cdots, \tag{59}$$

where $b_j(G)$ is the jth order bias correction term. Therefore, the expected value of the bootstrap estimate of the bias term is given by

$$\begin{aligned} E_X[b_B(\hat{G}_n)] &= E_X\left[b_1(\hat{G}_n) + \frac{1}{n}b_2(\hat{G}_n)\right] + o(n^{-1}) \\ &= b_1(G) + \frac{1}{n}\Delta b_1(G) + \frac{1}{n}b_2(G) + o(n^{-1}), \end{aligned} \tag{60}$$

where $\Delta b_1(G)$ is the bias of the first order bias correction term $b_1(G)$. This means that if $\Delta b_1(G) = 0$, the bootstrap estimate automatically yields the second order correction, namely it is the third order correct for the expected log-likelihood.

It is interesting to note that, in contrast to the above, the expected value of (11) in the GIC and (5) in TIC for the MLE are given by

$$E_X[b_1(\hat{G}_n)] = b_1(G) + \frac{1}{n}\Delta b_1(G) + o(n^{-1}). \tag{61}$$

In actual situations for which unbiasedness $\Delta b_1(G)$ is not assumed, we can estimate the second order correction term by bootstrapping. If an analytic expression for $b_1(G)$ is available, it is given by

$$\frac{1}{n}b_2^*(\hat{G}_n) = E_{X^*}\left[\log f(\boldsymbol{X}^*|\boldsymbol{T}(\hat{G}_n)) - b_1(\hat{G}_n) - nE_{Y^*}\log f(Y^*|\boldsymbol{T}(\hat{G}_n))\right]. \tag{62}$$

On the other hand, if an analytic expression is difficult to compute, then we can obtain the second order correction by double bootstrapping (Kitagawa and Konishi 2010),

$$\frac{1}{n}b_2^{**}(\hat{G}_n) = E_{X^*}\left[\log f(\boldsymbol{X}^*|\boldsymbol{T}(\hat{G}_n)) - b_B^*(\hat{G}_n) - nE_{Y^*}\log f(Y^*|\boldsymbol{T}(\hat{G}_n))\right], \quad (63)$$

where $b_B^*(G)$ is the bootstrap estimate of the first order correction term given by (19).

6 Regularization, L_1 Sparse Modeling and Bridge Regression

In recent years, the regularization method is used for the modeling of big data in many fields. In this section, we first consider application of GIC for the penalized log-likelihood method or the L_2 regularization problem. We then consider the generalization of the Bayesian information criterion BIC for the application to L_1 regularization and the bridge regression which involves a more general L_p regularization.

6.1 GIC for Penalized Log-Likelihood Method

The method based on maximizing the penalized log-likelihood function was originally introduced by Good and Gaskins (1971) in the context of density estimation. The Bayesian justification of the method and application to Bayesian modeling have been investigated by many authors such as Wahba (1978), Akaike (1980), Kitagawa and Gersch (1984), Silverman (1985) and Shibata (1989).

Here, we consider a penalized log-likelihood of the form

$$\ell_\lambda(\theta) = \sum_{\alpha=1}^{n}\log f(x_\alpha|\gamma,\sigma) - \frac{n}{2}\lambda\gamma' K\gamma, \quad (64)$$

where $\theta = (\gamma, \sigma)$ and K is a non-negative definite matrix. If we put $K = I_p$, $k \times k$ identity matrix, we obtained the simple L_2 regularization term.

Given the data $x_1, \ldots, x_n$, the maximum penalized log-likelihood estimates $\hat{\theta}$ is obtained as the solution to the implicit function

$$\sum_{\alpha=1}^{n}\psi(X_\alpha,\hat{\theta}) = \mathbf{0}, \quad (65)$$

where $\psi = (\psi_1, \cdots, \psi_p)^T$. Note that the penalized maximum likelihood estimator $\hat{\theta}_{PL}$ is obtained by putting

$$\psi(X_\alpha,\hat{\theta}) = \frac{\partial}{\partial\theta}\left\{\log f(X_\alpha|\hat{\theta}_{PL}) - \frac{\lambda}{2}\hat{\gamma}^T K\hat{\gamma}\right\}. \quad (66)$$

In the framework of the generalized information criterion GIC, the information criterion for the model $f(z|\hat{\theta})$ with the estimator $\hat{\theta}$ obtained as the solution of the (65) is given by

$$\mathrm{GIC_M} = -2\sum_{\alpha=1}^{n} \log f(X_\alpha|\hat{\theta}) + 2\mathrm{tr}\left\{J(\psi,\hat{G})^{-1}I(\psi,\hat{G})\right\}, \tag{67}$$

where

$$J(\psi,\hat{G}) = -\frac{1}{n}\sum_{\alpha=1}^{n}\frac{\partial\psi(X_\alpha,\hat{\theta})^T}{\partial\theta}, \qquad I(\psi,\hat{G}) = \frac{1}{n}\sum_{\alpha=1}^{n}\psi(X_\alpha,\hat{\theta})\frac{\partial\log f(X_\alpha|\hat{\theta})}{\partial\theta^T}. \tag{68}$$

6.2 Generalized BIC for Regularization Method

The BIC (Bayesian Information Criterion) proposed by Schwarz (1978)

$$\begin{aligned}\mathrm{BIC} &= -2\log f(\boldsymbol{x}_n|\hat{\boldsymbol{\theta}}) + k\log n \\ &\approx -2\log p(\boldsymbol{x}_n) = -2\log\left\{\int f(\boldsymbol{x}_n|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}\right\}\end{aligned} \tag{69}$$

is a model evaluation criterion based on the posterior probability of a model. Here, $\hat{\boldsymbol{\theta}}_i$ is the maximum likelihood estimator of the k-dimensional parameter vector $\boldsymbol{\theta}$ of the model $f(x|\boldsymbol{\theta})$. Consequently, from the r models that are estimated using the maximum likelihood method, the model that minimizes the value of BIC can be selected as the optimal model.

Konishi et al. (2004) developed generalized Bayesian information criterion, GBIC, for the evaluation of the models obtained by the maximum penalized likelihood method. In this subsection, a simplified version of GBIC is shown briefly. Let $f(x|\hat{\boldsymbol{\theta}}_P)$ be a statistical model estimated by the regularization method for the parametric model $f(x|\boldsymbol{\theta})$, and $\hat{\boldsymbol{\theta}}_P$ is obtained by maximizing the penalized log-likelihood function

$$\ell_\lambda(\boldsymbol{\theta}) = \log f(\boldsymbol{x}_n|\boldsymbol{\theta}) - \frac{n\lambda}{2}\boldsymbol{\theta}^T K\boldsymbol{\theta}, \tag{70}$$

where K is a $p\times p$ matrix. The penalized log-likelihood function can be rewritten as

$$\ell_\lambda(\boldsymbol{\theta}) = \log\left\{f(\boldsymbol{x}_n|\boldsymbol{\theta})\exp\left(-\frac{n\lambda}{2}\boldsymbol{\theta}^T K\boldsymbol{\theta}\right)\right\}. \tag{71}$$

Then, $\exp(-n\lambda/2\theta^T K\theta)$ in the above equation can be thought of as a prior distribution in which the smoothing parameter λ is a hyper-parameter,

$$\pi(\boldsymbol{\theta}|\lambda) = \frac{(n\lambda)^{p/2}|K|^{1/2}}{(2\pi)^{p/2}}\exp\left(-\frac{n\lambda}{2}\boldsymbol{\theta}^T K\boldsymbol{\theta}\right). \tag{72}$$

Given the data distribution $f(\boldsymbol{x}_n|\boldsymbol{\theta})$, and the prior distribution $\pi(\boldsymbol{\theta}|\lambda)$ with hyper-parameter λ, the marginal likelihood of the model can be rewritten as

$$\begin{aligned} p(\boldsymbol{x}_n|\lambda) &= \int f(\boldsymbol{x}_n|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\lambda)d\boldsymbol{\theta} \\ &= \int \exp\{nq(\boldsymbol{\theta}|\lambda)\}\, d\boldsymbol{\theta}, \end{aligned} \tag{73}$$

where

$$\begin{aligned} q(\boldsymbol{\theta}|\lambda) &= \frac{1}{n}\log\{f(\boldsymbol{x}_n|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\lambda)\} = \frac{1}{n}\{\log f(\boldsymbol{x}_n|\boldsymbol{\theta}) + \log\pi(\boldsymbol{\theta}|\lambda)\} \\ &= \frac{1}{n}\Big\{\log f(\boldsymbol{x}_n|\boldsymbol{\theta}) - \frac{n\lambda}{2}\boldsymbol{\theta}^T K\boldsymbol{\theta}\Big\} - \frac{1}{2n}\Big\{p\log(2\pi) - p\log(n\lambda) - \log|K|\Big\}. \end{aligned} \tag{74}$$

We note here that the mode, $\hat{\boldsymbol{\theta}}_P$, of $q(\boldsymbol{\theta}|\lambda)$ coincides with a solution obtained by maximizing the penalized log-likelihood function (70). By approximating it using Laplace's method for integrals, we have

$$\int \exp\{nq(\boldsymbol{\theta})\}d\boldsymbol{\theta} \approx \frac{(2\pi)^{p/2}}{n^{p/2}|J_\lambda(\hat{\boldsymbol{\theta}}_P)|^{1/2}} \exp\Big\{nq(\hat{\boldsymbol{\theta}}_P)\Big\}. \tag{75}$$

where

$$J_\lambda(\hat{\boldsymbol{\theta}}_P) = -\frac{1}{n}\frac{\partial^2 q(\boldsymbol{\theta}|\lambda)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\Big|_{\hat{\boldsymbol{\theta}}_P} = -\frac{1}{n}\frac{\partial^2 \log f(\boldsymbol{x}_n|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\Big|_{\hat{\boldsymbol{\theta}}_P} + \lambda K. \tag{76}$$

Taking the logarithm of this expression and multiplying it by -2, we obtain the generalized Bayesian information criterion GBIC (Konishi et al. 2004; Konishi and Kitagwa 2008),

$$\text{GBIC} = -2\log f(\boldsymbol{x}_n|\hat{\boldsymbol{\theta}}_P) + n\lambda\hat{\boldsymbol{\theta}}_P^T K\hat{\boldsymbol{\theta}}_P + \log|J_\lambda(\hat{\boldsymbol{\theta}}_P)| - p\log\lambda - \log|K|. \tag{77}$$

In the modeling by regularization method, the selection of the smoothing parameter λ is crucial and we select the λ that minimizes the GBIC as the optimal smoothing parameter.

By interpreting the regularization method based on the above argument from a Bayesian point of view, it can be understood that the regularized estimator agrees with the estimate that is obtained through the maximization (mode) of the following posterior probability depending on the value of the smoothing parameter;

$$\pi(\boldsymbol{\theta}|\boldsymbol{x}_n;\lambda) = \frac{f(\boldsymbol{x}_n|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\lambda)}{\int f(\boldsymbol{x}_n|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\lambda)d\boldsymbol{\theta}}, \tag{78}$$

where $\pi(\boldsymbol{\theta}|\lambda)$ is the density function resulting from (72) as a prior probability of the p-dimensional parameter $\boldsymbol{\theta}$ for the model $f(\boldsymbol{x}_n|\boldsymbol{\theta})$. For the Bayesian justification of the maximum penalized likelihood approach, we refer to Silverman (1985) and Wahba (1990).

Example: Regularization for the regression models. Suppose that n observations $\{(\boldsymbol{x}_\alpha, y_\alpha);\ \ \alpha = 1, 2, \cdots, n\}$ are observed in terms of a p-dimensional explanatory variable $\boldsymbol{x}$ and a response variable Y, and consider a simple regression model

$$y_\alpha = \sum_{j=1}^{p} \beta_j x_{\alpha j} + \varepsilon_\alpha, \quad \varepsilon_\alpha \sim N(0, \sigma^2), \tag{79}$$

where $\beta^T = (\beta_1, \ldots, \beta_m)$, $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \sigma^2)^T$ and $(y_\alpha, x_{\alpha 1}, \ldots, x_{\alpha,p})$, $\alpha = 1, \ldots, n$. If we estimate the parameter vector $\boldsymbol{\theta}$ by maximizing the penalized log-likelihood function (70), the estimators for $\boldsymbol{\beta}$ and σ^2 are respectively given by

$$\hat{\boldsymbol{\beta}} = (X^T X + n\lambda\hat{\sigma}^2 K)^{-1} X^T \boldsymbol{y}, \quad \hat{\sigma}^2 = \frac{1}{n}(\boldsymbol{y} - X\hat{\boldsymbol{\beta}})^T(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}), \tag{80}$$

where X is an $n \times m$ matrix given by $X = (\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n)^T$ and $x_\alpha = (x_{\alpha 1}, \ldots, x_{\alpha p})$.

By applying GBIC in (77), the evaluation criterion for the regularized regression model $f(y_\alpha|\boldsymbol{x}_\alpha; \hat{\boldsymbol{\theta}}_P)$ estimated by the regularization method is given by

$$\text{GBIC} = n\log\hat{\sigma}^2 + n\lambda\hat{\boldsymbol{\beta}}^T K\hat{\boldsymbol{\beta}} + n + n\log(2\pi) + \log|J_\lambda(\hat{\boldsymbol{\theta}}_P)| - \log|K| - m\log\lambda, \tag{81}$$

where $J_\lambda(\hat{\boldsymbol{\theta}}_P)$ is the $(m+1) \times (m+1)$ matrix

$$J_\lambda(\hat{\boldsymbol{\theta}}_P) = \frac{1}{n\hat{\sigma}^2}\begin{bmatrix} X^T X + n\lambda\hat{\sigma}^2 K & \frac{1}{\hat{\sigma}^2}X^T\boldsymbol{e} \\ \frac{1}{\hat{\sigma}^2}\boldsymbol{e}'X & \frac{n}{2\hat{\sigma}^2}, \end{bmatrix} \tag{82}$$

with the n-dimensional residual vector $\boldsymbol{e} = \left(y_1 - \hat{\boldsymbol{\beta}}^T\boldsymbol{x}_1, y_2 - \hat{\boldsymbol{\beta}}^T\boldsymbol{x}_2, \cdots, y_n - \hat{\boldsymbol{\beta}}^T\boldsymbol{x}_n\right)^T$.

6.3 L_1 Regularization and Bridge Regression

In recent years, with the advent of big data, modeling based on the L_1 regularization method has been widely used in many fields of science and technologies. The feature of the L_1 regularization method is that parameter estimation and variable selection can be performed at the same time and it is important as a method of extracting essential information from high dimensional data.

In this subsection, we will consider the evaluation of the bridge regression model. The bridge regression model (Frank and Frieman 1993; Fu 1998) has an L_p regularization term

$$\ell_{\lambda,p}(\beta, \sigma^2) = \ell(\beta, \sigma^2) - \frac{n\lambda}{2}\sum_{j=1}^{p}|\beta_j|^p, \tag{83}$$

and it becomes the ridge regression for $p = 2$ and Lasso for $p = 1$. For $0 < p \leq 1$, bridge regression method can perform the selection of variable and parameter estimation simultaneously. Therefore, the bridge regression can be considered as an estimation method that encompasses many estimation methods.

Kawano (2014) presents the GBIC for the bridge regression model

$$\begin{aligned} \mathrm{GBIC} = n \log \hat{\sigma}^2 + n\lambda \sum_{j \in A} |\hat{\beta}_j|^p + n + n \log(2\pi) + \log |J_\lambda| - 2|A| \log p \\ + 2|A|(1 + \frac{1}{p}) \log 2 - \frac{2|A|}{p} \log(n\lambda) + 2|A| \log \Gamma\left(\frac{1}{p}\right), \end{aligned} \tag{84}$$

where $A = \{j; \hat{\beta}_j \neq 0\}$, and J is the $(|A| + 1) \times (|A| + 1)$ matrix given by

$$J_\lambda(\hat{\boldsymbol{\theta}}_P) = \frac{1}{n\hat{\sigma}^2} \begin{bmatrix} X^T X + n\lambda\hat{\sigma}^2 p(p-1)K & \frac{1}{\hat{\sigma}^2} X^T \boldsymbol{e} \\ \frac{1}{\hat{\sigma}^2} \boldsymbol{e}' X & \frac{n}{2\hat{\sigma}^2}, \end{bmatrix}. \tag{85}$$

For $p < 1$, the influence function cannot differentiate, so GIC can not be directly applied. Matsui and Konishi (2011) use the SCAD penalty function to derive GIC and BIC. In addition, Umezu et al. (2015) derived AIC for the bridge regularization for $1 \geq p < 1$.

7 Summary

Due to the dramatic development of measuring instruments in recent years, a huge amount of large-scale data has been acquired in all research areas. Along with this, research method has changed, and data-driven methods are becoming important as the fourth scientific methodology. In the data-driven approach, the model is built according to the theory, knowledge, data, and further the purpose of the analysis. Once a model is built, useful information can be extracted from the data through the fitted model. In this data-driven method, it is crucial to use a good model. Therefore, the problem of developing good model evaluation criteria is a very important.

This paper outlined the model evaluation criteria such as AIC, GIC, EIC. Which are obtained by bias-correction of the log-likelihood of an estimated model. In particular, GIC can be applied to wide class of estimation procedures such as M-estimators, Bayes models and penalized likelihood methods. Bootstrap based information criterion EIC can be applied to various situation for which analytic methods are difficult to apply. Since L_1 regularization is important in recent data analysis, the evaluation of regularization model is also outlined.

References

Akaike, H.: Information theory and an extension of the maximum likelihood principle. In: Petrov, B.N., Csaki, F. (eds.) 2nd International Symposium in Information Theory, pp. 267–281 (1973). (Reproduced in Breakthroughs in Statistics, vol. I, Foundations and Basic Theory, S. Kots and N.L. Johnson, eds., pp. 610–624. Springer-Verlag, New York, 1992)

Akaike, H.: A new look at the statistical model identification. IEEE Trans. Autom. Control AC **19**, 716–723 (1974)

Akaike, H.: Likelihood and the Bayes procedure. In: Bernardo, N.J., DeGroot, M.H., Lindley, D.V., Smith, A.F.M. (eds.) Bayesian Statistics, Valencia, Spain, pp. 141–166. University Press (1980)

Akaike, H., Kitagawa, G. (eds.): The Practice of Time Series Analysis. Springer-Verlag, New York (1998)

Ayres, I.: Super Crunchers: Why Thinking-By-Numbers is the New Way To Be Smart. Bantam Books, New York (2007)

Bozdogan, H.: Proceeding of the First US/Japan Conference on the Frontiers of Statistical Modeling: An Informational Approach. Kluwer Academic Publishers, Netherlands (1994)

Cavanaugh, J.E., Shumway, R.H.: A bootstrap variant of AIC for state-space model selection. Statistica Sinica **7**, 469–473 (1997)

Davison, A.C.: Approximate predictive likelihood. Biometrika **73**, 323–332 (1986)

Drucker, P.F.: Post-capitalist Society. Routledge, London (1993)

Frank, L.E., Friedman, J.H.: A statistical view of some chemometrics regression tools. Technometrics **35**(2), 109–135 (1993)

Fu, W.J.: Penalized regressions: the bridge versus the Lasso. J. Comput. Graph. Stat. **7**(3), 397–416 (1998)

Good, I.J., Gaskins, R.A.: Nonparametric roughness penalties for probability densities. Biometrika **58**, 255–277 (1971)

Huber, P.J.: The behavior of maximum likelihood estimates under nonstandard conditions. In: Proceedings of the Fifth Berkley Symposium on Statistics, pp. 221–233 (1976)

Ishiguro, M., Sakamoto, Y., Kitagawa, G.: Bootstrapping log likelihood and EIC, an extension of AIC. Ann. Inst. Stat. Math. **49**(3), 411–434 (1997)

Kawano, S.: Selection of tuning parameters in bridge regression models via Bayesian information criterion. Stat. Pap. **55**(4), 1207–1223 (2014)

Kitagawa, G., Gersch, W.: A smoothness priors-state space modeling of time series with trend and seasonality. J. Am. Stat. Assoc. **79**(386), 378–389 (1984)

Kitagawa, G., Gersch, W.: Smoothness Priors Analysis of Time Series. Lecture Notes in Statistics, vol. 116. Springer-Verlag, Heidelberg (1996)

Kitagawa, G., Konishi, S.: Bias and variance reduction techniques for bootstrap information criteria. Ann. Inst. Stat. Math. **62**(1), 209–234 (2010)

Konishi, S., Ando, T., Imoto, S.: Bayesian information criteria and smoothing parameter selection in radial basis function networks. Biometrika **91**(1), 27–43 (2004)

Konishi, S., Kitagawa, G.: Generalized information criteria in model selection. Biometrika **83**(4), 875–890 (1996)

Konishi, S., Kitagawa, G.: Information Criteria and Statistical Modeling. Springer Series in Statistics. Springer, New York (2008)

Kullback, S., Leibler, R.A.: On information and sufficiency. Ann. Stat. **22**(22), 79–86 (1951)

Manyika, J., Chui, M., Bughin, J., Brown, B., Dobbs, R., Roxburgh, C., Byers, A.H.: Big Data: The Next Frontier for Innovation, Competition, and Productivity. McKinsey Global Institute, Washington (2011)

Matsui, H., Konishi, S.: Variable selection for functional regression models via the L1 regularization. Comput. Stat. Data Anal. **55**(12), 3304–3310 (2011)

Schwarz, G.: Estimating the dimension of a model. Ann. Stat. **6**, 461–464 (1978)

Shibata, R.: Statistical aspects of model selection. In: Willems, J.C. (ed.) From Data to Model, pp. 215–240. Springer-Verlag, New York (1989)

Shibata, R.: Bootstrap estimate of Kullback-Leibler information for model selection. Statistica Sinica **7**, 375–394 (1997)

Silverman, B.W.: Some aspects of the spline smoothing approach to nonparametric regression curve fitting (with discussion). J. Roy. Statist. Soc. Ser. B **47**, 1–52 (1985). Akaike's criterion

Takeuchi, K.: Distributions of information statistics and criteria for adequacy of models. Math. Sci. **153**, 12–18 (1976). (in Japanese)

Tierney, L., Kadane, J.B.: Accurate approximations for posterior moments and marginal densities. J. Am. Stat. Assoc. **81**, 82–86 (1986)

Umezu, Y., Shimizu, Y., Masuda, H., Ninomiya, Y.: AIC for non-concave penalized likelihood method. arXiv preprint arXiv:1509.01688 (2015)

Wahba, G.: Improper priors, spline smoothing and the problem of guarding against model errors in regression. J. Roy. Statist. Soc. Ser. B **40**, 364–372 (1978)

Wahba, G.: Spline Models for Observational Data. Philadelphia (1990)

http://www.springer.com/978-3-319-73149-0

Econometrics for Financial Applications

Anh, L.H.; Dong, L.S.; Kreinovich, V.; Thach, N.N. (Eds.)

2018, XIII, 1081 p. 176 illus., Hardcover
ISBN: 978-3-319-73149-0