Journal of Advances in Mathematics and Computer Science

28(3): 1-13, 2018; Article no.JAMCS.43344

ISSN: 2456-9968

(Past name: British Journal of Mathematics & Computer Science, Past ISSN: 2231-0851)

A Simulation Study for the AIC and Likelihood Cross-validation: The Case of Exponential Versus Weibull Distributions

Kunio Takezawa[1*]

[1] *Division of Informatics and Inventory, Institute for Agro-Environmental Sciences, National Agriculture and Food Research Organization, Kannondai 3-1-3, Tsukuba, Ibaraki 305-8604, Japan.*

Author's contribution

The sole author designed, analyzed and interpreted and prepared the manuscript.

Article Information

DOI: 10.9734/JAMCS/2018/43344

Editor(s):

(1) Dragos-Patru Covei, Professor, Department of Applied Mathematics, The Bucharest University of Economic Studies, Piata Romana, Romania.

Reviewers:

(1) Bayo H. Lawal, Kwara State University, Nigeria.

(2) Irshad Ullah, Education Government of Khyber Pakhtunkhwa, Pakistan.

(3) Haruhiko Ogasawara, Otaru University of Commerce, Japan.

Complete Peer review History: http://www.sciencedomain.org/review-history/25863

Original Research Article

Received: 16th July 2018

Accepted: 2nd August 2018

Published: 13th August 2018

Abstract

Various methods are available for choosing statistical models. It is difficult to know which model selection criterion is the best for specific data. This paper discusses a method for choosing the model selection criterion based on the characteristics of the data and models. As an example, we examined the choice between AIC and likelihood cross-validation as the model selection criterion with the exponential distribution and Weibull distribution as candidate models. First, we examined the characteristics of AIC and likelihood cross-validation using data generated from an exponential distribution or Weibull distribution; AIC and likelihood cross-validation show substantially different natures. Next, from the results of the numerical simulations, we propose an intuitive method for deciding whether to use AIC or likelihood cross-validation.

Keywords: AIC; cross-validation; expected log-likelihood; future data; exponential distribution; maximum likelihood estimator; Weibull distribution.

**Corresponding author: E-mail: nonpara@gmail.com;*

2010 Mathematics Subject Classification: 60G25, 62F10, 62M20.

1 Introduction

Model selection using AIC (Akaike's Information Criterion) ([1],[2], Chapter 3 in [3],Chapter 2 and Chapter 7 in [4], Chapter 2 in [5]) coincides asymptotically with selection using cross-validation ([6],[7],[8],[9],[10],[11]); a more recent and sophisticated study on this matter is [12]. However, this is an asymptotic result and is based on several assumptions; when dealing with practical data analysis, it is often not clear whether using AIC is the same as using cross-validation. Even if AIC functions in a similar manner to cross-validation, this does not necessarily mean that AIC gives a good approximation for the expected log-likelihood (Section 5.3 of [13]). To clarify this, we used numerical simulation to examine the characteristics of the two criteria and to construct a way of selecting which one is most in line with the nature of the data and the conditions governing the model selection. Use of real life data example may enhance the persuasiveness of comparison of statistics such as AIC and cross-validation. However, if we use real data without knowing the values of parameters, it is not easy to enable fair and square comparison of statistics.

In using numerical simulation to clarify the characteristics of AIC and cross-validation, whether to choose the exponential distribution or the Weibull distribution as the target distribution poses a problem. We used the results of our numerical simulations to propose a method of choosing the model selection criterion based on the nature of the data.

2 Outline of AIC and Likelihood Cross-validation

The equation for AIC is based on (page 55 in [3]):

$$E_{G(\mathbf{x})}\left[nE_{G(z)}\left[\log f(Z|\hat{\boldsymbol{\theta}}(\mathbf{X}))\right]\right] = E_{G(\mathbf{x})}\left[\log f(\mathbf{X}|\hat{\boldsymbol{\theta}}(\mathbf{X}))\right] - b(G). \tag{2.1}$$

Outline of AIC and likelihood cross-validation

where $\mathbf{x}(=(x_1, x_2, \ldots, x_n)^t)$ (available data, i.e., data at hand, a nonrandom variable) is one set of data consisting of n observations that are generated from the true model ($G(x)$). $\mathbf{x}$ are realizations of $\mathbf{X}(=(X_1, X_2, \ldots, X_n)^t)$. The random variable Z (future data) is an independent copy of X_1(not one set of data) that will be generated from the true model ($G(x)$) in the future. $E_{G(\mathbf{x})}$ stands for the expectation with respect to $\prod_{\alpha=1}^{n} G(x_\alpha) = G(\mathbf{x})$ (a joint distribution) that generates the available data and future data. $E_{G(z)}$ is the expectation with respect to $G(z)$ (true model, i.e., true probability distribution). $\hat{\boldsymbol{\theta}}(\mathbf{X})$ is the parameter estimator given by the maximum likelihood method using $\mathbf{X}$ as data. Then $E_{G(z)}\left[\log f(Z|\hat{\boldsymbol{\theta}}(\mathbf{X}))\right]$ is the expectation of the log-likelihood of the parameter estimator in the light of future data; the parameter estimator is given by the maximum likelihood method in the light of the available data. $b(G)$ is called "the bias of log-likelihood as an estimator of the expected log-likelihood" in (page 55 in [3]). Hence, the left-hand side of Eq. (2.1) is given by calculating the expectation of the log-likelihood of the available data in the light of future data and obtaining the expectation of the result; the former expectation is taken with respect to future data, and the latter expectation is taken with respect to available data. If a larger value is yielded by this computation, we draw the conclusion that the maximum likelihood method gives a better model, and therefore we should construct a model that makes this value larger.

$\log f(\mathbf{X}|\hat{\boldsymbol{\theta}}(\mathbf{X}))$ is the log-likelihood in the light of one set of available data consisting of n observations.

$\log f(\mathbf{X}|\hat{\boldsymbol{\theta}}(\mathbf{X}))$ is written formally (page 55 in [3]) as

$$\log f(\mathbf{X}|\hat{\boldsymbol{\theta}}(\mathbf{X})) = \sum_{\alpha=1}^{n} \log f(x_\alpha|\hat{\boldsymbol{\theta}}(\mathbf{X})). \tag{2.2}$$

That is, the left-hand side of this equation is the sum of the values of the log-likelihood of $\hat{\boldsymbol{\theta}}(\mathbf{X})$ in the light of one set of available data consisting of n observations.

If the values of the left-hand side of Eq. (2.1) for various models are estimated with high accuracy, an approximation to the right-hand side of Eq. (2.1) can be obtained. Then, a model which makes this value large is considered to give a large log-likelihood in the light of future data. However, the value given by the left-hand side of Eq. (2.1) cannot be obtained unless we have an infinite number of future data. Hence, we need to estimate the approximate value of the right-hand side of Eq. (2.1). It should be noted that $\log f(\mathbf{x}|\hat{\boldsymbol{\theta}}(\mathbf{x}))$ (i.e., realization of the value given by the right-hand side of Eq. (2.2)) is yielded by the available data (one set of data consisting of n observations) using the maximum likelihood method. Moreover, $E_{G(\mathbf{x})}\left[\log f(\mathbf{X}|\hat{\boldsymbol{\theta}}(\mathbf{X}))\right]$ is approximately identical to $\log f(\mathbf{x}|\hat{\boldsymbol{\theta}}(\mathbf{x}))$. Therefore, if the value of $b(G)$ is estimated with high accuracy, an accurate approximation to the left-hand side of Eq. (2.1) can be obtained. Therefore, if our purpose is to select an efficient model, we should aim to estimate the value of $b(G)$.

AIC adopts the equation

$$b(G) = p. \tag{2.3}$$

where p is the number of parameters contained in a model. Since Eq. (2.3) is derived by an analytical approximation procedure, usually we cannot estimate the accuracy of the approximation in our practical data analysis, so it is necessary to carry out a numerical simulation.

The approximation below is adopted in likelihood cross-validation (e.g., Eq. (3.43) on page 53 in [14],[15]).

$$nE_{G(z)}\left[\log f(Z|\hat{\boldsymbol{\theta}}(\mathbf{X}))\right] \approx \sum_{j=1}^{n} \log f(x_j|\hat{\boldsymbol{\theta}}(\mathbf{x}_{n(-j)})). \tag{2.4}$$

where $\log f(x_j|\hat{\boldsymbol{\theta}}(\mathbf{x}_{n(-j)}))$ is the log-likelihood of the estimator in the light of x_j; the maximum likelihood method derives these estimates using data which is given by deleting x_j from $\mathbf{x}$. This calculation is iterated by using each value of $j = 1, 2, \ldots, n$. Then, the resultant values are added to give an approximation to the left-hand side of Eq. (2.1). This is the procedure for likelihood cross-validation.

3 AIC and Likelihood Cross-validation in Exponential and Weibull Distributions

The probability density function of the exponential distribution is:

$$f_E(x) = \lambda \exp(-\lambda x). \tag{3.1}$$

where λ is a parameter. Then, the log-likelihood in the light of available data ($(x_1, x_2, \ldots, x_n)^t$) can be written as

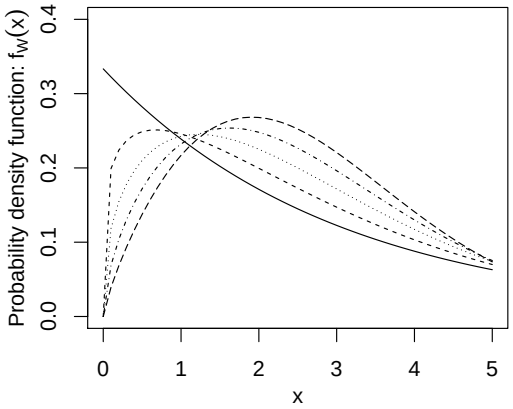

Fig. 1. Probability density function of the Weibull distribution; solid line: $\beta = 1.0$**, dashed line:** $\beta = 1.2$**, dotted line:** $\beta = 1.4$**, dot-dashed line:** $\beta = 1.6$**, long-dashed line:** $\beta = 1.8$**;** $\eta = 3$ **for all distributions**

$$l(\{x_i\}|\lambda) = n\log(\lambda) - \lambda\sum_{i=1}^{n} x_i. \tag{3.2}$$

Hence, the maximum likelihood estimator is:

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i}. \tag{3.3}$$

The probability density function of the Weibull distribution is as follows.

$$f_W(x) = \frac{\beta}{\eta}\left(\frac{x}{\eta}\right)^{\beta-1}\exp\left(-\left(\frac{x}{\eta}\right)^{\beta}\right). \tag{3.4}$$

where β and η are parameters. Fig. 1. shows the probability density function of the Weibull distribution when $\eta = 3$ and $\beta = 1.0, 1.2, 1, 4, 1.6, 1.8$. When $\beta = 1$, we have

$$f_W(x) = \frac{1}{\eta}\exp\left(-\frac{x}{\eta}\right). \tag{3.5}$$

Comparison of this equation with Eq. (3.1) indicates that the equation is the probability density function of the exponential distribution. Hence, the exponential distribution is a specific case of the Weibull distribution.

The log-likelihood of Eq. (3.4) in the light of the data x_i is as follows.

$$\begin{aligned}\log(f_W(x_i)) &= \log(\beta) - \log(\eta) + (\beta-1)\log(x_i) - (\beta-1)\log(\eta) - \left(\frac{x_i}{\eta}\right)^{\beta} \\ &= \log(\beta) - \beta\log(\eta) + (\beta-1)\log(x_i) - \left(\frac{x_i}{\eta}\right)^{\beta}. \end{aligned}\tag{3.6}$$

Therefore, the log-likelihood in the light of the available data $((x_1, x_2, \ldots, x_n)^t)$ can be written as

$$l(\{x_i\}|\eta,\beta) = n\log(\beta) - n\beta\log(\eta) + (\beta-1)\sum_{i=1}^{n}\log(x_i) - \frac{1}{\eta^{\beta}}\sum_{i=1}^{n} x_i^{\beta}. \tag{3.7}$$

Taking the derivative of Eq. (3.7) with respect to η and setting it equal to 0 yields

$$\frac{\partial l(\{x_i\}|\eta,\beta)}{\partial \eta} = -\frac{n\beta}{\eta} + \frac{\beta}{\eta^{\beta+1}}\sum_{i=1}^{n} x_i^{\beta} = 0. \tag{3.8}$$

Hence, we have the equation

$$\eta = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^{\beta}\right)^{\frac{1}{\beta}}. \tag{3.9}$$

That is,

$$\log(\eta) = \frac{1}{\beta}\log\left(\frac{1}{n}\sum_{i=1}^{n} x_i^{\beta}\right). \tag{3.10}$$

However, differentiating Eq. (3.7) with respect to β and setting it equal to 0 gives

$$\frac{\partial l(\{x_i\}|\eta,\beta)}{\partial \beta} = \frac{n}{\beta} - n\log(\eta) + \sum_{i=1}^{n}\log(x_i) + \frac{\log(\eta)}{\eta^{\beta}}\sum_{i=1}^{n} x_i^{\beta} - \frac{1}{\eta^{\beta}}\sum_{i=1}^{n}\log(x_i)x_i^{\beta} = 0. \tag{3.11}$$

Substituting Eq. (3.9) and Eq. (3.10) into Eq. (3.11) yields

$$\begin{aligned}\frac{\partial l(\{x_i\}|\eta,\beta)}{\partial \beta} &= \frac{n}{\beta} - n\frac{1}{\beta}\log\left(\frac{1}{n}\sum_{i=1}^{n} x_i^{\beta}\right) + \sum_{i=1}^{n}\log(x_i) \\ &+ \frac{\frac{1}{\beta}\log\left(\frac{1}{n}\sum_{i=1}^{n} x_i^{\beta}\right)}{\frac{1}{n}\sum_{i=1}^{n} x_i^{\beta}}\sum_{i=1}^{n} x_i^{\beta} - \frac{1}{\left(\frac{1}{n}\sum_{i=1}^{n} x_i^{\beta}\right)}\sum_{i=1}^{n}\log(x_i)x_i^{\beta} \\ &= \frac{n}{\beta} + \sum_{i=1}^{n}\log(x_i) - \frac{1}{\left(\frac{1}{n}\sum_{i=1}^{n} x_i^{\beta}\right)}\sum_{i=1}^{n}\log(x_i)x_i^{\beta} = 0.\end{aligned} \tag{3.12}$$

This equation contains β but not η. Hence, by solving an equation with one variable, we obtain the estimate for β ($= \hat{\beta}$). An estimate of η ($= \hat{\eta}$) can be derived by substituting the resultant $\hat{\beta}$ into Eq. (3.8).

4 True $b(G)$, $b(G)$ Given by AIC, and $b(G)$ Given by Likelihood Cross-validation

The $b(G)$ given by AIC and the $b(G)$ given by likelihood cross-validation were compared with the real $b(G)$. To achieve this, the value of the left-hand side of Eq. (2.1) when the data are sampled from a Weibull distribution was estimated using three methods: (1) use of real future data; (2) use of AIC; (3) use of likelihood cross-validation. Numerical simulations were carried out to compare these three estimates.

First, when we suppose that future data are at hand, an approximation to the value of the left-hand side of Eq. (2.1) can be calculated. It is approximated as follows.

$$E_{G(\mathbf{x})}\left[nE_{G(z)}\left[\log f_W(Z|\hat{\boldsymbol{\theta}}(\mathbf{X}))\right]\right] \approx \frac{1}{S}\sum_{s=1}^{S} n\frac{1}{M}\sum_{m=1}^{M}\log f_W(x_m^*|\hat{\eta}(\mathbf{x}_s),\hat{\beta}(\mathbf{x}_s)). \tag{4.1}$$

where the available data are one of $\{\mathbf{x}_s\}$ ($s = 1,2,3\ldots,S$). $\{x_m^*\}$ ($m = 1,2,3,\ldots,M$) are future data that are sampled from the same distribution as the available data, although they are

independent of the available data. $\log f_W(x_m^*|\hat{\eta}(\mathbf{x}_s), \hat{\beta}(\mathbf{x}_s))$ is the log-likelihood given by substituting x_m^* into the probability density function containing $\hat{\eta}$ and $\hat{\beta}$; $\hat{\eta}$ and $\hat{\beta}$ are obtained by the maximum likelihood method using $\mathbf{x}_s$. That is, the log-likelihood of the probability density function containing $\hat{\eta}$ and $\hat{\beta}$ is calculated in the light of x_m^*. When fitting an exponential distribution, $f_W(\cdot)$ is replaced with $f_E(\cdot)$.

If AIC is used, Eq. (4.1) is replaced with the equation

$$E_{G(\mathbf{x})}\Big[nE_{G(z)}\Big[\log f_W(Z|\hat{\boldsymbol{\theta}}(\mathbf{X}))\Big]\Big] \approx \frac{1}{S}\sum_{s=1}^{S}\Big(\log f_W(\mathbf{x}_s|\hat{\eta}(\mathbf{x}_s), \hat{\beta}(\mathbf{x}_s)) - 2\Big). \tag{4.2}$$

where the approximation below is used.

$$E_{G(\mathbf{x})}\Big[\log f(\mathbf{X}|\hat{\boldsymbol{\theta}}(\mathbf{X}))\Big] \approx \frac{1}{S}\sum_{s=1}^{S}\log f_W(\mathbf{x}_s|\hat{\eta}(\mathbf{x}_s), \hat{\beta}(\mathbf{x}_s)). \tag{4.3}$$

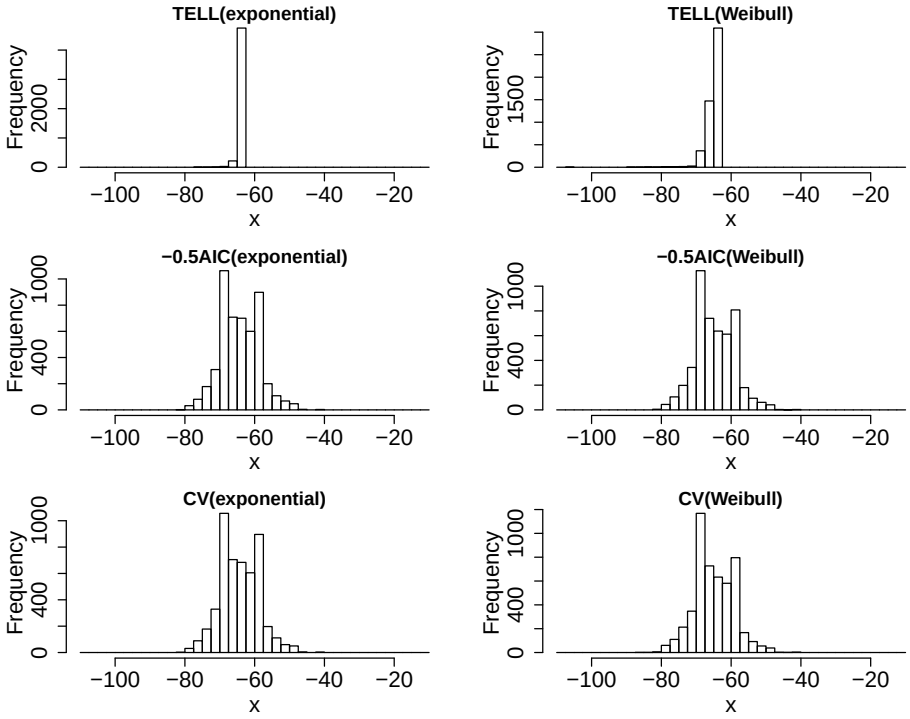

Fig. 2. Value of the left-hand side of Eq. (2.1); this value is provided by numerical simulation with the setting $n = 30$, $\eta = 3$, $\beta = 1$. The graph at top left shows the values of the right-hand side of Eq. (4.1) when fitting an exponential distribution. The graph at top right shows the values of the right-hand side of Eq. (4.1) when fitting a Weibull distribution. The graph on the middle left shows the values of the right-hand side of Eq. (4.2) when fitting an exponential distribution. The graph on the middle right shows the values of the right-hand side of Eq. (4.2) when fitting a Weibull distribution. The graph at bottom right shows the values of the right-hand side of Eq. (4.4) when fitting an exponential distribution. The graph at bottom left shows the values of the right-hand side of Eq. (4.4) when fitting a Weibull distribution.

$\log f_W(\mathbf{x}_s|\hat{\eta}(\mathbf{x}_s),\hat{\beta}(\mathbf{x}_s))$ is the log-likelihood given by substituting $\mathbf{x}_s$ (available data) into the probability density function with $\hat{\eta}$ and $\hat{\beta}$; $\hat{\eta}$ and $\hat{\beta}$ are obtained by the maximum likelihood method using $\mathbf{x}_s$. That is, the log-likelihood of the probability density function with $\hat{\eta}$ and $\hat{\beta}$ is obtained in the light of $\mathbf{x}_s$. (-2) in the right-hand side is due to the fact that the Weibull distribution contains two parameters. If fitting an exponential distribution, $f_W(\cdot)$ is replaced with $f_E(\cdot)$ and (-2) in the right-hand side is altered to (-1).

When likelihood cross-validation is used, Eq. (4.1) is replaced with

$$E_{G(\mathbf{x})}\Big[nE_{G(z)}\Big[\log f_W(Z|\hat{\boldsymbol{\theta}}(\mathbf{X}))\Big]\Big] \approx \frac{1}{S}\sum_{s=1}^{S}\Big(\sum_{j=1}^{n}\log f(x_{sj}|\hat{\eta}(\mathbf{x}_{s(-j)},\hat{\beta}(\mathbf{x}_{s(-j)}))\Big). \tag{4.4}$$

where x_{sj} is the j-th element of $\mathbf{x}_s$. The maximum likelihood method using $\mathbf{x}_{s(-j)}$ leads to $\hat{\eta}(\mathbf{x}_{s(-j)})$ (the estimator of η). The maximum likelihood method using $\mathbf{x}_{s(-j)}$ provides $\hat{\beta}(\mathbf{x}_{s(-j)})$ (the estimator of $b\eta$).

Fig. 2. shows the distributions of the approximated values of $nE_{G(z)}\Big[\log f_W(Z|\hat{\boldsymbol{\theta}}(\mathbf{X}))\Big]$ when numerical simulations were carried out with the setting $n = 30$, $\eta = 3$, $\beta = 1$, $S = 5,000$, and $M = 1,000$. That is, one set of data ($n = 30$) was generated as available data to estimate parameters, and the log-likelihood in the light of future data was calculated using three methods: (1) the log-likelihood of the estimated parameters in the light of future data calculated using real future data consisting of $1,000$ data (Eq. (4.1)); (2) use of AIC (Eq. (4.2)); (3) use of likelihood cross-validation (Eq. (4.4)). This graphs show the distributions of the estimates given by $5,000$ simulations using different pseudo random numbers as seeds. The R command "optimize()" implemented in R version 3.4.0 was used to derive $\hat{\beta}$ by fitting a Weibull distribution (Eq. (3.12)).

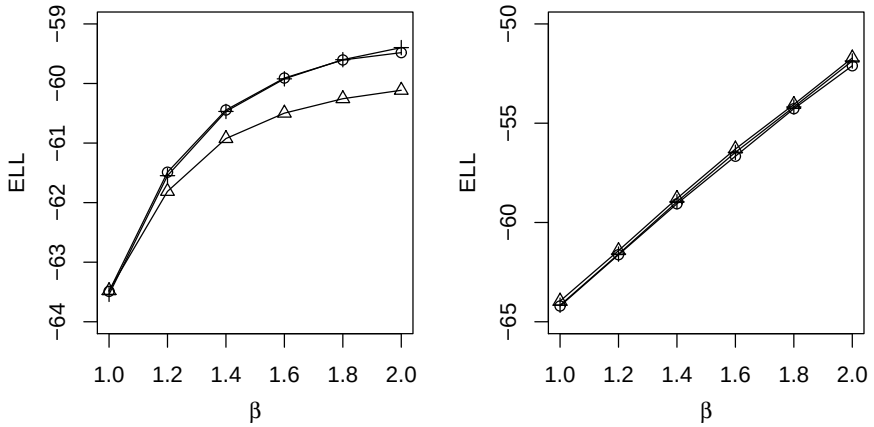

Fig. 3. Distributions of the real and approximated values of $nE_{G(z)}\Big[\log f_W(Z|\hat{\boldsymbol{\theta}}(\mathbf{X}))\Big]$. "◯" represents real values. "△" represents the estimates given by AIC. "+" represents the estimates given by likelihood cross-validation. The graph on the left shows the values given by fitting an exponential distribution. The graph on the right shows the values given by fitting a Weibull distribution.

Fig. 3. (left) shows the mean of the distributions of the real and approximated values of $nE_{G(z)}\Big[\log f_E(Z|\hat{\boldsymbol{\theta}}(\mathbf{X}))\Big]$. The value of β is set to one of $\beta = \{1, 1.2, 1.4, 1.6, 1.8, 2\}$. The other settings are the same as those used in Fig. 2. Thus this graph shows a comparison of the real values

with the approximated values using AIC (i.e., AIC·(−0.5)), and with the approximated values using likelihood cross-validation when fitting an exponential distribution. This graph indicates that the mean of the expected log-likelihood given by AIC gradually moves away from the mean of the true value of the expected log-likelihood. This phenomenon shows that AIC·(−0.5) functions less efficiently as an approximation for the expected log-likelihood when the value of β gets further away from 1 (i.e., the distribution becomes less like an exponential distribution.); when the value of β is 1, the generated data form an exponential distribution.

Fig. 3. (right) shows the results of fitting a Weibull distribution with the same conditions as in Fig. 3 (left). This graph indicates that nearly identical estimates are obtained by the three methods: (1) the true mean of the expected log-likelihood; (2) the mean of the expected log-likelihood given by AIC; (3) the mean of the expected log-likelihood given by likelihood cross-validation.

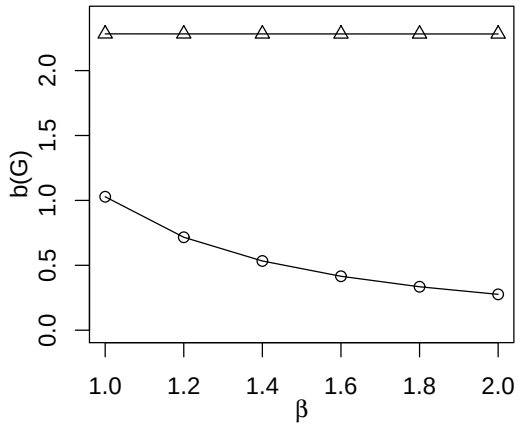

Fig. 4. Estimates of $b(G)$. Data are generated using one of $\beta = \{1, 1.2, 1.4, 1.6, 1.8, 2\}$ to fit an exponential distribution or a Weibull distribution.

Next, the true value of $b(G)$ (Eq. (2.1)) is estimated using Eq. (4.1) and Eq. (4.2) as follows.

$$\begin{aligned} b(G) &= E_{G(\mathbf{x})}\left[\log f(\mathbf{X}|\hat{\boldsymbol{\theta}}(\mathbf{X}))\right] - E_{G(\mathbf{x})}\left[nE_{G(z)}\left[\log f(Z|\hat{\boldsymbol{\theta}}(\mathbf{X}))\right]\right] \\ &\approx \frac{1}{S}\sum_{s=1}^{S}\log f_W(\mathbf{x}_s|\hat{\eta}(\mathbf{x}_s),\hat{\beta}(\mathbf{x}_s)) - \frac{1}{S}\sum_{s=1}^{S}\frac{n}{M}\sum_{m=1}^{M}\log f_W(x_m^*|\hat{\eta}(\mathbf{x}_s),\hat{\beta}(\mathbf{x}_s)). \end{aligned} \tag{4.5}$$

When an exponential distribution is employed, $f_W(\cdot)$ is replaced with $f_E(\cdot)$. Fig. 4 shows the estimates of $b(G)$ when the data are generated using one of $\beta = \{1, 1.2, 1.4, 1.6, 1.8, 2\}$ to fit an exponential distribution or a Weibull distribution. When AIC is used, we suppose $b(G) = 1$ for fitting an exponential distribution whatever the value of β may be. However, we suppose $b(G) = 2$ for fitting a Weibull distribution whatever the value of β may be. Therefore, the results of the numerical simulation indicate that the $b(G)$ given by AIC is a little less than the true value of $b(G)$ when the realizations of a Weibull distribution are fitted to Weibull distribution. That is, the $b(G) = 2$ given by AIC is slightly biased.

In contrast, when realizations of Weibull distribution with the setting $\beta = 1$ are fitted to an exponential distribution, the $b(G) = 1$ given by AIC is very close to the real value because the Weibull distribution is identical to the exponential distribution when we set $\beta = 1$. However, as

the value of β gets gradually larger than 1, the real value of $b(G)$ becomes less than 1. That is, the $b(G) = 1$ given by AIC becomes more biased. Some existing literature emphasizes that AIC works appropriately as an approximation for the expected log-likelihood even if the applicant model does not contain the data-generating model as a special case (page 369 in [4]). However, the results shown in Fig. 4. indicate that if we cannot assume, even approximately, that the applicant model (i.e., exponential distribution in this example) is identical to the data-generating model (Weibull distribution, which is somewhat different from the exponential distribution) or contains it as a special case, AIC$\cdot(-0.5)$ may not be regarded as a good approximation for the expected log-likelihood.

A similar situation occurs when a linear equation is fitted to the data generated from a linear equation with normal noise using the least squares method. That is, when a constant (special case of a linear equation) is fitted using least squares to the data generated from a linear equation with normal noise, $b(G)$ is not 1 but a value close to 0 (page 233 in [13]). That is, in these two examples, when the applicant model is substantially different from the data-generated model, the $b(G)$ given by AIC is larger than the true value. Therefore, the expected log-likelihood given by AIC is smaller than the true value of the expected log-likelihood. This tendency leads to an excusable mistake, because when the data-generated model is considerably different from the applicant model, that model is not selected by AIC. This theoretical consideration implies that this phenomenon may often occur in our real data analysis without knowing that it is the case. We do not know, however, whether this relationship always holds. Moreover, if all of the applicant models are considerably different from the true model, the values of AIC$\cdot(-0.5)$ are not good approximations for the expected log-likelihood. Hence, we are not sure that model comparison using AIC is appropriate. Since we do not know whether the applicant model contains a model that is very close to the true model, it is very difficult to decide whether model comparison using AIC leads to desirable results.

5 Selection Between AIC and Likelihood Cross-validation

The results of the numerical simulations showed that the characteristics of AIC and likelihood cross-validation were somewhat different. That is, although the value of $b(G)$ (Eq. (2.2)) has zero variance, it could be biased to a certain extent. However, $b(G)$ estimated by likelihood cross-validation has little if any bias. Therefore, we need to consider whether we should use AIC or likelihood cross-validation as the criterion.

We propose an intuitive method to choose between AIC (i.e., AIC$\cdot(-0.5)$) and likelihood cross-validation to approximate the value of the expected log-likelihood. Our method estimates the confidence interval of the expected log-likelihood given by likelihood cross-validation using a bootstraps method. Then, if the resulting confidence interval does not contain the expected log-likelihood estimated using AIC, we conclude that the expected log-likelihood estimated using AIC is not correct, and we adopt the expected log-likelihood given by likelihood cross-validation. However, if the resultant confidence interval contains the expected log-likelihood estimated using AIC, we do not reject the assumption that the expected log-likelihood estimated using AIC is correct, and we adopt it as an approximation for the expected log-likelihood.

Formally, this algorithm is written as:
(1) Obtain B bootstrap data ($\{\mathbf{x}^{(b)}\}$ ($\mathbf{x}^{(b)} = (x_1^{(b)}, x_2^{(b)}, \ldots, x_n^{(b)})^t$, $b = 1, 2, \ldots, B$)) by sampling n observations at random with replacement from $\mathbf{x}(= (x_1, x_2, \ldots, x_n)^t)$ (available data) B times.
(2) Calculate the expected log-likelihood ($ELL^{(b)}$) given by $\{\mathbf{x}^{(b)}\}$ using the equation below, which is obtained from Eq. (4.4).

$$ELL^{(b)} = \sum_{j=1}^{n} \log f(x_j^{(b)} | \hat{\eta}(\mathbf{x}_{(-j)}^{(b)}), \hat{\beta}(\mathbf{x}_{(-j)}^{(b)})). \tag{5.1}$$

where $\mathbf{x}_{(-j)}^{(b)}$ is obtained by deleting the j-th data from $\mathbf{x}^{(b)}$.
(3) Sort $\{ELL^{(b)}\}$ as $ELL^{(1)} \leq ELL^{(2)} \leq \cdots \leq ELL^{(B)}$.
(4) For example, when $B = 1,000$ is set, we assume that the interval between $ELL^{(25)}$ and $ELL^{(976)}$ represents the confidence interval of the expected log-likelihood estimated by likelihood cross-validation.

(5) If the expected log-likelihood estimated by AIC is located in the confidence interval derived in (4), we accept that the validity of the expected log-likelihood estimated by AIC is not rejected. Hence, we adopt the expected log-likelihood estimated by AIC. However, if the expected log-likelihood estimated by AIC is not positioned in the confidence interval derived in (4), we conclude that the validity of the expected log-likelihood estimated by AIC is rejected. Hence, we adopt the expected log-likelihood estimated by likelihood cross-validation.

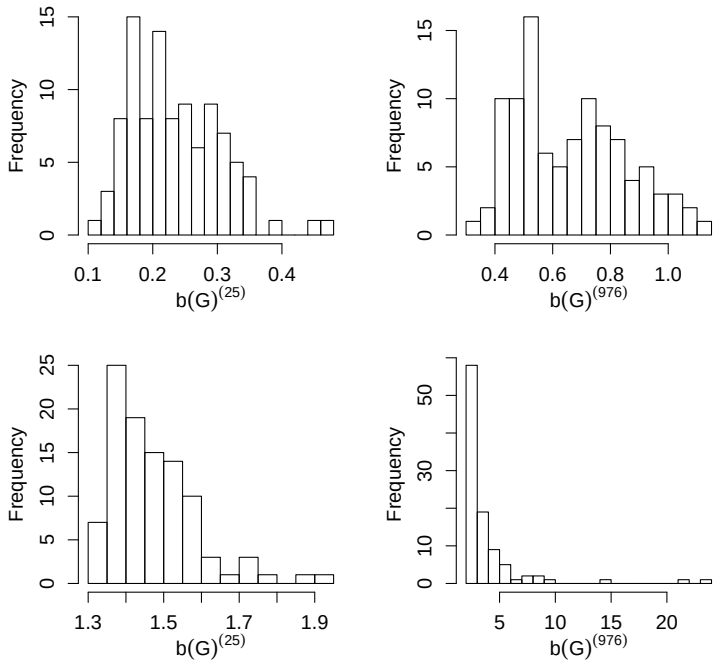

Fig. 5. Distribution of $b(G)^{(25)}$(top-left graph) and $b(G)^{(976)}$(top right) when $\beta = 1.6$ fits an exponential distribution. Distribution of $b(G)^{(25)}$(bottom left) and $b(G)^{(976)}$(bottom right) when $\beta = 1.6$ is set fits a Weibull distribution.

Numerical simulations were carried out 100 times using the above method, varying the initial value of the pseudo random numbers. One of the values of $\beta = 1, 1.2, 1, 4, 1.6, 1.8, 2$ was used. Fig 5 illustrates the distributions of $b(G)^{(25)}$ and $b(G)^{(976)}$ with the setting $\beta = 1.6$. $b(G)^{(25)}$ and $b(G)^{(976)}$ are defined as follows using Eq. (4.2).

$$b(G)^{(25)} = \frac{1}{S}\sum_{s=1}^{S} \log f_W(\mathbf{x}_s|\hat{\eta}(\mathbf{x}_s), \hat{\beta}(\mathbf{x}_s)) - ELL^{(25)}. \tag{5.2}$$

$$b(G)^{(976)} = \frac{1}{S}\sum_{s=1}^{S} \log f_W(\mathbf{x}_s|\hat{\eta}(\mathbf{x}_s), \hat{\beta}(\mathbf{x}_s)) - ELL^{(976)}. \tag{5.3}$$

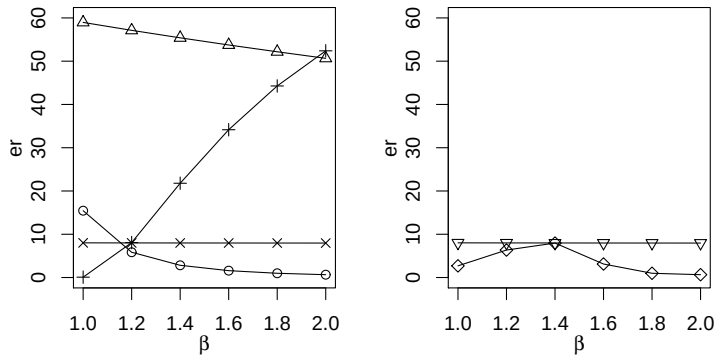

Fig. 6. Values of er(Eq. (5.4)). "◯" indicates the values of er when the data fit an exponential distribution and $b(G)$ is obtained using likelihood cross-validation. "△'" indicates the values of er when the data fit a Weibull distribution and $b(G)$ is obtained using likelihood cross-validation. "+" indicates the values of er when the data fit an exponential distribution and $b(G)$ is obtained using AIC. "×" indicates the values of er when the data fits a Weibull distribution and $b(G)$ is obtained using AIC. These four symbols are used in the left-hand graph. "◇" indicates the values of er when the data fits an exponential distribution and $b(G)$ is obtained using the suggested method. "▽" indicates the values of er when the data fits a Weibull distribution and $b(G)$ is obtained using the suggested method. These two symbols are used in the right-hand graph.

We compared the value of $b(G)$ estimated by the above method, which used AIC, and the value estimated by likelihood cross-validation by calculating the value defined below.

$$er = \sum_{k=1}^{100} (b(G)_k - \tilde{b}(G))^2. \tag{5.4}$$

where $\tilde{b}(G)$ represents the true value of $b(G)$, given by Eq. (4.1). $b(G)_k$ is the estimate derived using the above method; either AIC or likelihood cross-validation. Fig. 6. shows the values of er with the same settings as in the previous numerical simulation. These two graphs indicate that this method considerably reduces the value of er for both the exponential distribution and the Weibull distribution regardless of whether $\beta = 1, 1.2, 1, 4, 1.6, 1.8, 2$. Although this is a rather intuitive method without a clear background, the result of this numerical simulation implies the possibility of choice from among conventional model selection methods.

6 Conclusion

In this paper, we have discussed a method that allows us to choose between AIC and likelihood-cross validation in dealing with the model selection problem of whether the data fit a Weibull distribution or an exponential distribution. The results of the numerical simulation demonstrate that the characteristics of the two criteria are substantially different when the number of data is small. In light of this, we should consider the characteristics of the data and the applicable models when deciding which of the two methods to use. We propose an intuitive method based on the concept of confidence intervals to choose between the two methods.

Regarding the use of AIC, the available literature tells that "Akaike (1974) [2] stated that if the true distribution that generated the data exists near the specified parametric model, the bias associated

with the log-likelihood of the model based on the maximum likelihood method can be approximated by the number of parameters." (page 61 in [3]) However, this point is seldom paid attention to in actual data analysis, and if we were fully aware of this point in our model selection procedure we would seldom adopt AIC as the model selection criterion. This is because if AIC cannot be used unless all of applicable models are sure to be "near the specified parametric model," we need to know the quantitative definition of "near the specified parametric model.". Moreover, in most situations of data analysis, we do not know the exact appearance of "the true distribution." Therefore AIC is usually used without being conscious of the conditions for using it as an approximation for the expected log-likelihood.

The simple numerical simulations carried out here, however, indicates that AIC$\cdot(-0.5)$ is somewhat biased as an approximation for the expected log-likelihood. In this respect, likelihood cross-validation is preferable to AIC. Hence, we suppose that if we use AIC$\cdot(-0.5)$ unconditionally as an approximation for the expected log-likelihood, we may derive inappropriate results as a quantitative estimate to show the validity of the model. It should be noted, however, that even if AIC$\cdot(-0.5)$ is not regarded as an approximation for the expected log-likelihood, AIC performs well as a model selection criterion in some situations (page 242 in [13]).

In the age when it was difficult to examine the characteristics of estimators using numerical simulation because of relatively poor computer power, we had to investigate the characteristics of estimators with the assumption that the number of data to obtain asymptotic results is very large. In the current age of powerful computers, we can examine the characteristics of estimators from various perspectives by simulating the actual data analysis almost perfectly. This kind of research is unveiling the various aspects of estimators, most of which have not been shown by asymptotic studies. We should adopt this approach, taking advantage of both analytical methods and numerical simulations, to consider the conditions under which each model selection method works appropriately.

Acknowledgement

The author is very grateful to the referees for carefully reading the paper and for their comments and suggestions which have improved the paper.

Competing Interests

Author has declared that no competing interests exist.

References

[1] Akaike H. Information theory and an extension of the maximum likelihood principle. Proceedings of 2nd International Symposium on Information Theory (Petrov BN. Csaki F. (Eds.)), 267-281. Budapest: Akademiai Kiado; 1973.

[2] Akaike H. A new look at the statistical model identification. IEEE Transaction on Automatic Control. 1974;19(6):716-723.

[3] Konishi S, Kitagawa G. Information criteria and statistical modeling. New York: Springer; 2008.

[4] Burnham KP, Anderson DR. Model selection and multi-model inference: A practical information - theoretic approach (2nd Edition). New York: Springer; 2011.

[5] Claeskens G, Hjort NL. Model selection and model averaging. Cambridge: Cambridge University Press; 2008.

[6] Stone M. Cross-validatory choice and assessment of statistical predictions. Journal of the Royal Statistical Society, Series B. 1974;36(2):111-147.

[7] Geisser S. The predictive sample reuse method with applications. Journal of the American Statistical Association. 1975;70(350):320-328.

[8] Stone M. An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. Journal of the Royal Statistical Society. Series B (Methodological). 1977;39(1):44-47.

[9] Stoica P, Eykhoff P, Jansenn P, Söderstöm P. Model-structure selection by cross validation. International Journal of Control. 1986;43:1841-1878.

[10] Tong H. Akaike's approach can yield consistent order determination. Frontiers of statistical modeling: An information approach, (Bozdogan H. (Ed.)). Dordrecht: Kleuwer Academic Publication.1994;93-103.

[11] Syed A. A review of cross validation and adaptive model selection. Thesis, Georgia State University; 2011.

[12] Ogasawara H. Asymptotic biases of information and cross-validation criteria under canonical parametrization. Communications in Statistics - Theory and Methods; 2018. On-line published.

[13] Takezawa K. Learning regression analysis by simulation. Springer; 2014.

[14] Silverman BW. Density estimation for statistics and data analysis. London: Chapman & Hall/CRC; 1986.

[15] Horne JS, Garton EO. Likelihood cross-validation versus least squares cross-validation for choosing the smoothing parameter in Kernel home-range analysis. Journal of Wildlife Management 2006;70(3):641-648.

Peer-review history:
The peer review history for this paper can be accessed here (Please copy paste the total link in your browser address bar)
http://www.sciencedomain.org/review-history/25863