# On an Expression of Generalized Information Criterion

Zeng-Hua Lu

School of Mathematics and Statistics, University of South Australia, City West, GPO Box 2471, SA 5001, Australia (zen.lu@unisa.edu.au)

## Abstract

Choosing a model selection criterion for model search when there present many candidate models can be a controversy. This is not a surprise as different criteria are derived with different objectives in mind. However, it is generally agreed that the Bayesian Information Criterion (BIC) and its generalized version, the Generalized Information Criterion (GIC) possess the consistency property – choosing the correct model with probability 1 as the sample size goes to infinite, as opposed to others such as the Akaike Information Criterion (AIC). In this paper, we suggest a particular expression of the GIC as replacing $\log N$ in the penalty term of the BIC with $(\log N)^r$, $0 < r < \infty$. Justifications from the Bayes Factor point of view are provided. The strong consistency property of the proposed criterion is established. Our consistency results include the consistency of selecting the closest model when the true model is not presented and the consistency of selecting the true model with the smallest model dimension when there are more than one true models are presented. Discussions concerning a choice of *r* and simulation studies are provided.

KEYWORDS: Model selection; GIC; BIC; Consistency, Linear regression, Mixture models

# 1. INTRODUCTION

Criterion based approach for model selection problem remains a popular choice among approaches such as hypothesis testing, Bayesian posterior prediction and cross-validation (see e.g. Berger and Pericchi 2000, George 2000, Broman and Speed 2002, Miller 2002 for excellent review), as well as the newly developed shrinkage-type approach (see e.g. Tibshirani 1996, Fan and Li 2001, Efron, Hastie, Johnstone and Tibshirani 2004). This mainly owes to the simplicity of its applications for the problem.

There have been many criteria proposed in the literature (See Miller 2002 for the survey of this topic). Most criteria can be written in the form of

$$l_N(\theta) - \frac{d}{2}\lambda \tag{1}$$

where $l_N(\theta)$ is the sample log-likelihood function of a model that has parameter $\theta$ of dimension $d$, $N$ is sample size and $\lambda$ takes different forms in different criteria. The well-known Akaike Information Criterion (AIC) (Akaike 1973) and the Bayesian information criterion (BIC) (1978) take $\lambda = 2$ and $\log N$ respectively. In the criterion based approach for model selection, one chooses a model, which maximizes (1) among a finite number of candidate models. There has always been a controversy concerning which model criterion should be used. This is not a surprise as different criteria are derived with different objectives in mind. However, there is a general consent among statistical professionals that the BIC and its generalized version, the Generalized Information Criterion (GIC) (Rao and Wu 1989) where $\lambda$ satisfies $N^{-1}\lambda \to 0$ and $(\log\log N)^{-1}\lambda \to \infty$ as $N \to \infty$ enjoy the consistency property of correctly identifying the true model with probability 1. Such a property is not shared by the AIC or any criteria in which $\lambda$ is fixed (Foster and George 1994, Yang 2005, 2007).

This paper is concerned with a criterion possessing the consistency property. We suggest a particular form of expression of the GIC as

$$\lambda = c(\log N)^r . \tag{2}$$

The AIC corresponds to $c = 2$ and $r = 0$, and the BIC corresponds to $c = 1$ and $r = 1$. Furthermore, with $0 < r < \infty$, for any fixed $c$ (2) satisfies the GIC's conditions above. Therefore, it qualifies to be the GIC. As we are interested in the consistency property, we could fix $c$ say $c = 1$, so that the AIC can be reflected through $r = \log 2 / \log \log N$ for each fixed $N$.

In the next section, we justify our criterion from the Bayes Factor point of view (e.g. Berger, et al. 2000), In particular it is possible for $r$ to take a value other than 1 as an approximation to the posterior density function.

Section 3 studies the consistency property of our criterion that includes the BIC in a general setting. It is commonly believed that one of the assumptions for the consistency property of the BIC to hold is that the true model needs to be presented among the candidate models (e.g. Haughton 1988, Shao 1997). Our results reveal that when the assumption is violated, our criterion leads to consistently choosing the closest model to the true model among candidate models. Our consistency results also include the consistency of selecting the true model with the smallest model dimension when there are more than one true models are presented in candidate models.

Section 4 discusses an application of our criterion for covariate selection in linear regression models where it has been reported that the BIC causes underfitting while the AIC results in overfitting. We suggest take $r = 0.5$ as a practical choice to bridge the strengths of the AIC and BIC. Some justifications and simulation studies are provided.

Finally, some concluding remarks are made in Section 5.

## 2. VIEWPOINT OF BAYES FACTOR

Suppose there are finite $q$ models, defined by $m_i$, $i = 1,...,q$, whose joint density for observed data $Y$ is $f(Y;\theta_i)$, where $\theta_i$ is a vector of unknown model parameters. If we have available prior distributions concerning the model $m_i$ and the parameter $\theta_i$ as, $\pi(m_i)$ and $\pi(\theta_i)$, then by Bayes' theorem, the posterior probability of $m_i$ is obtained as,

$$P(m_i;Y) = \frac{P(Y;m_i)\pi(m_i)}{\sum_{i=1}^{q}\pi(m_i)P(m_i)}, \tag{3}$$

where

$$P(Y;m_i) = \int f(Y;\theta_i)\pi(\theta_i)d\theta_i, \tag{4}$$

is the marginal density of model $m_i$. Note that the functions $f$, $\pi$ and $P$ are in general different for each model, and $\pi$'s are different priors for $\theta_i$ and $m_i$, but we do not make it specific here for simplicity of presentation. The ratio of the posterior probabilities of the models $m_i$ and $m_j$, $i \neq j$, $i, j \in \{1,...,q\}$, is

$$\frac{P(m_i;Y)}{P(m_j;Y)} = \frac{P(Y;m_i)}{P(Y;m_j)} \cdot \frac{\pi(m_i)}{\pi(m_j)},$$

where the first term of the right hand side is called the Bayes Factor (BF), i.e.

$$B_{ij} = \frac{P(Y;m_i)}{P(Y;m_j)} = \frac{\int f(Y;\theta_i)\pi(\theta_i)d\theta_i}{\int f(Y;\theta_j)\pi(\theta_j)d\theta_j}. \tag{5}$$

Commonly researchers assign $P(m_i) = 1/q$, $\forall i$. Then,

$$B_{ij} = \frac{P(m_i;Y)}{P(m_j;Y)}.$$

Therefore, from the Bayesian point of view one chooses model $i$ if $B_{ij} > 1$; model $j$ if $B_{ij} < 1$. Computing BF in (5) involves evaluation of integrals, hence is usually computationally demanding. The Laplace approximation to the integral $P(Y;m_i)$ has been suggested (see e.g. Lindley 1980, Tierney and Kadane 1986, Kass and Raftery 1995). For the independent and identically distributed (iid) observations, $Y=(y_1,...y_N)'$, with the density $f(y_n;\theta_i)$, $n=1,...,N$, the approximation is

$$P(Y;m_i)=(\frac{2\pi}{N})^{\frac{d_i}{2}}\left|\Sigma(\theta_i)\right|^{\frac{1}{2}}L(\theta_i)\pi(\theta_i)[1+O(N^{-1})], \qquad (6)$$

where $L(\theta_i)=\prod_{n=1}^{N}f(y_n;\theta_i)$ and $\Sigma(\theta_i)=[-\partial^2\log f(y;\theta_i)/(\partial\theta_i\partial\theta_i)]^{-1}$.

The well-known BIC can be viewed as taking the leading terms of the logarithm of (6) in the view of BF as

$$BIC_i=l_N(\theta_i)-\frac{d_i}{2}\log N,$$

where $l_N(\theta_i)=\sum_{n=1}^{N}\log f_i(y_n;\theta_i)$. Obviously the approximation (6) is accurate to the order of $O(N^{-1})$. That means we may let the penalty term $(d_i/2)\log N$ be replaced by $(d_i/2)\log N+O(\log N)=(d_i/2)(\log N)^r$, $r>0$. Note that $r\le 0$ does not satisfy $(\log\log N)^{-1}(\log N)^r\to\infty$, hence it does not qualify to be GIC. Therefore, we have what we call Bayesian-like Generalized Information Criterion (BGIC),

$$BGIC_i=l(\theta_i)-\frac{d_i}{2}(\log N)^r. \qquad (7)$$

*Example 1:*

Kass and Wasserman (1995, Example 1). Let $y_n\sim N(\psi,1)$ and consider the normal unit-information prior $\psi\sim N(0,1)$, the exact posterior density function.

$$\log P(y;m)=l_0+\frac{N\overline{y}^2}{2}\frac{N}{N+1}-\frac{1}{2}\log(N+1)$$

where $l_0 = -\frac{N}{2}\log\frac{2\pi}{N} - \frac{N\sum_{n=1}^{N} y_n{}^2}{2}$ and $\overline{y} = \frac{\sum_{n=1}^{N} y_n}{N}$.

$$BIC = l_0 + \frac{N\overline{y}^2}{2} - \frac{1}{2}\log N .$$

In this case, in the view of (7), we have

$$\left(\log N\right)^r = \frac{N\overline{y}^2}{N+1} + \log\left(N+1\right). \tag{8}$$

There exists $r > 1$ for $N \geq 3$. To show this, let

$$\Delta(r) = \left(\log N\right)^r - \frac{N\overline{y}^2}{N+1} - \log\left(N+1\right),$$

it is obvious that $\Delta(r)$ is monotonically increasing function in $r$ and $\Delta(1) < 0$ and $\Delta(\infty) = \infty$ for $N \geq 3$.

## 3. CONSISTENCY

It has been shown that the AIC does not possess the consistency property while it has an advantage over the BIC in terms of risk inflation (see e.g. Yang 2005). The consistency property of the BIC and GIC has been shown for different models in the literature. Haughton (1988) proved the consistency of the BIC for the exponential distributions. Nishii (1984), Rao and Wu (1989) and Shao (1997) showed the consistency of the GIC for covariates selection in the context of linear regression models. The consistency property of the criterion based approach for estimating the number of mixture components in mixture models has been shown by a number of authors. Leroux (1992) showed that the estimate of number of mixture components is not under fitting its true value when the true model (with the true number of components) is among the candidate models. Keribin (2000) and Chambaz (2006) proved the consistency of the GIC for estimating the number of mixture components in mixture models. The consistency results under the non-iid setting have also been established by a number of authors such as Pötscher (1989) and Niu and Ang (2003).

In this section we study a general approach to the consistency property of our criterion that includes the BIC in the iid setting. The existing consistency results require the assumption that the true model is presented in candidate models, but we show that when the assumption is violated, the closest model to the true is selected with probability 1 as the sample size goes to infinite. We also show that, when there exist more than one true model, the true model with the smallest dimension is then selected with probability one as the sample size goes to infinite.

Suppose the iid observations $Y = (y_1, ..., y_N)'$ are generated from the true distribution function $G(y)$ and density function $g(y)$. Suppose there are a finite set of $q_k$ true models among candidate models, $M_g = \{m_k, k = 1, ..., q_k\}$, with the density function $g_k(y;\theta_k)$. To accommodate the situation of non-identifiable models, we allow there may exist a model $m_k \in M_g$ such that $g_k(y;\theta_{k01}) = g_k(y;\theta_{k02})$ for $\theta_{k01} \neq \theta_{k02}$, $\theta_{k01}, \theta_{k02} \in \Theta_{k0} \subset \Theta_k$. For example, suppose the true model is $y \sim N(0,1)$. If a candidate model is $y \sim N(\mu, \sigma^2)$, then there is an unique parameter point $(\mu, \sigma^2) = (0,1)$ such that the candidate model becomes the true model. But if a candidate model is a two-component mixture

$$\begin{cases} y \sim N(\mu_1, \sigma_1^2) & \text{with the probability } \alpha, \\ y \sim N(\mu_2, \sigma_2^2) & \text{with the probability 1-}\alpha, \end{cases}$$

then the true model can be recovered with $\alpha = 1$, $(\mu_1, \sigma_1^2) = (0,1)$, or $\alpha = 0$, $(\mu_2, \sigma_2^2) = (0,1)$, or $\mu_1 = \mu_2$, $\sigma_1^2 = \sigma_2^2$. We also suppose that there are a finite set of $q_{\tilde{k}}$ non-true models among candidate models, $M_f = \{m_{\tilde{k}}, \tilde{k} = 1, ..., q_{\tilde{k}}\}$, with the density functions $f_{\tilde{k}}(y;\theta_{\tilde{k}})$. Note that we use the subscript $k$ and $\tilde{k}$ to index the true and non-true candidate models, respectively, and $i$ to index candidate models, $i \in \{k\} \cup \{\tilde{k}\}$. Let $\theta_i \in \Theta_i \subset R^{d_i}$. Denote the expectation and empirical measure of, say $f$, as $Gf = \int f dG$ and

$G_N f = N^{-1}\sum_{n=1}^{N} f(y_n)$, respectively. Denote $\xrightarrow{a.s.}$ as convergence almost surely. Let $\hat{\theta}_i$ be the Maximum Likelihood Estimator (MLE) of the model $m_i$.

Assume that $g_k$ and $f_{\tilde{k}}$ are $\sigma-$finite measurable probability density functions with the regularity stated below. We state the regularity conditions for  below.

(C1) $\Theta_i, \forall i$, are compact.

(C2) $g_k$ and $f_{\tilde{k}}$ are dominated for $\forall k$ and $\forall \tilde{k}$, i.e. $|g_k| \le b_1(y)$ and $|f_{\tilde{k}}| \le b_2(y)$, where $b_1(y)$ and $b_2(y)$ are continuous on $Y$ and integrable with respect to $G$.

(C3) $g_k$ and $f_{\tilde{k}}$, $\forall k$, $\forall \tilde{k}$, are almost surely continuous on $Y \times \theta_k$, and $Y \times \theta_{\tilde{k}}$, respectively.

*Remarks*: The compactness condition (C1) may involve other restrictive conditions for models such as mixture normal models (see e.g. Hathaway 1985). The domination condition (C2) ensures the existence of the expectations $G \log g_k$, $G \log f_{\tilde{k}}$, $G|\log g_k|$ and $G|\log f_{\tilde{k}}|$, $\forall k$, $\forall \tilde{k}$. The nonidentifiability problem of the kind discussed above is allowed for $g_k$. However, such problem is not permitted for $f_{\tilde{k}}$; there exist no point $\theta_{\tilde{k}}$ such that $f_{\tilde{k}}(\theta_{\tilde{k}}) = g$.

Suppose one of candidate models is selected if the corresponding model selection criterion evaluated at the MLE, $l_N(\hat{\theta}_i) - p_{d_i,N}$ is maximized . The following conditions are assumed for the penalty function $p_{d_i,N}$.

(P1) $p_{d_{i_1},N} < p_{d_{i_2},N}$, $\forall N$ if $d_{i_1} < d_{i_2}$.

(P2) $\frac{p_{d_i,N}}{N} \xrightarrow{a.s.} 0$.

(P3) $\frac{p_{d_i,N}}{\log\log N} \xrightarrow{a.s.} \infty$.

*Theorem 1. Assume conditions (C1) – (C3) and (P1) – (P3) are satisfied.*

(i) If the set of candidate models is $M_f \cup m_k$ (i.e. $m_k$ is the only true model; $M_g = m_k$ and $q_k = 1$), then $\lim_{N\to\infty} \Pr(\hat{m}_i = m_k) = 1$, $i = 1,...,q_{\tilde{k}} + 1$, a.s..

(ii) If the set of candidate models is $M_f$ (i.e. it does not include a true model), then $\lim_{N\to\infty} \Pr(\hat{m}_i = m^*) = 1$, $i = 1,...,q_{\tilde{k}}$, a.s., where $m^*$ is the model whose density function $f^*$ is closest to $g$ in the Kullback-Leibler (KL) measure among the models in $M_f$.

(iii) If the set of candidate models is $M_f \cup M_g$ with $M_g = \{m_k, q_k > 1\}$ (i.e. it contains more than one true model), then $\lim_{N\to\infty} \Pr(\hat{m}_i = m_k^*) = 1$, $i = 1,...,q_{\tilde{k}} + q_k$, a.s., where $m_k^*$ is the true model with the smallest dimension $d_k^*$ among the models in $M_g$.

It is straightforward to verify that our criterion that includes the BIC satisfies Conditions (P1) – (P3). Our criterion as a generalized version of the BIC possesses the consistency results stated in Theorem 1.

*Remarks.*

1. There is common belief that the consistency of BIC relies on the true model being included in the candidate models (e.g. Haughton 1988, Shao 1997). Our results in (ii) do not require such assumption.
2. The results in (i) and (ii) hold regardless of size of model dimension, *d*, which implies the results hold even when the true model in (i) or the closest model in (ii) have a large model dimensionality.

3. $\lambda = O(\log\log N)$ proposed by Hannan and Quinn (1979) does not satisfy Condition P(3), hence does not give rise to the results in (iii).

4. $\lambda = O(N/\log N)$ suggested by Rao and Tibshirani (1997) satisfies all conditions P(1)-P(3). But as $(\log N)^r/(N/\log N) \to 0$ as $N \to \infty$, it leads to a more aggressive criterion than our BGIC for all $r > 0$. In fact it quickly goes well above 5 for a moderate sample size (see Table 1).

## 4. CHOICE OF $r$

### 4.1. Reporting the range of $r$ values

As we have seen in Section 2, different choices of priors on $\theta$ and $m$, as well as number of terms of the Laplace expansion being used have impacts on the value of $r$. It is generally difficult to decide what value $r$ should take. Even for the AIC, it is also derived in an approximate form (Akaike 1973). In fact, Zhang (1996) recommended $\lambda$ in (1) can takes a value from the range between 1 and 5 with a larger value leading to a more parsimony model. On the other hand, there exist situations, where one model is preferred over another for a range of values of $r$ based on the sample information. Therefore, we suggest one should report such a range of $r$ values whenever possible.

For example, suppose $l_N(\hat{f}_{i_1}) > l_N(\hat{f}_{i_2})$, $n \geq 3$ for two models $m_{i_1}$ and $m_{i_2}$. Model $m_{i_1}$ is preferred over $m_{i_2}$ for

$$0 < r < \frac{\log 2[l_N(\hat{f}_{i_1}) > l_N(\hat{f}_{i_2})] - \log(d_{i_1} - d_{i_2})}{\log\log N}$$

if $d_i > d_j$; for $0 < r < \infty$ if $d_i < d_j$.

One immediately knows whether the preference is suggested by AIC not BIC if $0 < r < 1$, or both AIC and BIC if $r \geq 1$.

Reporting $r$ range provides evidence of model fit in terms of how much penalty a preferred model can afford to. In some senses, it offers some sort of P-value ideas for model comparison; $r = \log 2 / \log\log N$ and 1 corresponding to the AIC and BIC offer benchmark values like the significance levels. It suggests to what extent the choice of a model is made by looking at how close the upper bound $r$ is away from $\log 2 / \log\log N$ (AIC) and 1 (BIC).

### 4.2. Choice of *r*

As demonstrated in Section 2, the derivation of the BIC involves an use of the Laplace approximation, accuracy of which closely relates to the model density function and prior distribution function. It implies that it might be difficult to find an uniform Bayesian type criterion that works well for in different model context. We suggest that a simulation based technique such as cross validation (CV) may be used for choosing an *r* value in our criterion. Such choice would provide some understanding of those factors influencing the BF in a particular model context. One would then use such value in a similar model context with a view that the consistency of model selection is held with a sample size approaches to infinity.

### 4.3. Linear regression

As covariate selection in linear regression is an important application of model selection criteria, (see e.g. Foster, et al. 1994, Shao 1997, George 2000, Yang 2005, 2007 and references therein), we discuss a choice of $r$ of our criterion in such applications. It has been reported in the literature (Shibata 1976, Shao 1997, Yang 2007), that while AIC is too conservative in selecting covariates, BIC on the other hand often is too liberal. In other words, while the AIC often results in selecting too many covariates, BIC often just retains those very influential ones. These observations naturally lead us to suggest to choose a value between $\log 2 / \log\log N$ and $1$. Shown in Table 1, $\log 2 / \log\log N \approx 0.508$ for $N = 50$ then slowly dies off as $N$ increases. It is about 0.220 when $N = 10^{10}$.

If we choose $r = 0.5$, from the comparison among the columns under, our criterion would roughly match AIC for a small sample size, while enjoying the consistency property for a large sample size. Although our criterion with $0 < r < 1$ has a slower rate of identifying the true model or the closest model compared to the BIC, it provides an opportunity to bridge the strengths of the AIC and BIC when sample size is large.

Foster and George (1994), Shao (1997) and Yang (2005, 2007) showed that the BIC has much greater model risk in terms of the estimated model deviating from the true model than the AIC for a large sample size. On the other hand, Zhang (1992) argued that $\lambda > 5$ is too great a penalty for overfitting consideration. From Table 1 it is clear that our criterion with $r = 0.5$, i.e. $\lambda = \sqrt{\log N}$ satisfies Zhang's suggestion, up to a very large N, $N = \exp(25) = 7.2 \times 10^{10}$; our criterion meets most practical needs.

### 4.4. Simulation studies

## APPENDIXEquation Section (Next)

We first show the following strong consistency. The results hold even when there exists the non-identifiability problem discussed in Section 3.

*Lemma 1. If the conditions (C1) - (C3) are satisfied, then* $G_N \log g_k(\hat{\theta}_k) \xrightarrow{a.s.} G \log g$ , $\forall k$ *, and* $G_N \log f_{\tilde{k}}(\hat{\theta}_{\tilde{k}}) \xrightarrow{a.s.} G \log f_{\tilde{k}}(\theta_{\tilde{k}0})$ , $\forall \tilde{k}$ *, where* $\theta_{\tilde{k}0}$ *is the maximum of* $G \log f_{\tilde{k}}(\theta_{\tilde{k}0})$ .

*Proof.* When a model does not involve the nonidentifiability problem, the standard MLE results ensure $G_N \log g_k(\hat{\theta}_k) \xrightarrow{a.s.} G \log g$ . For a misspecified model $m_{\tilde{k}} \in M_f$ , such strong consistency result also hold, i.e. $G_N \log f_{\tilde{k}}(\hat{\theta}_{\tilde{k}}) \xrightarrow{a.s.} G \log f_{\tilde{k}}(\theta_{\tilde{k}0})$ , (see e.g White 1994). strong consistency of MLE $\hat{\theta}_i \overset{a.s}{\rightarrow} \theta_i$ readily gives rise to the above results by the continuous mapping theorem.

When a model does involve the nonidentifiability problem for a model $m_k \in M_g$ , we now show $G_N \log g_k(\hat{\theta}_k) \xrightarrow{a.s.} G \log g$ . Our proof is based on Feng and McCulloch's (1996) idea, but provides an alternative approach, which extends a standard argument for the consistency in the ML context. Following the idea presented in, we define $A_\delta(\dot{\theta}_{k0}) \subset \Theta \setminus \Theta_{k0}$ as an open neighbourhood around $\Theta_{k0}$ such that $\left\| \theta_k - \dot{\theta}_{k0} \right\| < \delta$ , where $\|.\|$ is the Euclidean norm, $\delta > 0$ , $\theta_k \in A_\delta(\dot{\theta}_{k0})$ and $\dot{\theta}_{k0} \in \Theta_{k0}$ is the selected point such that it is the closest point in $\Theta_{k0}$ to $\theta_k$ , i.e., $\left\| \theta_k - \dot{\theta}_{k0} \right\| \leq \left\| \theta_k - \ddot{\theta}_{k0} \right\|$ for all $\ddot{\theta}_{k0} \in \Theta_{k0}$ , and $\ddot{\theta}_{k0} \neq \dot{\theta}_{k0}$ . Denote

$$c(y, \dot{\theta}_{k0}, \delta) = \sup_{\theta_k \in A_\delta(\dot{\theta}_{k0})} \left\| \log g_k(y, \theta_k) - \log g_k(y, \dot{\theta}_{k0}) \right\| .$$

Because of the compactness condition (C1), $A_\delta(\dot{\theta}_{k0})$ can be reduced to a finite number of open coverings $A_{\delta,j} = A_\delta(\dot{\theta}_{k0,j})$ , $j = 1,...,J$ , such that for each $A_{\delta,j}$ , we have

$$\begin{aligned}
&\left\| G_N \log g_k(\theta_k) - G\log g_k(\dot{\theta}_{k0}) \right\| \\
&= G_N \left\| \log g_k(\theta_k) - \log g_k(\dot{\theta}_{k0,j}) \right\| + \left\| G_N \log g_k(\dot{\theta}_{k0,j}) - G\log g_k(\dot{\theta}_{k0,j}) \right\| \\
&\qquad + \left\| G\log g_k(\dot{\theta}_{k0,j}) - G\log g_k(\dot{\theta}_{k0}) \right\| \\
&= \{G_N c(y,\dot{\theta}_{k0,j},\delta) - Gc(y,\dot{\theta}_{k0,j},\delta)\} + Gc(y,\dot{\theta}_{k0,j},\delta) \\
&\quad + \left\| G_N \log g_k(\dot{\theta}_{k0,j}) - G\log g_k(\dot{\theta}_{k0,j}) \right\| + \left\| G\log g_k(\dot{\theta}_{k0,j}) - G\log g_k(\dot{\theta}_{k0}) \right\|.
\end{aligned}$$

Following Tauchen (1985, p439), we have that for any $\delta > 0$, there exists $\omega > 0$, such that $E[c(y,\dot{\theta}_{k0,j},\delta(\dot{\theta}_{k0,j}))] \le \omega$, and $\left\| N^{-1} l_N(\log g_k(\theta_k)) - G\log g_k(\dot{\theta}_{k0}) \right\| \le 4\omega$, $\theta_k \in A_{\delta,j}$, whenever $N \ge N(\delta(\dot{\theta}_{k0,j}))$ a.s.. Thus $G_N \log g_k(\theta_k) \xrightarrow{a.s.} G\log g_k(\dot{\theta}_{k0})$. Finally by the standard argument, as $\hat{\theta}_k$ maximizing $G_N \log g_k(\theta_k)$, we have $G_N \log g_k(\hat{\theta}_k) \xrightarrow{a.s.} G\log g_k(\dot{\theta}_{k0})$.

*Proof of Theorem 1.*

(i) Define the KL measure (Kullback and Leibler 1951) as

$$K(g,f) = \int \log \frac{g}{f} dG$$

The non-negativity property of the KL measure gives $K(g,f) > 0$ if $f \ne g$ and $K(g,f) = 0$ if $f = g$..

By the strong Uniform Law of Large Numbers (ULLN)

$$\frac{l_N(g_k(\theta_k)) - l_N(f_{\tilde{k}}(\theta_{\tilde{k}}))}{N} \xrightarrow{a.s} K(g_k, f_{\tilde{k}}). \tag{A.1}$$

Because

$$\begin{aligned}
&\frac{1}{N}\{l_N(g(\hat{\theta}_k)) - p_{d_k,N} - [l_N(f_{\tilde{k}}(\hat{\theta}_{\tilde{k}})) - p_{d_{\tilde{k}},N}]\} \\
&= \frac{1}{N}\{l_N(g(\hat{\theta}_k)) - l_N(f_{\tilde{k}}(\hat{\theta}_{\tilde{k}}))\} - \frac{p_{d_k,N}}{N} + \frac{p_{d_{\tilde{k}},N}}{N} \\
&\xrightarrow{a.s} K(g, f_{\tilde{k}}) > 0,
\end{aligned}$$

the third line above follows the strong consistency property of Lemma 1 and Condition (P2).

Therefore, for $k = 1$, $m_{\tilde{k}} \in M_f$, $\forall \tilde{k}$,

$$\lim_{N\to\infty} \Pr(l_N(g(\hat{\theta}_k)) - p_{d_k,N} > l_N(f_{\tilde{k}}(\hat{\theta}_{\tilde{k}})) - p_{d_{\tilde{k}},N}) = 1 \text{ a.s.}$$

(ii) $K(g,f)$ also serves as a measure of the closeness of $f$ to $g$ (Akaike 1973), e.g. for two non-true densities $f_{\tilde{k}_1}$ and $f_{\tilde{k}_2}$, we have $K(g,f_{\tilde{k}_1}) < K(g,f_{\tilde{k}_2})$, if $f_{\tilde{k}_1}$ is closer to $g$ than $f_{\tilde{k}_2}$. Define

$$\begin{aligned} D(f_{\tilde{k}_1}, f_{\tilde{k}_2}) &= K(g,f_{\tilde{k}_1}) - K(g,f_{\tilde{k}_2}) \\ &= \int \log f_{\tilde{k}_2}\, dG(y) - \int \log f_{\tilde{k}_1}\, dG(y). \end{aligned}$$

Because $K(g,f) > 0$, if $f \neq g$, we have $D(f_{\tilde{k}_1}, f_{\tilde{k}_2}) < 0$, if $f_{\tilde{k}_1}$ is closer than $f_{\tilde{k}_2}$ to $g$; $D(f_{\tilde{k}_1}, f_{\tilde{k}_2}) > 0$, otherwise.

By the strong ULLN,

$$\frac{l_N(f^*(\theta^*)) - l_N(f_{\tilde{k}}(\theta_{\tilde{k}}))}{N} \xrightarrow{a.s} D(f_{\tilde{k}}, f^*) > 0$$

where $f^*$ is closest to $g$ among all non the true candidate models in $M_f$.

Therefore, together with the result of Lemma 1, we have

$$\frac{1}{N}\{l_N(f^*(\hat{\theta}^*)) - P_{d^*,N} - [l_N(f_{\tilde{k}}(\hat{\theta}_{\tilde{k}})) - P_{d_{\tilde{k}},N}]\} \xrightarrow{a.s} D(f_{\tilde{k}}, f^*) > 0,$$

i.e.,

$$\lim_{N\to\infty} \Pr(l_N(f^*(\hat{\theta}^*)) - p_{d^*,N} > l_N(f_{\tilde{k}}(\hat{\theta}_{\tilde{k}})) - p_{d_{\tilde{k}},N}) = 1 \text{ a.s.}$$

(iii) We will present two lemmas before we turn to show our result.

Consider a model $m_k \in M_g$. Let $\theta_k \in A_\delta(\dot{\theta}_{k0}) \subset \Theta \setminus \Theta_{k0}$ for any $\delta > 0$. Note that if $m_k$ is identifiable, then $\Theta_{k0}$ collapses to a single point. Let $h(y;\theta_k) = \log g_k(y;\theta_k) - \log g_k(y;\theta_{k0})$, $\theta_{k0} \in \Theta_{k0}$. The non-negativity property of the KL measure suggests $-Gh > 0$.

*Lemma 2. If* $\theta_k$ *maximizes* $G_N \log g_k(y;\theta_k)$ *,*

$$G_N h(y;\theta_k) \le \{(G_N - G)\frac{h(y;\theta_k)}{\sqrt{-Gh(y;\theta_k)}}\}^2 . \tag{A.2}$$

*Proof:*

The proof can be established by applying the technique presented in Chambaz (2006, Proposition A.1). Consider

$$G_N h - Gh = (G_N - G)h . \tag{A.3}$$

As $-Gh > 0$, (A.3) gives rise to

$$G_N h < \sqrt{-Gh(y;\theta_k)}(G_N - G)\frac{h(y;\theta_k)}{\sqrt{-Gh(y;\theta_k)}} . \tag{A.4}$$

Since $\theta_k$ maximizes $G_N \log g_k(y;\theta_k)$, $G_N h > 0$. Therefore (A.3) also gives rise to

$$-Gh < \sqrt{-Gh(y;\theta_k)}(G_N - G)\frac{h(y;\theta_k)}{\sqrt{-Gh(y;\theta_k)}} ,$$

$$\sqrt{-Gh(y;\theta_k)} < (G_N - G)\frac{h(y;\theta_k)}{\sqrt{-Gh(y;\theta_k)}} . \tag{A.5}$$

From (A.4) and (A.5), we have (A.2).

*Lemma 3.*

For $m_{k_1}, m_{k_2}, \in M_g$ ,

$$\frac{l_N(g_{k_1}(\hat{\theta}_{k_1})) - l_N(g_{k_2}(\hat{\theta}_{k_2}))}{p_{d_{k_s},N}} \xrightarrow{a.s.} 0 ,$$

where $s = 1, 2$ .

*Proof:*

From Lemma 2, we have for any $\theta_{k_1 0} \in \Theta_{k_1 0}$

$$\left| l_N(g_{k_1}(\hat{\theta}_{k_1})) - l_N(g_{k_1}(\theta_{k_1 0})) \right| = N \left| G_N h(y;\theta_{k_1}) \right|$$
$$\leq N\{(G_N - G)\frac{h(y;\theta_{k_1})}{\sqrt{-Gh(y;\theta_{k_1})}}\}^2.$$

By the bounded Law of Iterated Logarithm (van der Vaart 1998, p. 19), we have

$$\frac{\sqrt{N}(G_N - G)\frac{h(y;\theta_{k_1})}{\sqrt{-Gh(y;\theta_{k_1})}}}{\sqrt{\log\log N}} < c_1, \text{ a.s.}$$

where $0 < c_1 < \infty$.

Therefore,

$$\left| \frac{l_N(g_{k_1}(\hat{\theta}_{k_1})) - l_N(g_{k_1}(\theta_{k_1 0}))}{p_{d_{k_s},N}} \right| \leq \frac{N}{p_{d_{k_s},N}}\{(G_N - G)\frac{h(y;\theta_{k_1})}{\sqrt{-Gh(y;\theta_{k_1})}}\}^2,$$
$$\leq \frac{\log\log N}{p_{d_{k_s},N}} c_1{}^2.$$

Because of (P3), we have

$$\frac{l_N(g_{k_1}(\hat{\theta}_{k_1})) - l_N(g_{k_1}(\theta_{k_1 0}))}{p_{d_{k_s},N}} \xrightarrow{a.s.} 0.$$

Similarly, we have for any $\theta_{k_2 0} \in \Theta_{k_2 0}$

$$\frac{l_N(g_{k_2}(\hat{\theta}_{k_2})) - l_N(g_{k_2}(\theta_{k_2 0}))}{p_{d_{k_s},N}} \xrightarrow{a.s.} 0.$$

Finally, because $g_{k_1}(\theta_{k_1 0}) = g_{k_2}(\theta_{k_2 0})$, we have

$$\frac{l_N(g_{k_1}(\hat{\theta}_{k_1})) - l_N(g_{k_2}(\hat{\theta}_{k_2}))}{p_{d_{k_s},N}}$$
$$= \frac{l_N(g_{k_1}(\hat{\theta}_{k_1})) - l_N(g_{k_1}(\theta_{k_1 0}))}{p_{d_{k_s},N}} - \frac{l_N(g_{k_2}(\hat{\theta}_{k_2})) - l_N(g_{k_2}(\theta_{k_2 0}))}{p_{d_{k_s},N}},$$
$$\xrightarrow{a.s.} 0$$

We now turn to prove the result presented in Part (iii) of Theorem 1. For any two models $m_{k_1}, m_{k_2}, \in M_g$ with $d_{k_1} < d_{k_2}$. Consider

$$\Pr(l_N(\hat{g}_{k_1}) - p_{d_{k_1},N} > l_N(\hat{g}_{k_2}) - p_{d_{k_2},N}) = \Pr(\frac{l(\hat{g}_{k_1}) - l(\hat{g}_{k_2})}{p_{d_{k_2},N}} > \frac{p_{d_{k_1},N}}{p_{d_{k_2},N}} - 1) .$$

Because of Lemma 3, $\{l_N(\hat{g}_{k_1}) - l_N(\hat{g}_{k_2})\} / p_{d_{k_2},N} \xrightarrow{a.s.} 0$. By condition (P1), $d_{k_1} < d_{k_2}$ results in $p_{d_{k_1},N} / p_{d_{k_2},N} - 1 < 0$. Therefore,

$$\lim_{N\to\infty} \Pr(l(\hat{g}_{k_1}) - p_{d_{k_1},N} > l(\hat{g}_{k_2}) - p_{d_{k_2},N}) = 0 \text{ a.s.}$$

for all $m_{k_1}, m_{k_2} \in M_g$ with $d_{k_1} > d_{k_2}$. This implies,

$$\lim_{n\to\infty} \Pr(l(\hat{g}_k^*) - p_{d_k^*,N} > l(\hat{g}_k) - p_{d_k,N}) = 1 \text{ a.s.} \tag{A.6}$$

for any $m_k \in M_g$ with $d_k > d_k^*$.

Also because of the result in (i), we have

$$\lim_{N\to\infty} \Pr(l(\hat{g}_k^*) - p_{d_k^*,N} \geq l(\hat{f}_{\tilde{k}}) - p_{d_{\tilde{k}},N}) = 1 \text{ a.s.} \tag{A.7}$$

where $m_{\tilde{k}} \in M_f$. Combining (A.6) and (A.7) prove the results presented in (iii).

## ACKNOWLEDGMENTS

This research was supported in part by the Australian Research Council's Discovery project grants DP0666677 and DP0880796.

## References

*Akaike, H. (1973), "Information Theory and an Extension of the Maximum Likelihood Principle," in 2nd International Symposium on Information Theory, Tsahkadsor, Armenian SSR, pp. 267-281.*

*Berger, J., and Pericchi, L. (2000), "Objective Bayesian Methods for Model Selection: Introduction and Comparison," in Model Selection, ed. P. Lahiri, Institute of Mathematical Statistics.*

*Broman, K. W., and Speed, T. P. (2002), "A Model Selection Approach for the Identification of Quantitative Trait Loci in Experimental Crosses," Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64, 641-656.*

*Chambaz, A. (2006), "Testing the Order of a Model," The Annals of Statistics, 34, 1166-1203.*

*Dacunha-Castelle, D., and Gassiat , E. (1997), "Testing in Locally Conic Models, and Application to Mixture Models," Probability and Statistics, 1, 285-317.*

*Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least Angle Regression," The Annals of Statistics, 32, 407-499.*

*Fan, J., and Li, R. (2001), "Variable Selection Via Nonconcave Penalized Likelihood and Its Oracle Properties," Journal of the American Statistical Association, 96, 1348-1360.*

*Feng, Z. D., and McCulloch, C. E. (1996), "Using Bootstrap Likelihood Ratios in Finite Mixture Models," Journal of the Royal Statistical Society. Series B (Methodological), 58, 609-617.*

*Foster, D. P., and George, E. I. (1994), "The Risk Inflation Criterion for Multiple Regression," The Annals of Statistics, 22, 1947-1975.*

*George, E. I. (2000), "The Variable Selection Problem," Journal of the American Statistical Association, 95, 1304-1308.*

*Hannan, E. J., and Quinn, B. G. (1979), "The Determination of the Order of an Autoregression," Journal of the Royal Statistical Society. Series B (Methodological), 41, 190-195.*

*Hathaway, R. J. (1985), "A Constrained Formulation of Maximum-Likelihood Estimation for Normal Mixture Distributions," The Annals of Statistics, 13, 795-800.*

*Haughton, D. M. A. (1988), "On the Choice of a Model to Fit Data from an Exponential Family," The Annals of Statistics, 16, 342-355.*

*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, pp. 221-233.*

*Kass, R. E., and Raftery, A. E. (1995), "Bayes Factors," Journal of the American Statistical Association, 90, 773-795.*

*Kass, R. E., and Wasserman, L. (1995), "A Reference Bayesian Test for Nested Hypotheses and Its Relationship to the Schwarz Criterion," Journal of the American Statistical Association, 90, 928-934.*

*Keribin, C. (2000), "Consistent Estimation of the Order of Mixture Models," Sankhyā Series A 62, 49-66.*

*Kullback, S., and Leibler, R. A. (1951), "On Information and Sufficiency," The Annals of Mathematical Statistics, 22, 79-86.*

*Leroux, B. G. (1992), "Consistent Estimation of a Mixing Distribution," The Annals of Statistics, 20, 1350-1360.*

*Lindley, D. V. (1980), "Approximate Bayesian Methods," in Bayesian Statistics, eds. J. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, Valencia: Valencia University Press, pp. 223-237.*

*Miller, A. (2002), Subset Selection in Regression (2nd ed.), Chapman & Hall/CRC.*

*Nishii, R. (1984), "Asymptotic Properties of Criteria for Selection of Variables in Multiple Regression," Annals of Statistics, 12, 758-765.*

*Rao, C. R. (1973), Linear Statistical Inference and Its Applications (2 ed.), Wiley.*

*Rao, C. R., and Wu, Y. H. (1989), "A Strongly Consistent Procedure for Model Selection in a Regression Problem," Biometrika, 76, 369-374.*

*Rao, J. S., and Tibshirani, R. (1997), "Discussion To "An Asymptotic Theory for Model Selection" By Jun Shao," Statistica Sinica, 7, 249-252.*

*Schwarz, G. (1978), "Estimating the Dimension of a Model," The Annals of Statistics, 6, 461-464.*

*Shao, J. (1997), "An Asymptotic Theory for Linear Model Selection," Statistica Sinica, 7, 221-264.*

*Shibata, R. (1976), "Selection of the Order of an Autoregressive Model by Akaike's Information Criterion," Biometrika, 63, 117-126.*

*Tibshirani, R. (1996), "Regression Shrinkage and Selection Via the Lasso," Journal of the Royal Statistical Society. Series B (Methodological), 58, 267-288.*

*Tierney, L., and Kadane, J. B. (1986), "Accurate Approximations for Posterior Moments and Marginal Densities," Journal of the American Statistical Association, 81, 82-86.*

*van der Vaart, A. W. (1998), Asymptotic Statistics, Cambridge University Press.*

*Yang, Y. (2005), "Can the Strengths of Aic and Bic Be Shared? A Conflict between Model Indentification and Regression Estimation," Biometrika, 92, 937-950.*

*Yang, Y. (2007), "Prediction/Estimation with Simple Linear Models: Is It Really That Simple?," Econometric Theory, 23, 1-36.*

*Zhang, P. (1992), "On the Distributional Properties of Model Selection Criteria," Journal of the American Statistical Association, 87, 732-737.*

*Table* 1. *Table 1 Caption Here.*

| N | r(AIC)=log2/loglogN | r(BIC)=logN | logN^1/2 |
|---|---|---|---|
| 2 | -1.891 | 0.693 | 0.833 |
| 3 | 7.370 | 1.099 | 1.048 |
| 5 | 1.457 | 1.609 | 1.269 |
| 10 | 0.831 | 2.303 | 1.517 |
| 50 | 0.508 | 3.912 | 1.978 |
| 100 | 0.454 | 4.605 | 2.146 |
| 500 | 0.379 | 6.215 | 2.493 |
| 1000 | 0.359 | 6.908 | 2.628 |
| 10000 | 0.312 | 9.210 | 3.035 |
| 100000 | 0.284 | 11.513 | 3.393 |
| 1000000 | 0.264 | 13.816 | 3.717 |
| 10000000 | 0.249 | 16.118 | 4.015 |
| 1E+08 | 0.238 | 18.421 | 4.292 |
| 1E+09 | 0.229 | 20.723 | 4.552 |
| 1E+10 | 0.221 | 23.026 | 4.799 |
| 1.00E+15 | 0.196 | 34.539 | 5.877 |
| 1.00E+20 | 0.181 | 46.052 | 6.786 |