On an Expression of Generalized Information Criterion

Zeng-Hua Lu

School of Mathematics and Statistics
University of South Australia

The theme in this talk

We propose a Generalized Information Criterion for model Selection – the Bayesian-type Generalized Information Criterion

$$BGIC = l(\theta) - \frac{d}{2}(\log N)^r$$

where

$l(.)$ is the model log-likelihood function

d is the number of free parameters in the model

N is the sample size

$0 < r < \infty$.

An outline of the talk

- Introduction
- Justifications from the Bayes Factor point of view
- The consistency property of BGIC
- Simulation studies
- Outline of the proof of consistency
- Concluding remarks

Introduction

- Criterion based approach for model selection problem remains a popular choice due to the simplicity of its applications for the problem
- There have been many criteria proposed in the literature such as AIC and BIC
- Most criteria can be written in the form of

$$l_N(\theta) - 0.5 d\lambda$$

- 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

Introduction (continued)

- However, there is a general consent that the BIC enjoys the consistency property of correctly identifying the true model with probability 1 as $N \to \infty$
- In fact Rao and Wu (1989) showed that its generalized version, the Generalized Information Criterion (GIC), where λ satisfies $N^{-1}\lambda \to 0$ and $(\log\log N)^{-1}\lambda \to \infty$ possesses the consistency property in linear regression models.
- Such a property is not shared by the AIC or any criteria in which λ is constant (Foster and George 1994, Yang 2005, 2007).

Introduction (continued)

- However, it has been found that the BIC often is too liberal in selecting a model (underfitting), while the AIC on the other hand is too conservative (overriding)
- We propose what we call the Bayesian-type Generalized Information Criterion

$$BGIC = l(\theta) - 0.5d(\log N)^r$$

- When *r* = log2/loglogN, BGIC becomes AIC. When *r* = 1, BGIC becomes BIC. When *r* > 1, BGIC becomes more aggressive than BIC by introducing a heavier penalty

Introduction (continued)

- We then use the fivefold cross validation technique for choosing r value in the penalty term
- Our simulation studies suggest that our procedure performs well
- The consistency property is established in the context of the Maximum likelihood Estimation
- Our consistency results also apply to situations where there exist certain types of nonidentifiability problem such as in mixture models

Viewpoint of the Bayes factor

Suppose there are finite q models, defined by m_i, $i = 1, ..., q$, whose pdf for observed data Y is $f(Y;\theta_i)$, where θ_i is a vector of unknown model parameters. Let $\pi(m_i)$ and $\pi(\theta_i)$ be priors of m_i and θ_i. By Bayes' theorem, the posterior probability of m_i is obtained as,

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

Where $P(Y;m_i) = \int f(Y;\theta_i)\pi(\theta_i)d\theta_i$.

Viewpoint of the Bayes factor (continued)

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

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

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

Then,

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

Viewpoint of the Bayes factor (continued)

Therefore, from the Bayesian point of view one chooses model i over *j* if $B_{ij} > 1$; model j over *i* if $B_{ij} < 1$. Computing BF in involves evaluation of integrals, hence is usually computationally demanding. The Laplace approximation to the integral $P(Y;m_i)$ has been suggested in the literature. For the iid sample $Y = (y_1, ... y_N)'$, with the density $f(y_n;\theta_i)$, $n = 1, ..., N$, the approximation is

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

where $L(\theta_i) = \prod_{n=1}^{N} f(y_n;\theta_i)$ and

$\Sigma(\theta_i) = [-\partial^2 \log f(y;\theta_i)/(\partial\theta_i\partial\theta_i')]^{-1}$.

Viewpoint of the Bayes factor (continued)

The BIC can be viewed as taking logarithm of the leading terms as

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

where $l_N(\theta_i) = \sum_{n=1}^{N} \log f(y_n;\theta_i)$. The BIC approximation is accurate to the order of $O(N^{-1})$. If one wishes to take into account the higher order terms, one might consider using $(\log N)^r$ to replace $\log N$ in the BIC approximation. It therefore leads our BGIC,

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

The restriction $0 < r < \infty$ is for the consistency property.

An example

The following example from the literature demonstrates that an exact posterior probability function can be expressed by the BGIC.

Example: (Kass and Wasserman 1995)

Let $y_n \sim N(\psi, 1)$ and consider the normal unit-information prior $\psi \sim N(0,1)$, then the exact posterior density function is

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

where

$$l_0 = -\frac{N}{2} \log \frac{2\pi}{N} - \frac{N \sum_{n=1}^{N} y_n{}^2}{2},$$

and $\bar{y} = \sum_{n=1}^{N} y_n / N$.

An example (continued)

However, the BGIC is obtained as

$$BGIC = l_0 + \frac{N\overline{y}^2}{2} - \frac{1}{2}(\log N)^r \ ,$$

which suggests

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

Therefore, in this case, 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 $N \geq 3$.

Consistency

Let the true distribution function be $G(y)$ and density function be $g(y)$. Suppose there are a finite set of q_k true models among candidate models, $M_g = \{m_k, k = 1, \ldots, q_k\}$, with the density function $g_k(y; \theta_k)$. Suppose that there are a finite set of $q_{\tilde{k}}$ non-true models among candidate models, $M_f = \{m_{\tilde{k}}, \tilde{k} = 1, \ldots, q_{\tilde{k}}\}$, with the density functions $f_{\tilde{k}}(y; \theta_{\tilde{k}})$. 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.

Regularity conditions

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

(C1) Θ_i , $\forall i$, are compact.

(C2) g_k and $f_{\tilde{k}}$ are dominated for $\forall k$ and $\forall \tilde{k}$, i.e. $\left|g_k\right| \leq b_1(y)$ and $\left|f_{\tilde{k}}\right| \leq b_2(y)$, 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).

Conditions for the penalty

Rewrite the model selection criterion as $l_N(\theta_i) - p_{d_i,N}$, where $p_{d_i,N}$ represents the penalty term, is maximized . The following conditions are assumed for $p_{d_i,N}$.

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

Consistency results

Theorem 1. Under (C1) – (C3) and (P1) – (P3),

(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 $\Pr(\hat{m}_i = m_k) = 1$ a.s., $i = 1, ..., q_{\tilde{k}} + 1$.

(ii) If the set of candidate models is M_f (i.e. it does not include a true model), then $\Pr(\hat{m}_i = m^*) = 1$ a.s., $i = 1, ..., q_{\tilde{k}}$, 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 $\Pr(\hat{m}_i = m_k^*) = 1$ a.s., $i = 1, ..., q_{\tilde{k}} + q_k$, where m_k^* is the true model with the smallest dimension d_k^* among the models in M_g.

Remarks for the consistency results

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. $\lambda = O(N / \log N)$ 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 $r > 0$.

Reporting the range of values

Suppose that one wish to compare two models and $l_N(\hat{f}_{i_1}) > l_N(\hat{f}_{i_2})$, $N \geq 3$ 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.
- 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).

Model search: linear regression

If the primary goal to select a model among a number of candidate models, one could use the cross-validation (CV) technique to estimate *r*. For variable selection in linear regression models, we adopt the fivefold CV procedure as follows (Breiman 1995). Denote the training and test set as $N - N^v$ and N^v, respectively, for $v = 1, ..., 5$. For each *r* and *v*, we find all sub set estimator $\hat{\beta}^{(v)}(r)$ according to a criterion based on the training set $N - N^v$. Define the CV criterion as

$$CV(r) = \sum_{v=1}^{5} \sum_{(y_{n_v}, x_{n_v}) \in N^v} \{y_{n_v} - x'_{n_v} \hat{\beta}^{(v)}(r)\}^2 .$$

We find an $\hat{r}$ that minimizes $CV(r)$.

Model search: mixture models

In the context of order selection in mixture models, in stead of finding *r* that minimizes the prediction error we suggest to find *r* that maximizes the prediction likelihood, i.e.,

$$CV(r) = \sum_{v=1}^{5} \sum_{(y_{n_v}, x_{n_v}) \in N^v} \log f_K(y_{n_v}; x_{n_v}, \hat{\theta}^{(v)}(r)),$$

where $f_K(.)$ is *K*-component mixture density

function.

Simulation studies: Linear regression

Example : Linear regression. Consider the model,

$$y_n = \mathbf{x}_n' \beta + \sigma \varepsilon_n ,$$

where $\mathbf{x}_n = (x_{n1}, ..., x_{nD})'$ and $\{\varepsilon_n\}$ are independent and identically distributed as $N(0,1)$.

Our first simulation design follows that of Gunst and Mason (1980), Shao (1993, 1997, JASA). $D = 5$, $N = 40$, $x_{n1} = 1$, $\forall n$ and observations of four covariates are taken from an example in Gunst and Mason (1980), which was reproduced in Table 1 of Shao (1993).

Table 1. Simulation Results for the Linear Regression Model based on the First Simulation Design

True Model	Method	Avg. No. of 0 Coefs		CM(%)	MRME(%)
		Correct	Incorrect		
$\beta = (2,0,0,4,0)'$	AIC	2.435	0	55.4	75.97
	BIC	2.779	0	80.3	55.43
	RTC	2.998	0	99.8	36.84
	BGIC ($\hat{r} = 2.15$)	3	0	100	36.69
$\beta = (2,0,0,4,8)'$	AIC	1.611	0	66.0	86.25
	BIC	1.846	0	85.8	73.51
	RTC	1.999	0	99.9	64.24
	BGIC ($\hat{r} = 1.79$)	1.999	0	99.9	64.24
$\beta = (2,9,0,4,8)'$	AIC	0.817	0	81.7	95.24
	BIC	0.939	0	93.9	91.29
	RTC	0.996	0.014	98.4	88.43
	BGIC ($\hat{r} = 1.01$)	0.94	0	94.0	91.21
$\beta = (2,9,6,4,8)'$	AIC	0	0	100	100
	BIC	0	0	100	100
	RTC	0	0.051	94.9	100
	BGIC ($\hat{r} = 1.01$)	0	0	100	100
$\beta = (1,2,3,2,3)'$	AIC	0	0.615	39.1	100
	BIC	0	0.833	19.5	118.99
	RTC	0	1.488	0.0	229.11
	BGIC ($\hat{r} = 0.76$)	0	0.707	30.6	110.96

Correct: the average number of times, in which irrelevant covariates are correctly excluded from the model.
Incorrect: the average number of times, in which relevant covariates are incorrectly excluded from the model.
CM: select the correct model.
MRME: the median of relative model errors defined as the ratio of the model errors of the selected model over that of the full model, which includes all covariates

Simulation studies: Linear regression

Example : Linear regression (continued).

The second simulation design follows that of Tibshirani (1996) and Fan and Li (2001). In this design, there is no intercept in the model. The values of eight covariates ($D = 8$) are generated from the multivariate normal distribution with each covariate following $N(0,1)$ and the correlation between covariates x_a and x_b is $0.5^{|a-b|}$, $a, b \in \{q\}$. The true coefficient is $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)'$.

Table *2. Simulation Results for the Linear Regression Model based on the Second Simulation Design*

Method	Avg. No. of 0 Coefs		CM(%)	MRME(%)
	Correct	Incorrect		
		$N = 40,\ \sigma = 3$		
AIC	3.923	0.264	30.2	82.94
BIC	4.499	0.399	46.2	73.74
RTC	4.906	1.207	11.3	171.84
BGIC ($\hat{r} = 0.85$)	4.366	0.339	44.3	74.59
		$N = 40,\ \sigma = 1$		
AIC	3.987	0	36.9	77.16
BIC	4.589	0	67.8	55.33
RTC	4.993	0.001	99.3	34.16
BGIC ($\hat{r} = 1.57$)	4.955	0	95.6	36.55
		$N = 60,\ \sigma = 1$		
AIC	4.075	0	39.3	75.42
BIC	4.755	0	78.5	48.36
RTC	4.999	0	99.9	36.22
BGIC ($\hat{r} = 1.57$)	4.985	0	98.5	36.99
		$N = 100,\ \sigma = 1$		
AIC	4.128	0	40.7	73.75
BIC	4.829	0	84.4	45.24
RTC	5	0	100	35.67
BGIC ($\hat{r} = 1.57$)	4.992	0	99.2	36.36
		$N = 200,\ \sigma = 1$		
AIC	4.161	0	41	74.08
BIC	4.882	0	88.8	42.56
RTC	5	0	100	35.39
BGIC ($\hat{r} = 1.57$)	5	0	100	35.39

Simulation studies: Linear regression

Example : Linear regression (continued).

The third simulation design is to study the performance of the criteria when there is non true model presented among candidate models. We particularly consider the polynomial approximation to a nonlinear function (Shao 1997, Statistica Sinica). The values of y are generated according to $y_n = \exp(x_n) + \varepsilon_n$. where x_n are sampled from $N(0,1)$. We select a model from the class of linear models with $\mathbf{x}_n = (1, x_n, ..., x_n^{h-1})'$, $h = 5$.

Table 3. Probability of Selecting Polynomial Functions for Approximating an Exponential Function by Different Criteria

	$\hat{h}$			
	2	3	4	5
		N = 40		
AIC	0	0	0.436	0.564
BIC	0	0.003	0.604	0.393
RTC	0	0.12	0.829	0.051
BGIC ($\hat{\lambda} = 0.01$)	0	0	0.271	0.729
		N = 60		
AIC	0	0	0.089	0.911
BIC	0	0	0.225	0.775
RTC	0	0	0.842	0.158
BGIC ($\hat{\lambda} = 0.01$)	0	0	0.032	0.968
		N = 100		
AIC	0	0	0.035	0.965
BIC	0	0	0.141	0.859
RTC	0	0	0.949	0.051
BGIC ($\hat{\lambda} = 0.01$)	0	0	0.015	0.985
		N = 200		
AIC	0	0	0.002	0.998
BIC	0	0	0.021	0.979
RTC	0	0	0.985	0.015
BGIC ($\hat{\lambda} = 0.01$)	0	0	0.002	0.998
		N = 300		
AIC	0	0	0	1
BIC	0	0	0.001	0.999
RTC	0	0	0.933	0.067
BGIC ($\hat{\lambda} = 0.01$)	0	0	0	1

Simulation studies: Mixture models

Example (Mixture normal linear regression). Consider the finite *K*-component mixture model

$y_n \sim \sum_{i=1}^{K} \alpha_i \phi_i(y_n; \mathbf{x}'_n \beta_i, \sigma_i^2)$, where $\phi_i(.)$ is the normal density function and the probability $\alpha_i \geq 0$, $i = 1, \ldots, K$ and $\sum_{i=1}^{K} \alpha_i = 1$.

The first simulation design: *K* = 2, and

$y_n = 1 + x_{1n} + e_{1n}$, with probability 0.5,

$y_n = 2 - x_{2n} + e_{2n}$, with probability 0.5,

where e_{1n}, $e_{2n} \sim i.i.dN(0,1)$.

Other Criteria studied

Biernacki and Govaert (1997),

$$\mathrm{CLC} = l_N(\hat{\theta}) + \sum_{k=1}^{K} \sum_{n=1}^{N} \hat{\tau}_{kn} \ln(\hat{\tau}_{kn}) ,$$

Biernacki, Celeux and Govaert (2000)

$$\mathrm{ICL} = \mathrm{CLC} - 0.5\dot{q} \ln N ,$$

Naik, Shi and Tsai (2007).

$$\mathrm{MRC} = \sum_{k=1}^{K} \hat{m}_k \ln(\pi^2 / 3) + \sum_{k=1}^{K} \frac{\hat{m}_k(\hat{m}_k + \hat{p}_k)}{\hat{m}_k - \hat{p}_k - 2} - 2\sum_{k=1}^{K} \hat{m}_k \ln(\hat{\alpha}_k)$$

Table 4. Probability of Selecting number of mixture components

	$\hat{K}$			
	1	2	3	4
		N = 40		
AIC	0	0.25	0.45	0.3
BIC	0	0.7	0.2	0.1
CLC	0	0.1	0.25	0.65
ICL-BIC	0	0.6	0.3	0.1
MRC	0	0.25	0.2	0.55
RTC	0	0.9	0.1	0
BGIC ($\hat{\lambda} = 1.3$)	0	0.85	0.15	0

Outline of the proof of consistency

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 $G \log f_{\tilde{k}}(\theta_{\tilde{k}0})$.

Remars: we paprticularly show based on Feng and McCulloch's (1996) idea that the results apply even when ther exists the nonidentifiability problem, 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 $\mu_1 = \mu_2$, $\sigma_1^2 = \sigma_2^2$.

Proof of Theorem 1.

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 $f \neq g$ and $K(g, f) = 0$ if $f = g$ ..

Because

$$\frac{1}{N}\{l_N(g(\hat{\theta}_k)) - p_{d_k,N} - [l_N(f_{\tilde{k}}(\hat{\theta}_{\tilde{k}})) - p_{d_{\tilde{k}},N}]\}$$

$$= \frac{1}{N}\{l_N(g(\hat{\theta}_k)) - l_N(f_{\tilde{k}}(\hat{\theta}_{\tilde{k}}))\} - \frac{p_{d_k,N}}{N} + \frac{p_{d_{\tilde{k}},N}}{N}$$

$$\xrightarrow{a.s} K(g, f_{\tilde{k}}) > 0,$$

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

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

Proof of Theorem 1 (continued)

(ii) Define

$$D(f_{\tilde{k}_1}, f_{\tilde{k}_2}) = K(g, f_{\tilde{k}_1}) - K(g, f_{\tilde{k}_2})$$

$$= \int \log f_{\tilde{k}_2} dG(y) - \int \log f_{\tilde{k}_1} dG(y).$$

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, 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.,

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

Proof of Theorem 1 (continued)

(iii) The proof is more involved. The key results are that, 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 $\{l_N(\hat{g}_{k_1}) - l_N(\hat{g}_{k_2})\} / p_{d_{k_2},N} \xrightarrow{a.s.} 0$. By condition (P1), $d_{k_1} < d_{k_2}$ results in $p_{d_{k_1},N} / p_{d_{k_2},N} - 1 < 0$. Therefore,

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

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

$$\Pr(l(\hat{g}_k^*) - p_{d_k^*,N} > l(\hat{g}_k) - p_{d_k,N}) = 1 \text{ a.s.}$$

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

Concluding remarks

- we propose a particular form of the Generalized Information Criterion
- The consistency property of our criterion is studied
- The cross validation technique is suggested for estimating the penalty parameter
- Simulation studies suggest our proposed procedure works well, particularly in small sample size
- Our findings also provide some insight on understanding why the well known criteria such as AIC and BIC can fail to perform
- The drawback of our method is intensity of computation inherited from the cross validation technique
- Future studies are needed