Journal of Statistical Planning and Inference 142 (2012) 301–311

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

A jackknife type approach to statistical model selection

Hyunsook Lee [a,*], G. Jogesh Babu [b], C.R. Rao [b]

[a] KISTEP, DongWon Bldg 9Fl., Mabang-gil 68, Seocho-gu, Seoul 137-130, South Korea
[b] Department of Statistics, Penn State, 326 Thomas Bldg., University Park, PA 16802, United States

ARTICLE INFO

Article history:
Received 9 September 2009
Received in revised form
25 February 2011
Accepted 9 July 2011
Available online 29 July 2011

Keywords:
Jackknife
Unbiased estimation
Kullback–Leibler divergence
Model selection
Information criterion
Maximum likelihood estimation

ABSTRACT

Procedures such as Akaike information criterion (AIC), Bayesian information criterion (BIC), minimum description length (MDL), and bootstrap information criterion have been developed in the statistical literature for model selection. Most of these methods use estimation of bias. This bias, which is inevitable in model selection problems, arises from estimating the distance between an unknown true model and an estimated model. Instead of bias estimation, a bias reduction based on jackknife type procedure is developed in this paper. The jackknife method selects a model of minimum Kullback–Leibler divergence through bias reduction. It is shown that (a) the jackknife maximum likelihood estimator is consistent, (b) the jackknife estimate of the log likelihood is asymptotically unbiased, and (c) the stochastic order of the jackknife log likelihood estimate is O(log log n). Because of these properties, the jackknife information criterion is applicable to problems of choosing a model from separated families especially when the true model is unknown. Compared to popular information criteria which are only applicable to nested models such as regression and time series settings, the jackknife information criterion is more robust in terms of filtering various types of candidate models in choosing the best approximating model.

1. Introduction

Model selection, in general, is a procedure for exploring candidate models and choosing the one among candidates that mostly minimizes the distance from the true model. One of the well known measures is the Kullback–Leibler (KL) divergence (Kullback and Leibler, 1951),

$$D_{KL}[g;f_\theta] := \int g(x)\log g(x)\,dx - \int g(x)\log f_\theta(x)\,dx = E_g[\log g(X)] - E_g[\log f_\theta(X)] \geq 0, \tag{1}$$

where $g(X)$ is the unknown true model, $f_\theta(X)$ is the estimating model with parameter θ, and X is a random variable of observations. The purpose of model selection is to find the best model, from a set of candidates $\{f_\theta : \theta \in \Theta^p \subset R^p\}$ that describes the data best. It is not necessary that this parametric family of distributions contains the true density function g but it is assumed that θ_g exists such that f_{θ_g} is closest to g in Kullback–Leibler divergence. A fitted model, $f_{\hat\theta}$, is opted to be estimated and assessing the distance between $f_{\hat\theta}$ and g automatically provides a measure to be compared with other fitted models. Since $E_g[\log g(X)]$ is unknown but constant, we only need to obtain the estimator of θ_g that maximizes $E_g[\log f_\theta(X)]$.

* Corresponding author.
E-mail addresses: hlee@kistep.re.kr (H. Lee), babu@psu.edu (G. Jogesh Babu), crr1@psu.edu (C.R. Rao).

0378-3758/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2011.07.017

Akaike (1973, 1974), first, introduced the model selection criterion (AIC) that leads to the model of the minimum KL divergence. For an unknown density g, however, AIC chooses a model that maximizes the estimate of $E_g[\log f_\theta(X)]$. Yet, the bias cannot be avoided from estimating the likelihood as well as model parameters with the same data set. The estimated bias appears as a penalty term in AIC. In fact, as θ_g is unknown, it is replaced by its maximum likelihood estimate. The resulting bias is then estimated using $E[\sum_i \log f_{\hat\theta}(X_i) - nE_g(\log f_{\hat\theta}(Z)]$, where random variable Z is independent of the data and has density g. Here the model evaluation is done from the stand point of prediction (Konishi and Kitagawa, 2008). However, the jackknife based approach in this article gets around this bias estimation problem, without explicit estimation. Regardless of their theoretical and historical background, BIC, MDL, and other modified model selection criteria that are similar to AIC are popular among practitioners.

Despite the simplicity of AIC, a few drawbacks were reported; the tendency of picking an overfitted model (Hurvich and Tsai, 1989), the lack of consistency of choosing the correct model (McQuarrie et al., 1997), and the limited application on models from the same parametric family of distributions (Konishi and Kitagawa, 1996). As remedies for these shortcomings of AIC, several information criteria with improved bias estimation were proposed. In order to adjust the inconsistency of AIC, especially for small sample size, Hurvich and Tsai (1989) and McQuarrie et al. (1997) introduced the corrected Akaike information criterion (AICc) and the unbiased Akaike information criterion (AICu), respectively. The penalty terms in their information criteria estimate the bias more consistently so that AICc and AICu select the correct model more often than AIC. On the other hand, AICc and AICu are only applied to regression models with normal error distributions. To relax the assumptions on bias estimation, other information criteria were presented, such as Takeuchi information criterion (TIC, Takeuchi, 1976), the generalized information criterion (GIC, Konishi and Kitagawa, 1996), and the information complexity (ICOMP, Bozdogan, 2000).

Apart from the bias estimation of the expected log likelihood, bias reduction through statistical resampling methods has fabricated model selection criteria without assumptions such as a parametric family of distributions. Popular resampling methods are cross-validation, jackknife, and bootstrap. Stone (1977) showed that estimating the expected log likelihood by cross-validation is asymptotically equivalent to AIC. Several attempts with the bootstrap method were made for model selection; however, the bootstrap information criterion was no better than AIC (Chung et al., 1996; Shibata, 1997; Ishiguro et al., 1997). While the asymptotic properties of the cross-validation and the bootstrap model selection, and their applications were well studied, the application of the jackknife resampling method to model selection is hardly found in the literature.

The jackknife estimator is expected to reduce the bias of the expected log likelihood from estimating the KL divergence. Assume that the bias b_n based on n observations arises from estimating the KL divergence and satisfies the expansion (Firth, 1993),

$$b_n = \frac{1}{n}a_1 + \frac{1}{n^2}a_2 + \mathrm{O}(n^{-3})$$

for some constants a_i of $\mathrm{O}(n^{-1})$. Let $b_{n,-i}$ be the estimate of bias without the ith observation. By denoting b_J as the jackknife bias estimate, the order of the jackknife bias estimate is $\mathrm{O}(n^{-2})$, which is reduced from the order of an ordinary bias $\mathrm{O}(n^{-1})$ since

$$b_J = \frac{1}{n}\sum_{i=1}^{n}(nb_n-(n-1)b_{n,-i}) = a_1 + \frac{1}{n}a_2 - \left(a_1 + \frac{1}{n-1}a_2\right) + \mathrm{O}(n^{-3}) = -\frac{1}{n(n-1)}a_2 + \mathrm{O}(n^{-3}).$$

Furthermore, this bias reduction allows to alleviate model assumptions toward estimating bias in the same parametric family of distributions or the nested candidate models that typically are listed in those popular model selection criteria and therefore, without these assumptions for these candidate model the statistics for model selection utilizing the jackknife method becomes robust, whose detail account will follow in the later section.

Very few studies on the jackknife principle were found in model selection compared to the bootstrap and the cross-validation methods, possibly caused by the perception that jackknife is similar to either bootstrap or cross-validation asymptotically as similarities among those resampling methods were reported in diverse statistical problems. In this study, we developed the jackknife information criterion (JIC) for model selection based on the KL divergence measure as a competing model selection criterion. Section 2 describes necessary assumptions and discusses the strong convergence of the maximum likelihood estimator when the true model is unspecified. Section 3 presents the uniformly strong convergence of a jackknife maximum likelihood estimator toward the new model selection criterion with the jackknife resampling scheme. Section 4 provides the definition of JIC and its asymptotic properties under the regularity conditions. Section 5 provides comparisons between JIC and popular information criteria when the true model is unknown. Lastly, Section 6 discusses the consistency of JIC in comparison with other model selection methods.

2. Preliminaries

Suppose that the data points, $X_1, X_2, \ldots, X_n$ are *iid* with a common density g and $\mathcal{M} = \{f_\theta : \theta \in \Theta^p\}$ is a class of candidate models. Let the log likelihood, $\log f_\theta(X_i)$ of X_i be denoted by $l_i(\theta)$. The log likelihood function of all the observations and the

log likelihood function without the ith observation are denoted by $L(\theta)=\sum_{i=1}^{n} l_i(\theta)$ and $L_{-i}(\theta)=\sum_{j\neq i} l_j(\theta)$, respectively. Throughout this paper, we assume the following five conditions:

(J1) The p dimensional parameter space Θ^p is a compact subset of $\mathbb{R}^p$.
(J2) For any $\theta\in\Theta^p$, each $f_\theta\in\mathcal{M}$ is distinct, i.e. if $\theta_1\neq\theta_2$, then $f_{\theta_1}\neq f_{\theta_2}$.
(J3) A unique parameter θ_g exists in the interior of Θ^p and satisfies

$$\theta_g=\arg\max_{\theta\in\Theta^p} E_g[l_i(\theta)] \quad \text{and} \quad E_g[\nabla_\theta l_i(\theta_g)]=0.$$

(J4) The log likelihood $l_i(\theta)$ is continuous in $\theta\in\Theta^p$ and thrice continuously differentiable with respect to θ in the interior of Θ^p, of which derivatives are denoted by $\nabla_\theta l_i(\theta)$, $\nabla_\theta^2 l_i(\theta)$, and $\nabla_\theta^3 l_i(\theta)$, in a given order. Also, $|l_i(\theta)|$, $|(\partial/\partial\theta_k)l_i(\theta)|$, $|(\partial^2/\partial\theta_k\partial\theta_l)l_i(\theta)|$, and $|(\partial^3/\partial\theta_k\partial\theta_l\partial\theta_m)l_i(\theta)|$ are dominated by $h(X_i)$, which is non-negative, does not depend on θ, and satisfies $0<E[h(X_i)]<\infty$ $(k,l,m=1,\ldots,p)$.
(J5) For any $\theta\in\Theta^p$, $E_g[\nabla_\theta l_i(\theta)\nabla_\theta^T l_i(\theta)]$ and $-E_g[\nabla_\theta^2 l_i(\theta)]$ are finite and positive definite $p\times p$ matrices. If $g=f_{\theta_g}$, then $E_g[\nabla_\theta l_i(\theta)\nabla_\theta^T l_i(\theta)]=-E_g[\nabla_\theta^2 l_i(\theta)]$. In particular, for $\theta=\theta_g$, $\Lambda=-E_g[\nabla_\theta^2 l_i(\theta_g)]$, where Λ is a non-singular matrix.

Remark on (J3). If $\mathcal{M}$ contains g, not only θ_g maximizes $E_g[\log f_\theta(X_i)]$ but also $D_{KL}[g;f_{\theta_g}]$ in (1) becomes zero. However, g is likely to be unknown so that the estimate of the density f_{θ_g} that minimizes the KL divergence is sought as a surrogate model of g. Also, note that the parameter θ_g satisfies $E_g[\nabla_\theta l_i(\theta_g)]=0$ and $E_g[\log f_\theta(X_i)]<E_g[\log f_{\theta_g}(X_i)]$ for any $\theta\in\Theta^p$ and $\theta\neq\theta_g$. Here, the strict inequality holds due to the uniqueness of θ_g by (J3).

Hitherto, we show that the maximum likelihood estimator $\hat\theta_n$ of θ_g converges almost surely to the parameter of interest θ_g that minimizes the KL divergence.

Theorem 1. *If $\hat\theta_n$ is a function of $X_1,\ldots,X_n$ such that $\nabla_\theta L(\hat\theta_n)=0$, then*

$$\hat\theta_n\xrightarrow{a.s.}\theta_g.$$

The maximum likelihood estimator $\hat\theta_n$, in another words, the minimum discrepancy estimator of θ_g is credited to minimize the KL divergence.

Proof. It is sufficient to prove that $\mathrm{B}: P\{\lim_{n\to\infty}\|\hat\theta_n-\theta_g\|<\epsilon\}=1$ for $\forall\epsilon>0$, where $\hat\theta_n$ satisfying

$$\mathrm{A}: \prod_{i=1}^{n} f_{\hat\theta_n}(X_i)\geq \prod_{i=1}^{n} f_\theta(X_i), \quad \forall\theta\in\Theta^p,\ \forall n. \tag{2}$$

We will prove this almost sure convergence by negating $\mathrm{A}\Rightarrow\mathrm{B}$. Assume that B does not hold. There exists a subsequence $\{\hat\theta_{n_k}\}$ of $\hat\theta_n$ such that $\hat\theta_{n_k}\to\overline{\theta}$ a.e. for $\overline{\theta}\neq\theta_g$ and $\overline{\theta}\in\Theta^p$. Consider $\epsilon>0$ such that $\|\overline{\theta}-\theta_g\|\geq\epsilon$. In addition, from the assumptions (J1–J3), for some $\delta>0$,

$$E[l_1(\overline{\theta})-l_1(\theta_g)]=-\delta<0. \tag{3}$$

By the strong law of large numbers (SLLN), and (3),

$$\frac{1}{n}\sum_{i=1}^{n}[l_i(\overline{\theta})-l_i(\theta_g)]\xrightarrow{a.s.}-\delta.$$

Thus, for all large n,

$$\frac{1}{n}\sum_{i=1}^{n}[l_i(\overline{\theta})-l_i(\theta_g)]<-\delta/2 \quad \text{a.s.} \tag{4}$$

As $\hat\theta_{n_k}\xrightarrow{a.s.}\overline{\theta}$, for all large k,

$$\|\hat\theta_{n_k}-\overline{\theta}\|<\frac{\delta}{8E[h(X_1)]} \quad \text{a.s.},$$

where $E[h(X_1)]$ is bounded as described in (J4) and therefore, by the SLLN,

$$\left|\frac{1}{n_k}\sum_{i=1}^{n_k}[l_i(\hat\theta_{n_k})-l_i(\overline{\theta})]\right|\leq\|\hat\theta_{n_k}-\overline{\theta}\|\frac{1}{n_k}\sum_{i=1}^{n_k}h(X_i)<\frac{\delta}{8E[h(X_1)]}\frac{1}{n_k}\sum_{i=1}^{n_k}h(X_i)<\frac{\delta}{4} \quad \text{a.s.} \tag{5}$$

for all large k. Thus (4) and (5) lead to

$$\frac{1}{n}\sum_{i=1}^{n}[l_i(\hat{\theta}_n)-l_i(\theta_g)] < -\delta/4 \quad i.o.\ \text{a.s.}$$

and therefore,

$$P\left(\sum_{i=1}^{n}[l_i(\hat{\theta}_n)-l_i(\theta_g)] < 0 \ i.o.\right) = 1.$$

Equivalently,

$$P\left(\prod_{i=1}^{n} f_{\hat{\theta}_n}(X_i) < \prod_{i=1}^{n} f_{\theta_g}(X_i) \ i.o.\right) = 1,$$

which violates (2) for $\theta = \theta_g$. This completes the proof. □

Under the assumptions (J1–J3), the maximum likelihood estimator $\hat{\theta}_n$ converges almost surely to θ_g even if the true model is unspecified. This strong consistency is generally assumed in statistical model selection studies with an unspecified/misspecified true model (e.g. Nishii, 1988; Sin and White, 1996). For a particular case $g = f_{\theta_g}$, this strong convergence is easily proved with Wald's (1949) approach. When θ_g is the true model parameter, Stone (1977) proved that the cross-validation is equivalent to AIC by assuming that the maximum likelihood estimate of $\hat{\theta}_n$ converges to θ_g in probability. Information criterion based methods such as AIC evaluate a fitted model to the true model in terms of minimizing the KL divergence through the maximum likelihood estimation.

Another popular model selection criterion BIC assumes equal priors for candidate models as well as an exponential distribution family. Naturally, general scientific models beyond these assumptions have to be excluded for model selection. Nonetheless, these assumptions are not considered carefully upon collecting candidate models. Without having a clue to the true model, the popular model selection criteria are practiced on these collected candidate models blindly irrespective of assumptions. One of the key objectives in model selection is assessing the unknown true model reasonably and objectively without introducing bias before declaring one.

Although the maximum likelihood estimator $\hat{\theta}_n$ is strongly consistent with the parameter of interest θ_g (the parameter that minimizes the KL divergence), there is no guarantee that this $\hat{\theta}_n$ produces an unbiased likelihood estimate. Similar problems arise in other well known model selection criteria. In order to rectify this, unlike the other information criteria that take the bias estimation approach in the likelihood estimation, we will take the bias reduction approach based on the jackknife resampling method. The jackknife bias reduction is not a new concept in estimating models as given in Firth (1993) for generalized linear models, but it has not appeared rigorously from the best model selection perspective.

3. Jackknife maximum likelihood estimator

Jackknife is well known for bias reduction and relatively inexpensive cost in computation compared to bootstrap. The strong consistency of the maximum likelihood estimator has been discussed extensively in the literature (e.g. Wald, 1949; LeCam, 1953; Huber, 1967). To begin developing a model selection criterion from the jackknife approach, first we will show the uniform strong convergence of the jackknife maximum likelihood 'pseudo-estimators' $\hat{\theta}_{-i}$ to θ_g.

Theorem 2. *Let $\hat{\theta}_{-i}$ be a function of $X_1,\ldots,X_{i-1},X_{i+1},\ldots,X_n$ and satisfy $\nabla_\theta L_{-i}(\hat{\theta}_{-i}) = 0$ for any $i \in [1,n]$. Then, uniformly in i, $\hat{\theta}_{-i}$ converges to θ_g almost surely. That is,*

$$P\left(\lim_{n\to\infty}\max_{1\le i\le n}\|\hat{\theta}_{-i}-\theta_g\| \le \epsilon\right) = 1, \quad \forall \epsilon > 0.$$

Proof. Let $\tau_n = \max_{1\le i\le n}\|\hat{\theta}_{-i}-\theta_g\|$, where $\hat{\theta}_{-i}$ satisfies

$$\text{A}: \prod_{j=1, j\neq i}^{n} f_{\hat{\theta}_{-i}}(X_j) \ge \prod_{j=1, j\neq i}^{n} f_{\theta}(X_j), \quad \forall \theta \in \Theta^p, \ \forall i \in [1,n]. \tag{6}$$

It is sufficient to show that τ_n converges to zero almost surely (B). The proof is similar to that of Theorem 1 via negating $\text{A} \Rightarrow \text{B}$ to $\neg\text{B} \Rightarrow \neg\text{A}$.

Consider $\tau_n \nrightarrow 0$, a.e. and $\|\hat{\theta}_{-i(n)}-\theta_g\| \nrightarrow 0$, a.e. for all $i = i(n) : 1 \le i \le n$. Let $\overline{\theta}$ be a limit point of $\{\hat{\theta}_{-i(n)}\}$. For some $\delta > 0$ as in (3) and $E[h(X_1)] < \infty$ as in (J4),

$$\|\hat{\theta}_{-i(n)}-\overline{\theta}\| < \frac{\delta}{8E[h(X_1)]} \quad i.o.$$

Then,

$$\left|\frac{1}{n-1}\sum_{j=1,j\neq i(n)}^{n}[l_j(\hat{\theta}_{-i(n)})-l_j(\overline{\theta})]\right| \leq \|\hat{\theta}_{-i(n)}-\overline{\theta}\|\frac{1}{n-1}\sum_{j=1,j\neq i(n)}^{n}h(X_j) < \frac{\delta}{8E[h(X_1)]}\frac{n}{n-1}\frac{1}{n}\sum_{j=1}^{n}h(X_j) < \frac{\delta}{4} \quad i.o.\ \text{a.s.}$$

Hence,

$$P\left(\frac{1}{n-1}\sum_{j=1,j\neq i(n)}^{n}[l_j(\hat{\theta}_{-i(n)})-l_j(\theta_g)] < 0 \ i.o\right) = 1$$

and

$$P\left(\prod_{j=1,j\neq i(n)}^{n} f_{\hat{\theta}_{-i(n)}}(X_j) < \prod_{j=1,j\neq i(n)}^{n} f_{\theta}(X_j) \ i.o.\right) = 1. \tag{7}$$

This result (7) contradicts (6) for $\theta=\theta_g$. This completes the proof. □

We show that the jackknife maximum likelihood estimator converges to θ_g, the parameter that minimizes the KL divergence. Based on the regularity conditions and the strong convergence of the maximum likelihood estimators discussed so far, we shall take the jackknife approach to a model selection criterion in the next section.

4. Jackknife information criterion

In this section, we define the jackknife information criterion (JIC) and investigate the asymptotic characteristics of the jackknife model selection principle. In contrast to the traditional theoretical formulas, the jackknife does not explicitly require assumptions on candidate models (Shao and Tu, 1995). We only assumed regularity conditions to ensure the strong consistency of maximum likelihood estimators and to regulate the asymptotic behavior of the jackknife model selection.

Deriving an information criterion starts from estimating the expected log likelihood. Prior to discussing JIC, we define the jackknife estimator of the log likelihood function. In addition, we show that this jackknife estimator is asymptotically unbiased to the expected log likelihood function.

Once the bias corrected jackknife estimator $\overline{\tau}$ is defined as

$$\overline{\tau}=\hat{\tau}-b_J,$$

b_J is a jackknife bias estimator and $\hat{\tau}$ is a target estimator. The jackknife bias estimator is then obtained by

$$b_J=(n-1)\frac{1}{n}\sum_{i=1}^{n}(\hat{\tau}_{-i}-\hat{\tau})$$

with $\hat{\tau}_{-i}$, the same estimator as $\hat{\tau}$ without the ith observation. Similarly, we formulate the jackknife estimator of the log likelihood.

Definition 1. Let the jackknife estimator of the log likelihood J_n be

$$J_n = nL(\hat{\theta}_n)-\sum_{i=1}^{n}L_{-i}(\hat{\theta}_{-i}). \tag{8}$$

Hitherto, the strong convergence of $\hat{\theta}_n$ and $\hat{\theta}_{-i}$ leads to the following theorem.

Theorem 3. *Let $X_1,X_2,\ldots,X_n$ be iid random variables from a density function g of an unknown distribution and $\mathcal{M}=\{f_\theta:\theta\in\Theta^p\}$ be a class of models. If the $(g,\mathcal{M})$ meets (J1)–(J5), then the jackknife log likelihood estimator J_n satisfies*

$$J_n = nL(\hat{\theta}_n)-\sum_{i=1}^{n}L_{-i}(\hat{\theta}_{-i}) = L(\theta_g)+\frac{1}{n}\sum_{i\neq j}\nabla l_i(\theta_g)^T\Lambda^{-1}\nabla l_j(\theta_g)+\epsilon_n, \tag{9}$$

where $\Lambda=-E_g[\nabla_\theta^2 l_i(\theta_g)]$ and ϵ_n is a random variable satisfying $\lim_{n\to\infty}E_g[\epsilon_n]=0$ and $\|\epsilon_n\|\xrightarrow{a.s}0$. Additionally, the fact that $E_g[(1/n)\sum_{i\neq j}\nabla l_i(\theta_g)^T\Lambda^{-1}\nabla l_j(\theta_g)]=0$ leads to J_n as an asymptotically unbiased estimator of the log likelihood function.

Proof. We begin with forming groups of components in J_n.

$$J_n=\sum_{i=1}^{n}[L(\hat{\theta}_n)-L_{-i}(\hat{\theta}_{-i})]=\sum_{i=1}^{n}[L_{-i}(\hat{\theta}_n)+l_i(\hat{\theta}_n)-L_{-i}(\hat{\theta}_{-i})-l_i(\theta_g)+l_i(\theta_g)]=\sum_{i=1}^{n}[L_{-i}(\hat{\theta}_n)-L_{-i}(\hat{\theta}_{-i})]+\sum_{i=1}^{n}[l_i(\hat{\theta}_n)-l_i(\theta_g)]+\sum_{i=1}^{n}l_i(\theta_g).$$

By the Taylor expansion, the first and the second term respectively become

$$\sum_{i=1}^{n}[L_{-i}(\hat{\theta}_n)-L_{-i}(\hat{\theta}_{-i})]=\sum_{i=1}^{n}\nabla L_{-i}(\hat{\theta}_{-i})(\hat{\theta}_n-\hat{\theta}_{-i})+\frac{1}{2}\sum_{i=1}^{n}(\hat{\theta}_n-\hat{\theta}_{-i})^T\nabla^2 L_{-i}(\eta_i)(\hat{\theta}_n-\hat{\theta}_{-i})=\frac{1}{2}\sum_{i=1}^{n}(\hat{\theta}_n-\hat{\theta}_{-i})^T\nabla^2 L_{-i}(\eta_i)(\hat{\theta}_n-\hat{\theta}_{-i})$$

and

$$\sum_{i=1}^{n}[l_i(\hat{\theta}_n)-l_i(\theta_g)]=L(\hat{\theta}_n)-L(\theta_g)=-(\theta_g-\hat{\theta}_n)\nabla L(\hat{\theta}_n)-\frac{1}{2}(\theta_g-\hat{\theta}_n)^T\nabla^2 L(\xi)(\theta_g-\hat{\theta}_n)=-\frac{1}{2}(\theta_g-\hat{\theta}_n)^T\nabla^2 L(\xi)(\theta_g-\hat{\theta}_n),$$

where $\xi=\theta_g+\delta(\hat{\theta}_n-\theta_g)$ and $\eta_i=\hat{\theta}_n+\gamma_i(\hat{\theta}_{-i}-\hat{\theta}_n)$ for some δ and γ_i such that $0<\delta,\gamma_i<1$. The last term $\sum_{i=1}^{n} l_i(\theta_g)$ simplifies to $L(\theta_g)$. Hence,

$$J_n=L(\theta_g)+\frac{1}{2}\sum_{i=1}^{n}(\hat{\theta}_n-\hat{\theta}_{-i})^T\nabla^2 L_{-i}(\eta_i)(\hat{\theta}_n-\hat{\theta}_{-i})-\frac{1}{2}(\theta_g-\hat{\theta}_n)^T\nabla^2 L(\xi)(\theta_g-\hat{\theta}_n). \tag{10}$$

First, consider $\sum(\hat{\theta}_n-\hat{\theta}_{-i})^T\nabla^2 L_{-i}(\eta_i)(\hat{\theta}_n-\hat{\theta}_{-i})$ of Eq. (10). From Theorems 1 and 2, the maximum likelihood estimates $\hat{\theta}_n$ and $\hat{\theta}_{-i}$ are confined such that for any $\epsilon>0$, $\max\{\|\hat{\theta}_n-\theta_g\|,\max_{1\le i\le n}\|\hat{\theta}_{-i}-\theta_g\|\}\le\epsilon$ for all large n. Also, note that $\max_{1\le i\le n}\|\eta_i-\theta_g\|\le\max(\|\hat{\theta}_n-\theta_g\|,\max_{1\le i\le n}\|\hat{\theta}_{-i}-\theta_g\|)$. Thus, the following is established:

$$\max_{1\le i\le n}\left\|\frac{1}{n}\nabla^2 L_{-i}(\eta_i)+\Lambda\right\|\le\max_{1\le i\le n}\frac{1}{n}\|\nabla^2 L_{-i}(\eta_i)-\nabla^2 L(\theta_g)\|+\left\|\frac{1}{n}\nabla^2 L(\theta_g)+\Lambda\right\|. \tag{11}$$

Here, $\max_{1\le i\le n}(1/n)\|\nabla^2 L_{-i}(\eta_i)-\nabla^2 L(\theta_g)\|\xrightarrow{a.s.}0$ and $\|(1/n)L(\theta_g)+\Lambda\|\xrightarrow{a.s.}0$ by (J4) and the SLLN. Thus,

$$\max_{1\le i\le n}\left\|\frac{1}{n}\nabla^2 L_{-i}(\eta_i)+\Lambda\right\|\xrightarrow{a.s.}0. \tag{12}$$

In addition, as $\nabla L(\hat{\theta}_n)=\nabla L_{-i}(\hat{\theta}_{-i})=0$,

$$-\nabla l_i(\hat{\theta}_n)=\nabla L(\hat{\theta}_n)-\nabla L_{-i}(\hat{\theta}_{-i})-\nabla l_i(\hat{\theta}_n)=\nabla L_{-i}(\hat{\theta}_n)-\nabla L_{-i}(\hat{\theta}_{-i})=\nabla^2 L_{-i}(\eta_i)(\hat{\theta}_n-\hat{\theta}_{-i}).$$

By (12) and (J4), the above equation is

$$(\hat{\theta}_n-\hat{\theta}_{-i})=\frac{1}{n}\Lambda^{-1}\nabla l_i(\hat{\theta}_n)+n^{-1}\epsilon_n, \tag{13}$$

uniformly in i. Here and in what follows $\{\epsilon_n\}$ denotes a generic sequence of random variables satisfying $\lim_{n\to\infty}E[\epsilon_n]=0$ and $\epsilon_n\xrightarrow{a.s.}0$, and may represent different values in different equations. Since $|\nabla^2 l_i(\eta)|$ is bounded and $\|\hat{\theta}_n-\theta_g\|=\epsilon_n$,

$$\nabla l_i(\hat{\theta}_n)=\nabla l_i(\theta_g)+\nabla^2 l_i(\eta)(\hat{\theta}_n-\theta_g)$$

is equivalent to

$$\nabla l_i(\hat{\theta}_n)=\nabla l_i(\theta_g)+\epsilon_n. \tag{14}$$

Therefore, combining Eqs. (12)–(14) gives

$$\sum_{i=1}^{n}(\hat{\theta}_n-\hat{\theta}_{-i})^T\nabla^2 L_{-i}(\eta_i)(\hat{\theta}_n-\hat{\theta}_{-i})=-\frac{1}{n}\sum_{i=1}^{n}\nabla^T l_i(\theta_g)\Lambda^{-1}\nabla l_i(\theta_g)+\epsilon_n.$$

Similarly, the last term of Eq. (10) becomes

$$(\theta_g-\hat{\theta}_n)^T\nabla^2 L(\xi)(\theta_g-\hat{\theta}_n)=-\frac{1}{n}\nabla^T L(\theta_g)\Lambda^{-1}\nabla L(\theta_g)+\epsilon_n.$$

Since $\|\xi-\theta_g\|\le\|\hat{\theta}_n-\theta_g\|$,

$$\frac{1}{n}\|\nabla^2 L(\xi)-\nabla^2 L(\theta_g)\|=\epsilon_n,$$

and by the SLLN,

$$\nabla^2 L(\xi)=\nabla^2 L(\theta_g)+n\Lambda-n\Lambda-\nabla^2 L(\theta_g)+\nabla^2 L(\xi)=n\left(\frac{1}{n}\nabla^2 L(\theta_g)+\Lambda\right)+n\left(\frac{\nabla^2 L(\xi)-\nabla_\theta^2 L(\theta_g)}{n}\right)-n\Lambda=-n(\Lambda+\epsilon_n). \tag{15}$$

Additionally, we have

$$\nabla L(\theta_g)=\nabla L(\theta_g)-\nabla L(\hat{\theta}_n)=\nabla^2 L(\xi)(\theta_g-\hat{\theta}_n) \tag{16}$$

and replacing Eq. (15) in Eq. (16) gives

$$(\theta_g-\hat{\theta}_n)=-\frac{1}{n}\Lambda^{-1}\nabla L(\theta_g)+n^{-1}\epsilon_n.$$

Hitherto,

$$(\theta_g-\hat{\theta}_n)^T\nabla^2 L(\xi)(\theta_g-\hat{\theta}_n)=-\frac{1}{n}\nabla^T L(\theta_g)\Lambda^{-1}\nabla L(\theta_g)+n^{-1}\epsilon_n.$$

Consequently, Eq. (10) becomes

$$\begin{aligned}J_n &= L(\theta_g)-\frac{1}{2n}\sum_{i=1}^{n}\nabla^T l_i(\theta_g)\Lambda^{-1}\nabla l_i(\theta_g)+\frac{1}{2n}\nabla^T L(\theta_g)\Lambda^{-1}\nabla L(\theta_g)+\epsilon_n\\ &= L(\theta_g)-\frac{1}{2n}\sum_{i=1}^{n}\nabla^T l_i(\theta_g)\Lambda^{-1}\nabla l_i(\theta_g)+\frac{1}{2n}\sum_{i=1}^{n}\nabla^T l_i(\theta_g)\Lambda^{-1}\nabla l_i(\theta_g)+\frac{1}{2n}\sum_{i\neq j}\nabla^T l_i(\theta_g)\Lambda^{-1}\nabla l_j(\theta_g)+\epsilon_n\\ &= L(\theta_g)+\frac{1}{2n}\sum_{i\neq j}\nabla^T l_i(\theta_g)\Lambda^{-1}\nabla l_j(\theta_g)+\epsilon_n. \qquad (17)\end{aligned}$$

Note that J_n is an asymptotically unbiased estimator of the expected log likelihood;

$$E\left[\sum_{i\neq j}\nabla l_i(\theta_g)^T\Lambda^{-1}\nabla l_j(\theta_g)\right]=E\left[E\left[\sum_{i\neq j}\nabla l_i(\theta_g)^T\Lambda^{-1}\nabla l_j(\theta_g)|X_i\right]\right]=\sum_{i\neq j}E[\nabla l_i(\theta_g)^T\Lambda^{-1}]E[\nabla l_j(\theta_g)]=0.$$

This completes the proof. □

Asymptotically, no bias term is involved in J_n in contrast to the cross-validation and the bootstrap estimator of the log likelihood (Shao, 1993; Chung et al., 1996; Shibata, 1997). Therefore, the model selection criterion obtained from the jackknife method is expected to perform differently in asymptotics compared to the selection criteria of other resampling methods. Here we investigate the jackknife principle further as a model selection criterion by proposing the JIC as minus twice J_n since the popular information criteria involve estimators of negative twice expected log likelihood.

Definition 2. The jackknife information criterion, JIC is defined as minus twice J_n.

$$JIC=-2J_n=-2nL(\hat{\theta})+2\sum_{i=1}^{n}L_{-i}(\hat{\theta}_{-i}).$$

Note that multiplying -2 does not change the asymptotically unbiased property of J_n. Besides, the actual behavior of a maximum likelihood type estimator is predicted through the convergence rate, obtained from the Taylor expansion and the central limit theorem (CLT). Since the regularity conditions guarantee the Taylor expansion, investigating the limiting distributions expanded from J_n specifies the convergence rate. Besides, Miller (1974) commented that no one succeeded in assessing the limiting distribution of a delete-one jackknife estimator. While the CLT requires the consistent variance estimator, the delete-one jackknife is known for inconsistent variance estimators for non-smooth estimators (Shao and Wu, 1989). Theorem 3 only assures the existence of the asymptotically unbiased delete-one jackknife estimator of the expected log likelihood. Instead of getting the limiting distribution of J_n by the CLT, the stochastic rate of J_n is presented by means of the law of iterated logarithm (LIL) in order to understand the asymptotic behavior of J_n.

Theorem 4. *Let $X_1, X_2, \ldots, X_n$ be iid random variables from the unknown distribution with density g and $\mathcal{M}$ be a class of models such that $\mathcal{M}=\{f_\theta : \theta\in\Theta^p\}$. Under* (J1)–(J5), *the stochastic orders relating to J_n are*:

(1) $J_n = L(\theta_g)+\mathrm{O}(\log\log n)$ *a.s.*
(2) $(1/n)J_n = \int \log f_{\theta_g}(x)g(x)\,dx+\mathrm{O}(\sqrt{n^{-1}\log\log n})$ *a.s.*

Proof. By the LIL and (J5), with $Var[\nabla l_i(\theta_g)]<\infty$, we have

$$\nabla L(\theta_g)=\mathrm{O}(\sqrt{n\log\log n})\quad \text{a.s.}$$

so that the stochastic order of J_n is given by

$$\begin{aligned}J_n &= L(\theta_g)+\frac{1}{2n}\sum_{i\neq j}\nabla^T l_i(\theta_g)\Lambda^{-1}\nabla l_j(\theta_g)+\epsilon_n = L(\theta_g)+\frac{1}{2n}\nabla^T L(\theta_g)\Lambda^{-1}\nabla L(\theta_g)+\frac{1}{2n}\sum_{i=1}^{n}\nabla^T l_i(\theta_g)\Lambda^{-1}\nabla l_i(\theta_g)+\epsilon_n\\ &= L(\theta_g)+\mathrm{O}(\log\log n)\ \text{a.s.}, \qquad (18)\end{aligned}$$

where ϵ_n is as in the proof of Theorem 3. The result from (18) and the LIL leads to

$$\frac{1}{n}J_n = \frac{1}{n}L(\theta_g) + \frac{1}{n}\{J_n - L(\theta_g)\} = \int \log f_{\theta_g}(x) g(x)\, dx + O(\sqrt{n^{-1}\log\log n}) \quad \text{a.s.} \tag{19}$$

since $(1/n)L(\theta_g) = \int \log f_{\theta_g}(x) g(x)\, dx + O(\sqrt{n^{-1}\log\log n})$ (Nishii, 1988). This completes the proof. □

Although Theorem 3 shows that J_n is asymptotically unbiased, JIC might not pick a good model, particularly, among the nested models of which log likelihoods $L(\theta_g)$ are asymptotically identical. The stochastic rates of the maximum likelihood type estimators under model misspecification investigated by Nishii (1988) imply that the randomness of the estimators is bounded by a function of sample size, not by the number of parameters and model complexity. This randomness can be diluted by adopting some penalty terms appeared in other information criteria, which may enhance the consistency of jackknife model selection. We propose a slight modification of JIC.

Proposition 1. *The corrected JIC (JICc) selects a model consistently, which is defined as*

$$JICc = JIC + C_n$$

where C_n satisfying $C_n/n \to 0$ and $C_n/\log\log n \to \infty$.

The penalty term C_n from Proposition 1 was adapted from the data oriented penalty by Rao and Wu (1989) and Bai et al. (1999). Penalty terms only can be estimated when the true model is known. In the current study, however, we consider cases when the true model is unspecified or misspecified. We expect that employing bias reduction methods improves the result when the penalty is not estimable. No model complexity has been reflected in JIC, although such complexity further penalizes the model with many parameters. Therefore, we only can bound the order of the penalty term C_n that should be larger than the stochastic rate of J_n and that requires to be smaller than the sample size. Like MDL, it seems reasonable to adopt the degree of model complexity into C_n. If a part of candidate models is nested and those partially nested models have the complexity proportional to the number of parameters, then $p\log n$ would be a good choice for the penalty term since $p\log n$ is higher than $\log\log n$ and less than n in order. Depending on the choice of candidate models and the structure of their parameter spaces, in addition to bias corrected maximum log likelihoods J_n, the penalty terms adopted and modified from popular model selection criterion, could improve the decision making process for more versatile candidate models without sacrificing robustness.

The following are mere suggestions for the modification of JIC, when some candidate models share their parameter spaces, or are nested. First, the penalty term in AIC is added to JIC. We denote this information criterion as aJIC,

$$aJIC = JIC + 2p.$$

Note that the penalty term $2p$ does not satisfy the data oriented penalty term in Proposition 1. Second, bJIC is defined by adding the penalty term of BIC,

$$bJIC = JIC + p\log n.$$

Because of the popularity of AIC and BIC, these two penalties are chosen without rigorous derivation. With a complex collection of candidate models, assessing the model complexity p properly could be a challenge since the penalty terms must satisfy the conditions in Proposition 1. We expect that bJIC is a more consistent model selection criterion than aJIC or JIC because if p is plainly taken via counting the number of parameters, then $p\log n$ automatically satisfies those constraints $C_n/n \to 0$ and $C_n/\log\log n \to \infty$. Defining C_n is quite circumstantial and we omit this case specific discussion here, except a general discussion of choosing a model from a small set of non-nested/nested candidate models.

Remark on (J2). In the context of estimating model complexity, the identifiability constraint in parameters is admittedly very restrictive. However, this strong condition eliminates long disputes on the problems caused by parameters near the boundary. When candidate models are nested, for example, in mixtures and regression models, it is hard to distinguish whether the additional component is close to zero or the parameters of the additional component are close to those of the smaller models. Due to this ambiguity in the parameter space, traditional model selection cannot be compared directly to the jackknife criterion. The discussion on these break downs of parameters near the boundary are found in Bozdogan (2000). A practical account for this boundary issue from astronomy is given in Protassov et al. (2002).

5. When candidate models are nested

In this section we discuss circumstances where JIC is preferable. Consider k candidates models, $M_1, \ldots, M_k$ where M_1 is the smallest model and M_k is the largest. These models are nested in a given order. Let the corresponding parameter space of each model be $\Omega_1, \ldots, \Omega_k$. It is obvious that $\Omega_i \subset \Omega_j$ for $i < j$. It is probable that θ_g, the parameter of interest satisfying $E[\nabla L(\theta_g)] = 0$, lies in Ω_m for a given m but does not live in the space Ω_i where $i < m$. The virtue of consistent model selection methods is praised when the method correctly identify m, neither $m-1$ nor $m+1$. Traditional model selection criteria have penalty terms to control neither underestimated nor overestimated model to be chosen, whereas the jackknife type model selection criterion has no penalty terms.

Unless $\theta_g \in \Omega_k \backslash \Omega_{k-1}$, the jackknife information criterion is unable to discriminate the best model among nested models. The event $\theta_g \in \Omega_m$ implies $\theta_g \in \Omega_i$ for $m \le i \le k$ so that the jackknife maximum likelihood estimates of $M_m, \ldots, M_k$ are stochastically equivalent ($J_n(M_i) - J_n(M_j) = o(1)$ for $m \le i,j \le k$ and $i \ne j$). On the other hand, if θ_g only resides in the largest parameter space, the maximum likelihood approach estimates the parameter of interest θ_i of each model M_i that minimizes the KL divergence in its own parameter space. Yet the resulting KL divergence by θ_i is larger than the KL divergence by θ_g. When the true model is unspecified, bias due to model estimation cannot be assessed and the penalty terms in popular model selection criteria such as $2p$ in AIC and $p \log n$ in BIC would take different forms reflecting the complexity of candidate models (Spiegelhalter et al., 2002). Thus, the bias reduction approach of estimating and minimizing the KL divergence allows to measure the relative distances among candidate models objectively. When the parameter space of each candidate model occupies a unique space, regardless of the unknown true model, the JIC will select a model with the smallest KL divergence. Only difficulty arises when some candidate models are nested and the true model is unknown. Fig. 1 provides the schematic description of relationships among candidate models and their parameter spaces, where $g(x)$, in particular, denotes the unknown true model that generates data points.

We expect that there exist parameters θ_g, θ_a, and θ_b of non-nested candidate models that minimize the KL divergences from the unknown true model g and these parameters reside in Θ_2, Θ_A, and Θ_B, respectively as illustrated in Fig. 1. Let θ_g be the parameter of the target model $f_{\theta \in \Theta_2}$ that minimizes the KL divergence mostly among other candidate models such as $f_{\theta \in \Theta_1}$, $a_{\theta_A \in \Theta_A}$, $b_{\theta_B \in \Theta_B}$. Then, the jackknife log likelihood estimate J_n is the asymptotically unbiased log likelihood (Theorem 3) and the (jackknife) maximum likelihood estimator of θ_g is the consistent estimator of θ_g (Theorems 1 and 2). This target model, of which parameter space is Θ_2, is the best approximating model that describes the true model in terms of the KL divergence compared to the other candidate models of Θ_A and Θ_B since the jackknife maximum likelihood estimate J_{n,θ_g} shows the smallest KL divergence compared to other distances: $D_{KL}(g; f_{\theta_g}) < D_{KL}(g; b_{\theta_B}) < D_{KL}(g; a_{\theta_A})$. By the theorems, among non-nested models, JIC correctly identifies the model of minimum KL divergence. Only the trouble comes in when we have nested candidate models.

Suppose $M_1 = \{f_\theta : \theta \in \Theta_1\}$ and $M_2 = \{f_\theta : \theta \in \Theta_2\}$ are the candidate models whose parameter space are Θ_1 and Θ_2. Consider Θ_1 is the smaller parameter space of the nested candidate model M_1 such that $\Theta_1 \subset \Theta_2$. Obviously, M_1 is nested to M_2. In Fig. 1, this subspace Θ_1 is indicated by the shaded set that subsides in Θ_2. If θ_g, the parameter that minimizes the KL divergence mostly, resides on Θ_1, the jackknife log likelihoods of both models (M_1 and M_2) are asymptotically equal and the jackknife method does not rectify any additional penalty terms to choose the parsimonious model M_1 because the jackknife only corrects bias, not estimate bias like AIC or GIC. On the contrary, as illustrated in Fig. 1, when θ_g does not live in Θ_1 and θ_1 is the parameter of interest that minimizes the KL divergence with the given model M_1 within Θ_1, only M_2 is the best approximating model to g that the jackknife maximum likelihood approach can identify. If we fail to include the candidate model M_2 then θ_B and b_{θ_B} become the best approximating model ($D_{KL}(g; b_{\theta_B}) < D_{KL}(g; f_{\theta_1}) < D_{KL}(g; a_{\theta_A})$). Considering only the parsimonious model M_1 and other candidate models like $M_A = \{a_\theta : \theta \in \Theta_A\}$ and $M_B = \{b_\theta : \theta \in \Theta_B\}$ in Fig. 1 lead to choosing the model $b_{\theta \in \Theta_B}$ according to the information theory type model selection criteria. Disappointingly, without a priori knowledge about the data and candidate models, there is no guarantee that the (jackknife) maximum likelihood estimator distinguishes the uniqueness of θ_g only to the larger model M_2.

As a matter of fact, practically, we are more interested in choosing a model among several different parametric models, or non-nested models (Bozdogan, 2000), while the true data generating function is unspecified. Of course, the distribution of the parameter vectors of the data generating function is unknown and the average of the maximized log likelihood is not guaranteed to converge to the expected value of the parametrized log likelihood under the true model. Therefore, the consistency of model selection with the jackknife method cannot be tested as the consistency of AIC or BIC is tested: it is known that BIC is consistent and parsimonious and AIC tends to choose larger models when candidate models are nested.

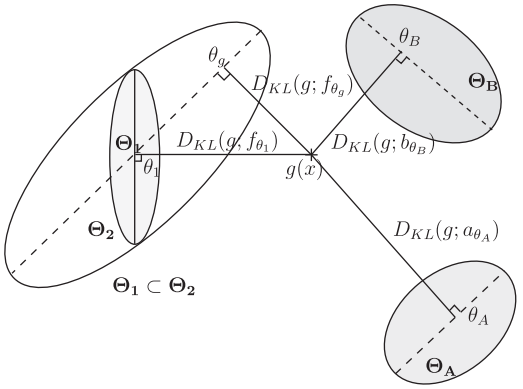

Fig. 1. An illustration of model parameter space: $g(x)$ is the unknown true model and Θ_A, Θ_B, Θ_1, and Θ_2 are parameter space of different models. Note that $\Theta_1 \subset \Theta_2$ and θ_g is the parameter of the model that mostly minimizes the KL divergence than parameters of other models (θ_A, θ_B, and θ_1). The solid lines indicate KL divergences from the true model to candidate models, f_{θ_g}, f_{θ_1}, a_{θ_A}, and b_{θ_B}, denoted by $D_{KL}(g; f_{\theta_g})$, $D_{KL}(g; f_{\theta_1})$, $D_{KL}(g; a_{\theta_A})$, and $D_{KL}(g; b_{\theta_B})$, respectively. Further detail is given in the text.

However, we would like to emphasize that traditional model selection methods should not be applied when the model selection process concerns non-nested candidate models or the unspecified true model.

6. Discussions

We showed that JIC is the asymptotically unbiased estimator of the maximized log likelihood so that the model of minimum JIC value indicates the unbiased minimum KL divergence model. Such model is the best approximating model among candidates, that accounts for the unknown true model in an asymptotically objective way.

Nonetheless, JIC cannot be applied universally since the estimates of JIC cannot be compared when candidate models are nested and the true model is specified. This traditional model selection condition that the truth is known or candidates are nested, has been the primary driving force to develop well known information criteria such as AIC, AICc, HQ, BIC, MDL and many variable selection methods. For example, in the field of time series, determining the correct and parsimonious order is the goal of model selection problems to avoid adopting the complex full order time series model. Choosing a parsimonious model is the result of trade-offs between model bias and model variance, both of which quantities are estimable only when the true model is specified. Thus, by using these well known model selection criteria, no comparison between autoregressive models of different orders and wavelet basis models occurs for choosing the best approximating model. On the contrary, JIC compares these two models and other time series fitting models whose parameter spaces are mutually exclusive but their likelihoods are estimable.

Nowadays, statistical model selection is applied in various disciplines and often candidate models are not nested to each other. These candidate models take various functional forms as their parameters have different topologies based on expertise in the subject matter. In addition, the true model is likely inaccessible so that choosing a working model, closest to the (unknown) true model has been the best strategy. Therefore, the model selection methods that minimize the distance between the unknown true model and candidates strongly depend on the choice of distance metrics. Namely, information criteria rely on the recipes of unbiased estimation of these metrics. Nevertheless, model selection methodologies and applications observe less interests in estimating/approximating the distributions of candidate models and their parameters, whereas the focus of AIC or BIC type information criteria has been verifying the model selection consistency with respect to the true model.

We have chosen the most celebrated KL divergence to approach model selection problems thanks to the diversity in information theory: the KL divergence is equivalent to Boltzmann's entropy and Shannon's information, and it enters naturally in maximum likelihood estimators. For the unbiased estimation purpose, we choose the jackknife, which is known for its bias reduction. With a different distance measure, Herzberg and Tsukanov (1985, 1986) concluded that the jackknife model selection approach provides a better criterion than Mallows' (1973) C_p when the correct model does not belong to the set of models under the verification process.

The most popular statistical resampling method is bootstrapping, which has been studied often in statistical model selection (see Chung et al., 1996; Ishiguro et al., 1997; Shibata, 1997; Shao, 1996; Cavanaugh and Shumway, 1997; Djurić, 1997; Feng and McCulloch, 1996; Hjorth, 1994; Neath and Cavanaugh, 2000). These studies did not considered the unspecified true model so that we cannot compare our JIC to the results from these studies. Yet, a few studies presented that the bias of bootstrap model selection is proportional to the number of parameters in a model (Chung et al., 1996; Ishiguro et al., 1997; Shibata, 1997). Notably, Chung et al. (1996) added that bootstrap after bootstrap provides an asymptotically unbiased estimate of the maximum log likelihood. It is worthwhile to point out that JIC is quite simple and inexpensive computation-wise as well as asymptotically unbiased, compared to bootstrap after bootstrap.

After the debut of AIC and other model selection criteria, consistency in selecting parsimonious model has been the primary interests. As discussed, when the candidate models are nested, JIC cannot achieve desirable consistency for the parsimonious model, the most sought-after property for a model selection criterion. A promising idea is that from Eq. (9), case-specifically JIC could be improved by (a) finding an estimator of $(1/n)\sum_{i\neq j}\nabla l_i(\theta_g)^T \Lambda^{-1}\nabla l_j(\theta_g)$, (b) considering 2nd order bias,[1] or (c) adding a data oriented penalty term. The virtue of JIC is that through the resampling technique, the asymptotically unbiased estimate of the maximum log likelihood can be obtained so that more general candidate models from various topological space can be tested to assess the true data generating model.

No single model selection criterion has the most desirable property compared to other model selection criteria. There are theoretical results showing that some of model selection criteria become optimal under proper conditions (Kuha, 2004). Different conditions lead to different conclusions, and arguably none of them capture the full complexity of real model selection problems. In addition, Sin and White (1996) pointed out that a smaller quantity from a model selection criterion does not conclude a correctly specified model; nonetheless, a correct model attains the smallest quantity of the information criterion. Information criterion for the model selection purpose is not a sufficient but a necessary condition of choosing a right model. Thus, a sole reliance to information criterion estimates without investigating a data set by its nature may mislead the conclusion. This weakness in model selection criteria was accentuated by Burnham and Anderson (2002), who made a statement in their textbook that the model selection methods become successful after better understanding of data sets and restricting possible candidates.

[1] Adams et al. (1971) studied the properties of delete one jackknife and delete two jackknife methods in parameter estimation.

The JIC may suffer inconsistency when candidate models are nested but it has wider possibilities than traditional model selection criteria in terms of applicability to non-nested models with an unspecified true model. Moreover, we are more likely to confront the situations that (a) the true model is unknown and (b) the best approximating model, minimizing the given distance metric, is more desired to be chosen instead the true model is assumed to be specified. Therefore, the bias reducing jackknife approach for estimating the KL divergence and the asymptotically unbiased maximum likelihood and its estimator is worth to be recognized as a model selection criterion.

Acknowledgments

We thank the referee for helpful comments. This work was supported in part by NSF Grant AST-0707833 (P.I.: G.J. Babu).

References

Adams, J.E., Gray, H.L., Watkins, T.A., 1971. An asymptotic characterization of bias reduction by jackknifing. Ann. Math. Statist. 42 (5), 1606–1612.
Akaike, H., 1973. Information theory and an extension of the likelihood ratio principle. In: Petrov, B.N., Csaki, F. (Eds.), Proceedings of the Second International Symposium of Information Theory. Akademiai Kiado, Budapest, pp. 257–281.
Akaike, H., 1974. A new look at statistical model identification. IEEE Trans. Automat. Control 19, 716–723.
Bai, Z.D., Rao, C.R., Wu, Y., 1999. Model selection with data-oriented penalty. J. Statist. Plann. Inference 77, 103–117.
Bozdogan, H., 2000. Akaike's information criterion and recent developments in information complexity. J. Math. Psychol. 44 (1), 62–91.
Burnham, K.P., Anderson, D.R., 2002. Model Selection and Inference, A practical Information-Theoretic Approach, 2nd ed. Springer-Verlag, New York.
Cavanaugh, J.E., Shumway, R.H., 1997. A bootstrap variant of AIC for state-space model selection. Statist. Sinica 7, 473–496.
Chung, H., Lee, K., Koo, J., 1996. A note on bootstrap model selection criterion. Statist. Probab. Lett. 26, 35–41.
Djurić, P.M., 1997. Using the bootstrap to select models. In: ICASSP'97, vol. 5, pp. 3729–3732.
Feng, Z.D., McCulloch, C.E., 1996. Using bootstrap likelihood ratios in finite mixture models. J. Roy. Statist. Soc. B 58, 609–617.
Firth, D., 1993. Bias reduction of maximum likelihood estimates. Biometrika 80 (1), 27–38.
Herzberg, A.M., Tsukanov, A.V., 1986. A note on modifications of the jackknife criterion for model selection. Utilitas Math. 29, 209–216.
Herzberg, A.M., Tsukanov, A.V., 1985. The Monte-Carlo comparison of two criteria for the selection of models. J. Statist. Comput. Simul. 22, 113–126.
Hjorth, J.S.U., 1994. Computer Intensive Statistical Methods: Validation, Model Selection and Bootstrap. Chapman & Hall, London.
Huber, P.J., 1967. The behavior of maximum likelihood estimates under nonstandard conditions. In: Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, vol 1, pp. 221–233.
Hurvich, C.M., Tsai, C., 1989. Regression and time series model selection in small samples. Biometrika 76 (2), 297–307.
Ishiguro, M., Sakamoto, Y., Kitagawa, G., 1997. Bootstrapping log likelihood and EIC, an extension of AIC. Ann. Inst. Statist. Math. 49 (3), 411–434.
Konishi, S., Kitagawa, G., 2008. Information Criteria and Statistical Modeling. Springer Series in Statistics, Springer, New York.
Konishi, S., Kitagawa, G., 1996. Generalised information criteria in model selection. Biometrika 83 (4), 875–890.
Kuha, J., 2004. AIC and BIC. Soc. Meth. Res. 33 (2), 188–229.
Kullback, S., Leibler, R.A., 1951. On information and sufficiency. Ann. Math. Statist. 22 (1), 79–86.
LeCam, L., 1953. On Some Asymptotic Properties of Maximum Likelihood estimates and Related Bayes Estimates. Univ. California Pub. Statist. 1, 277–330.
Mallows, C.L., 1973. Some comments on C_p. Technometrics 15, 661–675.
McQuarrie, A., Shumway, R., Tsai, C., 1997. The model selection criterion AICu. Statist. Probab. Lett. 34, 285–292.
Miller, G., 1974. The jackknife—a review. Biometrika 61 (1), 1–15.
Neath, A.A., Cavanaugh, J.E., 2000. A regression model selection criterion based on bootstrap bumping for use with resistant fitting. Comput. Statist. Data Anal. 35 (2), 155–169.
Nishii, R., 1988. Maximum likelihood principle and model selection when the true model is unspecified. J. Multivariate Anal. 27 (2), 392–403.
Protassov, R., et al., 2002. Statistics handle with care: detecting multiple model components with the likelihood ratio. Astrophys. J. 571, 545–559.
Rao, C.R., Wu, Y., 1989. A strongly consistent procedure for model selection in a regression problem. Biometrika 76, 369–374.
Shao, J., 1996. Bootstrap model selection. J. Amer. Statist. Assoc. 91, 655–665.
Shao, J., Tu, D., 1995. The Jackknife and Bootstrap. Springer-Verlag, New York.
Shao, J., 1993. Linear model selection by cross-validation. J. Amer. Statist. Assoc. 88, 486–494.
Shao, J., Wu, C.F.J., 1989. A general theory for jackknife variance estimation. Ann. Statist. 17 (3), 1176–1197.
Shibata, R., 1997. Bootstrap estimate of Kullback–Leibler information for model selection. Statist. Sinica 7, 375–394.
Sin, C., White, H., 1996. Information criteria for selecting possibly misspecified parametric models. J. Econometrics 71, 207–225.
Spiegelhalter, D.J., et al., 2002. Bayesian measures of model complexity and fit. J. Roy. Statist. Soc. B 64 (4), 583–639.
Stone, M., 1977. An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. J. Roy. Statist. Soc. B 39 (1), 44–47.
Takeuchi, K., 1976. Distribution of information statistics and a criterion of model fitting. Suri-Kagaku 153, 12–18 (in Japanese).
Wald, A., 1949. Note on the consistency of the maximum likelihood estimate. Ann. Math. Statist. 20 (4), 595–601.