Generalized Bayesian Information Criterion

Jae-Young Kim

Department of Economics

SUNY - Albany

April 1995

Abstract

Model Selection Criteria

Jae-Young Kim[1]
Department of Economics
State University of New York - Albany
and
Department of Economics
Hong Kong University of Science and Technology

Abstract

This paper reconsiders model selection criteria in the presence of nonstationarity. Following regular ways of deriving the model selection criteria we have found that the existing forms of AIC, C_p, and $\bar{R}^2$ do not change in the case of nonstationarity while it is not the case with the Bayesian rule. It implies that the other criteria than Bayesian give the same weight to a nonstationary component of the model as to stationary components while the Bayesian rule differentiates them. Certainly, this feature of the criteria other than Bayesian is not a desirable property for model selection in the presence of possible nonstationarity. Not only for this reason but also for the following reasons we recommend to use the Bayesian approach to model selection in the presence of nonstationarity. First, because of theoretical merits of Bayesian model selection rule found in this paper and in the literature. This paper has found that the exact, non-approximated, Bayesian rule leads to the choice of the model with minimum entropy distance to the true model when the choice set does not include the true model. It is also found that the non-approximated Bayesian rule leads to the choice of the true model with unit probability when the choice set includes the true model. Second, approximated Bayesian model selection rules are available that are simple to apply for many interesting cases of econometric work. We have found that, compared with other criteria, the approximated Bayesian rules have better asymptotic properties. Finally, Monte Carlo studies show that finite sample performance of the approximated Bayesian rules is satisfactory.

Keywords: Model selection, entropy distance, Bayesian rule.

[1]Correspondence: Department of Economics, State University of New York - Albany, Albany, NY 12222. e-mail: jykim@csc.albany.edu

1 Introduction

Model selection problem has been an important subject in econometrics as well as in many other science fields. Aside from theoretical *a priori* considerations in model building, the subject of how well the model fits the data is usually taken as the most important guide to model selection. Recently, researchers such as Sims (1988), Bai and Perron (1997) and Phillips (1996), Phillips and Ploberger (1996) noted that econometric model selection strategy needs to be reconsidered in the presence of nonstationarity. This paper is concerned with reexamination and redevelopment of model selection criteria in the presence of nonstationarity.

Following regular ways of deriving the model selection criteria we have found that generally the existing forms of AIC, C_p, and $\bar{R}^2$ do not change in the presence of nonstationarity while it is not the case with the Bayesian criterion. Certainly, this feature of the criteria other than Bayesian is not a desirable property for model selection. Nonstationary variables have different properties than stationary variables. Different weights have to be given to nonstationary variables from stationary variables for model selection. The Bayesian rule gives a correct weight to each component of the model while the other criteria do not, as is noted by Sims (1988), Phillips (1995), Phillips and Ploberger (1996), and Kim (1998).

Bayesian model selection rules are discussed by the above mentioned authors in the presence of possible nonstationarity. For theoretical and practical reasons, these authors study approximated forms of the Bayesian decision rule for model selection. Those approximated Bayesian decision rules are simple to apply for many interesting cases of econometric work. We have studied in this paper some properties of the approximated Bayesian rules relative to the other criteria. We have found that, compared with other criteria, the approximated Bayesian rules have better asymptotic properties. Monte Carlo studies show that finite sample performance of the approximated Bayesian rules is satisfactory.

In this paper, we also study the Bayesian decision rule in its non-approximated

form. The decision rule is interpreted as the one that selects the model with the highest posterior probability out of a set of alternatives. As is found in this paper, the exact non-approximated Bayesian rule leads to the choice of the model with minimum entropy distance to the true model when the choice set does not include the true model. It is also found that the non-approximated Bayesian rule leads to the choice of the true model with unit probability when the choice set includes the true model. Certainly, the non-approximated Bayesian rule has practical importance as well as theoretical merit. It has small sample justifications for application while most Bayesian criteria in approximated forms are justified for large sample applications.

Our discussion in this paper goes as follows. Section 2 reexamines each of the model selection criteria of interest. Section 2 also studies an exact, non-approximated, Bayesian model selection rule. Section 3 compares the model selection criteria in parsimony, power, consistency, etc. Section 4 performs Monte Carlo studies to examine finite sample performance of the criteria.

2 Model Selection Rules in General Frameworks

Let $(\Omega, \mathcal{F}, P)$ be a probability space and $\{\mathcal{F}\}_{t\geq 0}$ be an increasing family of sub σ-fields of $\mathcal{F}$. Let $\{y_t(\cdot)\}$ be a stochastic process defined on $(\Omega, \mathcal{F}, P)$ that is adapted to $\mathcal{F}_t$. Denote by $Y_n = \{y\}_{t=1}^n$, the n-segment of $\{y_t\}$. Assume that Y_n has a distribution function $P_n(\theta, Y_n)$ whose density is denoted by $p_n(\theta, Y_n)$ for $\theta \in I\!R^p$. Let $(\mathcal{V}, \mathcal{G}, P_\nu)$ be a probability space on which θ is defined.

A family $\mathcal{M}$ consists of candidate models for Y_n in the presence of uncertainty of the true model. A model $m_i \in \mathcal{M}$ is associated with a parameter space Θ^i of dimension p_i for $i \in \mathcal{I}$ where $\mathcal{I} = \{1, ..., I\}$. Assume that for each m_i a family $Q_n^i(\theta^i, Y_n)$ of distribution functions, with a density $q_n^i(\theta^i, Y_n)$, is defined on the measurable space $(\mathcal{V}, \mathcal{G}) \times (\Omega, \mathcal{F})$.

We assume some regular conditions on the density $q_n^i(\theta^i, Y_n)$ defined on $(\mathcal{V}, \mathcal{G}) \times (\Omega, \mathcal{F})$: (a) For each n, $\int_\Omega \log(p_n) dP_n$ exists and $|\log q_n^i(\theta^i, Y_n)| \leq k^i(Y_n)$ for all

$\theta^i \in \Theta^i$ for $i \in \mathcal{I}$, where k is integrable with respect to P_n. (b) For each n, and m_i $\int_\Omega \log(p_n/q_n^i)dP_n$ has a unique minimum at $\theta_{n,0}^i \in \Theta^i$. (c) For each n, $\int_\mathcal{Y} \log(p_n)dP_n$ exists and $|\log q_n^i(\theta^i, Y_n)| \leq \kappa^i(\theta)$ a.e. in Ω, where κ^i is integrable with respect to $P_n(\theta)$. (d) For each n $\int_\mathcal{Y} \log(p_n/q_n^i)dP_n$ has a minimum at $i = i^*(n)$ for an $i^*(n) \in \mathcal{I}$.

2.1 Bayesian Information Criterion

2.1.1 A General Framework

A natural approach to model selection in the Bayesian framework is to choose a model m_i for which the posterior probability is the largest. Thus, let $\Pr(m_i|Y_n)$ be the posterior probability that m_i is true. By the Bayes' rule

$$\Pr(m_i|Y_n) = \frac{q_n(Y_n|m_i)\Pr(m_i)}{\sum_{j\in\mathcal{I}} q_n(Y_n|m_j)\Pr(m_j)} \tag{2.1}$$

where $\Pr(m_i)$ is the prior probability that m_i is true and $q_n(Y_n|m_j) = q_n^i(Y_n)$. But, notice that

$$q_n(Y_n|m_j) = \int q_n(Y_n|\theta^j, m_j)\varphi(\theta^j|m_j)d\theta^i = E_j[q_n(Y_n|\theta^j)], \tag{2.2}$$

where $\varphi(\theta^j|m_j)$ is the prior density associated with the model m_j. If we further assume that $\Pr(m_j)$ is the same for all j, the model selection rule is to choose m_i for which $E_i[q_n(Y_n|\theta^i)]$ is the largest. Phillips (1996) provides another dimension of justification of Bayesian approach for model selection based on the notion of a Bayesian model measure.

We can show that the above criterion (2.2) can be also justified by the notion of Kullback-Leibler information. Let $\pi_n(\theta|Y_n)$ be the true posterior density of θ from the true likelihood $p_n(Y_n|\theta)$ and a prior $\pi(\theta)$. Also, let $\varphi_n^i(\theta^i|Y_n)$ be the posterior density of θ^i from the likelihood q_n^i and a prior $\varphi^i(\theta^i)$. The Kullback-Leibler distance between the posterior densities π_n and φ_n^i is defined as

$$K\left(\pi_n, \varphi_n^i\right) = \int \ln\left(\pi_n/\varphi_n^i\right)\pi_n d\theta.$$

We can show that BIC virtually chooses the model that has minimum Kullback-Leibler distance from the true model under some condition.

Theorem 2.1 *Let* $\varphi_n^{i^*} = \text{argmin}_{i\in\mathcal{I}} K\left(\pi_n, \phi_n^i\right)$ *and* $q_n^{i^*}(\mathrm{y}_n)\varphi^{i^*} = \text{argmax}_{i\in\mathcal{I}} E_i[q_n^i(Y_n|\theta^i)]$. *Assume that for* $i, j \in \mathcal{I}$ *such that*

$$K\left(\pi_n, \varphi_n^i\right) - K\left(\pi_n, \varphi_n^j\right) \geq 0,$$

$$\begin{aligned} (2.3) \qquad & \int \ln\left((p_n\pi)/(q_n^i\varphi^i)\right)\pi_n d\theta - \int \ln\left((p_n\pi)/(q_n^j\varphi^j)\right)\pi_n d\theta \\ & \geq K\left(\pi_n, \varphi_n^i\right) - K\left(\pi_n, \varphi_n^j\right) \end{aligned}$$

with equality only if $i = j$. *Then,* $\varphi_n^{i^*}$ *is such that*

$$\varphi_n^{i^*}(\theta_i|Y_n) = \frac{q_n^{i^*}(Y_n|\theta^i)\varphi^{i^*}(\theta^i)}{\int q_n^{i^*}(Y_n|\theta^i)\varphi^{i^*}(\theta^i)d\theta^i}.$$

Therefore, BIC selects the model with the shortest Kullback-Leibler distance from the true model out of a set of alternatives under the condition (2.3). The condition (2.3) generally holds in some probabilistic sense for a sufficiently large sample.[2]

Theorem 2.2 *Assume that for* $i \in \mathcal{I}$

$$(2.4) \qquad \int \ln\left(/(p_n\pi)/(q_n^i\varphi^i)\right)\pi_n d\theta \geq K\left(\pi_n, \varphi_n^i\right)$$

with equality only if $\varphi_n^i = \pi_n$. *Then, for any* $i \in \mathcal{I}$,

$$E_i[q_n^i(Y_n|\theta^i)] \leq E_\theta[p_n(Y_n|\theta)],$$

with equality only if $q_n^i\varphi^i = p_n\pi$. *That is,* $\Pr[m_i|Y_n] \leq \Pr[\text{true model}|Y_n]$ *for all* $i \in \mathcal{I}$ *with equality only if* m_i *is the true model.*

An important implication of the above theorem is that BIC chooses the true model under the condition (2.4) if the true model is contained in the choice set. The condition (2.4) generally holds in some probabilistic sense for a sufficiently large sample by the same reason as for the condition (2.3).

[2]Under some regularity conditions we can show that $\log(\varphi_n), \log(\pi_n) = O_p(\log(n))$ while $\log(q_n), \log(p_n) = O_p(n)$, so that $(\varphi_n^i/(q_n^i\varphi^i)), (\pi_n/(p_n\pi)) \to 0$. For example, for a normal model with i.i.d. observations $\log(\pi_n) = O_p(\log(|-L_n''(\hat{\theta})|)) = O_p(\log(n))$ where $L_n''(\theta) = \partial^2 \log(p_n(\theta))/\partial\theta\partial\theta'$ while $\log(p_n) = O_p(-\log\sigma^T) = O_p(T)$ where σ^2 is the variance of the variable.

2.1.2 Approximations

The criterion (2.2) involves computation of an integral of $q_n \times \varphi$ with respect to θ^i in $\mathbb{R}^{p_i}$. Certainly this computation is not an easy task even with a very fast computer. Also, the choice of the range of θ is another problem for the computation. Chib (1995), among others, applies the Gibbs sampling method to compute the marginal likelihood $q_n(Y_n)$. The Gibbs sampling method is a powerful approach to computing a density that can be written as a product of several conditional densities. Sometimes, however, the result from the Gibbs sampler is sensitive to the setup of the simulation or 'sampling'. Also, it is necessary that all integrating constants of the full conditional distributions in the Gibbs sampler be known (p.1314, Gibbs (1995)). On the other hand, the marginal likelihood $q_n(Y_n)$ itself depends on the prior density, so that model selection based on a direct computation of $q_n(Y_n)$ yields different results depending on the choice of the prior.

In the following we provide an approximation to the integral in (2.2) that is valid for a sufficiently large sample. It is computationally simple to handle and yet has sound theoretical justification. Also, the approximation provides a convenient method for the decision making that can be used before applying a more sophisticated method such as the Gibbs sampler.

We study approximation of $E_\theta[p_n(Y_n|\theta)]$ for the true model in the following. The same analysis applies to $E_j[q_n^j(Y_n|\theta^j)]$ for $j \in \mathcal{I}$. Let $N(\hat{\theta}_n, \delta_n)$, $n = 1, ..., \infty$, be such that

$$N(\hat{\theta}_n, \delta_n) = \{\theta : |\theta_1 - \hat{\theta}_{n1}|^2/\delta_{n1}^2 + \cdots + |\theta_k - \hat{\theta}_{nk}|^2/\delta_{nk}^2 < 1\}$$

where $\hat{\theta}_{ni}$ is the i^{th} element of $\hat{\theta}_n$, the MLE of θ based on Y_n; $\delta_n = (\delta_{n1}, ..., \delta_{nk})'$ is a k-vector of real numbers; $|\cdot|$ denotes the usual Euclidean norm. We consider a sequence $\{\delta_n\}$ such that δ_n becomes smaller and smaller as $n \nearrow \infty$, so that $N(\hat{\theta}_n, \delta_n)$ shrinks as n gets larger. Also, δ_n may depend on $\omega \in \Omega$.

Assume that the log-likelihood $L_n(\theta) = \log p_n(\theta)$ is twice differentiable with respect to θ in $\bigcup_{n=1}^{\infty} N(\hat{\theta}_n, \delta_n)$. Denote by $L_n''(\theta)$ the second derivative of the log-likelihood. Also, denote by $\|\cdot\|$ the matrix norm: For an $m \times m$ matrix A,

$\|A\| = \sup |Ax|/|x|$, where $|Ax|$ is the usual Euclidean norm on $\mathbb{R}^m$.

Now, consider the following conditions (C1) and (C2).

(C1)(a) *Let* $M_n(\hat{\theta}_n(\omega), \delta_n) = \sup_{\theta \in N(\hat{\theta}_n, \delta_n)} \|[L_n''(\hat{\theta}_n)]^{-1}[L_n''(\theta) - L_n''(\hat{\theta}_n)]\|$. *There exists a positive sequence* $\{\delta_n\}_{n=1}^{\infty}$ *such that* $\lim_{n \nearrow \infty} P[M_n(\hat{\theta}_n(\omega), \delta_n) < \epsilon] = 1$ *for each* $\epsilon > 0$.

(b) *Let* $\Sigma_n = [-L_n''(\hat{\theta}_n)]^{-1}$. *For* δ_n *satisfying (C1)(a) the absolute value of each element of the vector* $\Sigma_n^{-1/2}\delta_n$ *tends to infinity as* $n \nearrow \infty$ *in* P*-probability.*

(C2) For δ_n satisfying (C1),

$$\int_{\Theta \setminus N(\hat{\theta}_n, \delta_n)} \pi_n(\theta|Y_n)d\theta \longrightarrow 0$$

as $n \nearrow \infty$ in P-probability, i.e., θ concentrates in $N(\hat{\theta}_n, \delta_n)$ as $n \nearrow \infty$.

Conditions (C1) and (C2) cover a very wide variety of models containing nonstationary components. Kim (1998) showed that a posterior formed from a likelihood satisfying (C1) and (C2) above is asymptotically normal with the first moment of the distribution equal to $\hat{\theta}_T$ and the second moment $[-L_T''(\hat{\theta}_T)]^{-1}$.

Under the above conditions, we derive the following result:

Lemma 2.1 *Assume that the prior* $\pi(\theta)$ *is continuous in* Θ *and bounded at* θ_0. *Then, under the assumptions (C1) and (C2)*

$$\begin{aligned} &\log E_j[q_n^j(Y_n|\theta^j)] \\ &= \log(p_n(Y_n|\hat{\theta}_n)) + (1/2)\log(|\Sigma_n|) + (k/2)\log(2\pi) + \log(\pi(\theta_0)) + R_0, \end{aligned}$$

where R_0 *is of* $o_p(1)$.

From Lemma 2.1, a general form of Bayesian information criterion is

$$\begin{aligned} &(GBIC(I)) \quad \textit{choose the model } j \textit{ that maximizes} \\ &\log(p_n(Y_n|\hat{\theta}_n)) + (1/2)\log(|\Sigma_n|) + (k/2)\log(2\pi). \end{aligned} \tag{2.5}$$

Notice that the above criterion obtained from an approxiamtion of $E_i[p_n(Y_n|\theta)]$ is the same as PIC in Phillips (1995,1996), Phillips and Ploberger (1994) that is obtained

based on a new notion of 'Bayes model measure'. Therefore, the approximated BIC in (2.5) is justified by a sound theoretical basis discussed in Phillips (1996).

We can derive an alternative form of Bayesian information criterion under some condition: Let $s_i(\cdot)$ for $i = 1, ..., k$ be a real valued function defined on $\mathbb{N} \equiv \{n : n = 1, ..., \infty\}$ and $\mathcal{D}(n) = diag(s_1(n)^{-1}, ..., s_k(n)^{-1})$.

Lemma 2.2 *Suppose that there exists a real-valued function s_i defined on $\mathbb{N}$ for each $i = 1, ..., k$ such that*

$$\mathcal{D}(n)L_n''(\theta_0)\mathcal{D}(n) = O_p(1). \tag{2.6}$$

Then,

$$\log(|\Sigma_n|) = -2\log\Big(\prod_{i=1}^k s_i(n)\Big) + R_1 \quad \text{for } R_1 = O_p(1).$$

$s_i(T)$ is the rate of convergence of $\hat{\theta}_i$, the i^{th} component of the MLE of θ $\hat{\theta}$. For example, $s_i(T) = T^{1/2}$ for a stationary component, $s_i(T) = T$ for a unit root component, $s_i(T) = T^{3/2}$ for a trend component, etc.

From Lemma 2.2, an alternative form of Bayesian information criterion (2.5) is

$$\begin{aligned} &(GBIC(II)) \quad \textit{choose the model } j \textit{ that maximizes} \\ &\log(p_n(Y_n|\hat{\theta}_n)) - \log\Big(\prod_{i=1}^k s_i(n)\Big) + (k/2)\log(2\pi). \end{aligned} \tag{2.7}$$

As an example, consider a linear regression model

$$y_t = \beta' x_t + \varepsilon_t, \tag{2.8}$$

where x_t is a vector of variables, and β is a vector of parameters. The vector $(y_t, x_t')'$ may contain nonstationary components. The disturbance ε_t is an identical and independent random variable with normal distribution $N(0, \sigma^2)$.

Denote $X = (x_1, ..., x_n)'$. For the model (2.8), the criteria (2.5) and (2.7) are as follows:

$$\begin{aligned} &(GBIC1) \quad \textit{choose the model } j \textit{ that minimizes} \\ &(n - k_j)\log(\hat{\sigma}_j^2)/T + \log(|X_j'X_j|)/n - k_j\log(2\pi)/n. \end{aligned} \tag{2.9}$$

(*GBIC2*) *choose the model j that minimizes*

$$(n-k_j)\log(\hat{\sigma}_j^2)/n + 2\log\left(\prod_{i=1}^{k_j} s_i(n)\right)/n - k_j\log(2\pi)/n. \tag{2.10}$$

Notice that Schwarz Bayesian information criterion, which selects a model j that minimizes

$$\log(\hat{\sigma}_j^2) + k_j\log(n)/n, \tag{2.11}$$

is a special case of our GBIC2: if $s_i(n) = n^{1/2}$ for all $i = 1, ..., k$, that is, all the components of x_t are stationary, then our GBIC2 reduces to Schwarz criterion (2.11). This implies that Schwarz criterion is a valid asymptotic Bayesian criterion when all components of x_t are stationary. On the other hand, Schwarz criterion is *not* a valid Bayesian criterion when x_t contains *nonstationary* component(s) since $s_i > n^{1/2}$ for the nonstationary component(s).

2.2 Akaike Information Criterion

Consider two competing models m_0 and m_1 where m_1 has a zero restriction on the parameter θ:

$$m_1 : \theta = (\theta_1', 0')',$$

where θ_1 is a k_1-vector of parameters and 0 is a $(k-k_1)$-vector of zeroes.

Akaike (1973)'s criterion is derived based on the following loss function

$$\ell(\theta_{1,0}, \hat{\theta}_1) = -\frac{2}{n}\int \log\frac{p_n(Y_n, \hat{\theta}_1)}{p_n(Y_n, \theta_{1,0})} p_n(Y_n, \theta_{1,0}) dY_n \tag{2.12}$$

where $\hat{\theta}_1$ is treated as a constant in the integration. Because ℓ depends on the unknown parameter, Akaike initially proposes using $\tilde{\ell}_1$ in place of ℓ:

$$\tilde{\ell}_1 = -\frac{2}{n}\log\frac{p_n(Y_n, \hat{\theta}_1)}{p_n(Y_n, \hat{\theta})}. \tag{2.13}$$

Notice that $|\tilde{\ell}_1 - \ell| \to 0$ as $n \to 0$ under H_0. Akaike suggests using

$$\tilde{\ell}_2 \equiv \tilde{\ell}_1 - \|\hat{\theta} - \theta_0\|^2 + 2\|\hat{\theta}_1 - \theta_{1,0}\|^2 \tag{2.14}$$

as an improvement over $\tilde{\ell}_1$ where

$$\|\hat{\theta}-\theta_0\|^2 = -(\hat{\theta}-\theta_0)' n^{-1} E\left[\frac{\partial^2 \log p_n}{\partial\theta\partial\theta'}\,|_{\theta_0}\right](\hat{\theta}-\theta_0). \tag{2.15}$$

Based on (1) Taylor expansion of $\ell(\theta_{1,0},\hat{\theta})$: $\ell(\theta_{1,0},\hat{\theta}) = \|\hat{\theta}_1-\theta_{1,0}\|^2 + O(T^{-3/2})$, (2) Taylor expansions of $\log p_n(Y_n,\theta_{1,0})$ and $\log p_n(Y_n,\theta_0)$ around $\hat{\theta}_1$) and $\hat{\theta}$), respectively, (3) under the assumption that H_0 is true, and (4) under the premise that $E\|\hat{\theta}-\theta_0\|^2 = k/n$ and $E\|\hat{\theta}_1-\theta_{1,0}\|^2 = k_1/n$, Akaike derived the criterion

$$AIC = -\frac{2}{n}\log p_n(Y_n,\hat{\theta}_1) + \frac{2k_1}{n}. \tag{2.16}$$

from the loss (2.14).

There are two crucial prerequisites for the premise in (4) above to be valid. First, the existence of

$$E\left[\frac{\partial^2 \log p_n}{\partial\theta\partial\theta'}\,|_{\theta_0}\right] \quad \text{and} \quad E\left[\frac{\partial^2 \log p_n}{\partial\theta_1\partial\theta_1'}\,|_{\theta_{1,0}}\right], \tag{2.17}$$

and second, asymptotic normality of $(\hat{\theta}-\theta_0)$ (or similar conditions) for the following result to hold,

$$E\|\hat{\theta}-\theta_0\|^2 = \frac{k}{n}; \quad E\|\hat{\theta}_1-\theta_{1,0}\|^2 = \frac{k_1}{n}. \tag{2.18}$$

In the presence of nonstationarity neither of (2.17) nor asymptotic normality of $(\hat{\theta}-\theta_0)$ is available. However, for a regression model

$$y_t = \theta' x_t + \varepsilon_t, \quad \varepsilon \sim N(0,\sigma^2), \tag{2.19}$$

if ε_t and x_t are independent, then we can show that (2.18) is true regardless of whether or not there exists a nonstationary variable in the model:

Lemma 2.3 *Assume that ε_t and x_t are independent in (2.19). Then, it is true that*

$$E\|\hat{\theta}-\theta_0\|^2 = \frac{k}{n}; \quad E\|\hat{\theta}_1-\theta_{1,0}\|^2 = \frac{k_1}{n}.$$

Therefore, if ε_t and x_t are independent, AIC (2.16) is a valid Akaike's criterion.

When ε_t and x_t are not independent, however, (2.18) is not true in general, and AIC (2.16) is not a valid Akaike criterion. For example, if the presence of nonstationary variables in x_t that is correlated with ε_t causes the distribution of $(\hat{\theta} - \theta_0)$ to have flatter tails than a normal distribution, the second term in AIC is not correct any more. In this case the exact quantity for the second term of AIC is

$$2E_*\|\hat{\theta}_1 - \theta_{1,0}\|_*^2 \equiv 2\int -(\hat{\theta}_1 - \theta_{1,0})' n^{-1} \left[\frac{\partial^2 \log p_n(\theta_{1,0})}{\partial\theta_1 \partial\theta_1'}\right] (\hat{\theta}_1 - \theta_{1,0}) p_n(\theta_{1,0}, Y_n) dY.$$

AIC is derived based on an implicit assumption that each model under study is the true model, as the loss (2.12) implies. Akaike (1973) implicitly assumes that the distance between θ_0 and the subspace defined by H_1 is of order $n^{-3/8}$, which is a condition for the validity of applying (2.12) for each model. Therefore, when the 'distance' between the true model and an assumed model is large, AIC will not provide a reliable result for model selection. For example, a model with possible nonstationary components with the dimension of the nonstationarity being unknown, decision making in favor of one dimension against another based on AIC will not be reliable. See the example in Section 4.2.

2.3 Other Model Selection Criteria

There are other criteria for model selection such as Theil's (1961) $\bar{R}^2$, Conditional Prediction Criterion C_p, Unconditional Prediction Criterion (UPC) and FIC of Wei (1992). Since unconditional mean squared error depends on the sample size in the presence of nonstationarity, a standard form of UPC will not be available. Thus, we do not consider it in this paper. Also, since FIC is basically in the category of Bayesian information criterion we do not investigate it further here.

2.3.1 Theil's $\bar{R}^2$ criterion

Since both $\bar{R}^2$ and C_p are adequate for regression models, we confine our discussion to a linear regression model:

$$y_t = \beta' x_t + \varepsilon_t, \quad \varepsilon \sim (0, \sigma^2). \tag{2.21}$$

Consider two competing models m_0 and m_1 where m_1 has a zero restriction on the parameter β:

$$m_1 : \beta = (\beta_1', 0')'$$

where β_1 is a k_1-vector of parameters and 0 is a $(k - k_1)$-vector of zeroes. Let b_j be the least square estimator of β in model j. Since $\bar{R}_j^2 = 1 - \frac{\sum_{t=1}^n \hat{\varepsilon}_{t,j}^2/(n-k_j)}{\sum_{t=1}^n (y_t - \bar{y}_n)^2/(n-1)}$ where $\hat{\varepsilon}_{t,j} = y_t - b_j' x_{t,j}$ and $\bar{y}_n = n^{-1} \sum_{t=1}^n y_t$,

$$\max_j \bar{R}_j^2 = \min_j \sum_{t=1}^n \hat{\varepsilon}_{t,j}^2/(n - k_j) \tag{2.22}$$

given $\{y_t\}$. We can interprete $\bar{R}^2$ as a special case of $\tilde{\ell}_1$ defined in (2.13) with

$$p_n(\theta) = \exp\left(- \sum_{t=1}^n (y_t - \beta' x_t)^2 \right) \tag{2.23}$$

or

$$p_n(\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2} \sum_{t=1}^n (y_t - \beta' x_t)^2 \right)$$

with $\sigma^2 =$ [constant] accross different models. This implies that $\bar{R}^2$ criterion can be interpreted as a special case of Akaike's criterion where the approximated loss function $\tilde{\ell}_1$ is used with p_n as in (2.23). Notice that the form of $\bar{R}^2$ does not change in the presence of nonstationarity.

2.3.2 Conditional Prediction criterion C_p

The conditional prediction criterion C_p is defined as

$$\rho(Xb_1, X\beta) = E[(Xb_1 - X\beta)'(Xb_1 - X\beta)]$$

where b_1 is the OLS estimator of the restricted parameter β_1. Under the assumption that the error term is independent of the regressor, C_p is derived as

$$(2.24) \qquad C_p = \rho(Xb_1, X\beta) = \frac{(n-k_1)\hat{\sigma}_1^2}{\hat{\sigma}^2} + (2k_1 - n) = \frac{(n-k_1)(1-\bar{R}_1^2)}{1-\bar{R}^2} + (2k_1 - n)$$

where $\hat{\sigma}_1^2 = \sum_{t=1}^n (y_t - x_t' b_1)^2/(n-k_1)$ and $\hat{\sigma}^2 = \sum_{t=1}^n (y_t - x_t' b)^2/(n-k)$. By the same way as in Lemma 2.3 we can show that (2.24) is true regardless of whether there exists a nonstationary component in (x_t', y_t) if the error term is independent of the regressors.

3 Comparisons

3.1 Parsimony

All the criteria considered above can be compared with a likelihood ratio (LR) test of one model against another. It facilitates the analysis of comparing parsimony and power of different criteria if we have the exact relationship between LR and each of the criteria.

For simplicity, but without loss of generality, we consider a linear regression (2.19). We are to choose between the model with k regressors X, denoted by m_0, and the model with its k_1 subset X_1, denoted by m_1. Then, we can easily establish the relationship between LR and each of the criteria studied above.

For BIC we have

$$(3.1) \qquad \begin{aligned} LR &= 2(\log q_n^0(\hat{\theta}) - \log q_n^1(\hat{\theta}_1)) \\ &= 2(GBIC(I)_0 - GBIC(I)_1) - \log(|\Sigma_n^0|/|\Sigma_n^1|) \\ &= 2(GBIC(II)_0 - GBIC(II)_1) + 2\log\left(\Pi_{i=k_1+1}^k s_i(T)\right). \end{aligned}$$

For AIC we have

$$(3.2) \qquad \begin{aligned} LR &= 2(\log q_n^0(\hat{\theta}) - \log q_n^1(\hat{\theta}_1)) \\ &= -n(AIC_0 - AIC_1) + 2n(E_*\|\hat{\theta} - \theta_0\|_*^2 - E_*\|\hat{\theta}_1 - \theta_{1,0}\|_*^2) \end{aligned}$$

where in the case of the error term and the resgressor being independent

$$LR = -n(AIC_0 - AIC_1) + 2(k - k_1).$$

To compare parsimony of BIC and AIC we compare Pr[choose m_1 over m_0] for BIC and AIC. For GBIC(I)

$$\begin{aligned} (3.3) \qquad & \Pr[\text{choose } m_1 \text{ over } m_0] = \Pr[GBIC(I)_0 < GBIC(I)_1] \\ & = \Pr[(\log q_n^0(\hat{\theta}) - \log q_n^1(\hat{\theta}_1)) \leq -\log(|\Sigma_n^0|/|\Sigma_n^1|)] \\ & = \Pr[LR \leq -2\log(|\Sigma_n^0|/|\Sigma_n^1|)], \end{aligned}$$

and for GBIC(II)

$$\begin{aligned} (3.4) \qquad & \Pr[GBIC(II)_0 < GBIC(II)_1] \\ & = \Pr[(\log q_n^0(\hat{\theta}) - \log q_n^1(\hat{\theta}_1)) \leq 2\log\left(\Pi_{i=k_1+1}^k s_i(T)\right)] \\ & = \Pr[LR \leq 4\log\left(\Pi_{i=k_1+1}^k s_i(n)\right)]. \end{aligned}$$

Likewise, for AIC

$$\begin{aligned} (3.5) \qquad & \Pr[\text{choose } m_1 \text{ over } m_0] = \Pr[nAIC_1 < nAIC_0] \\ & = \Pr[2(\log q_n^0(\hat{\theta}) - \log q_n^1(\hat{\theta}_1)) \leq 2n(E_*\|\hat{\theta} - \theta_0\|_*^2 - E_*\|\hat{\theta}_1 - \theta_{1,0}\|_*^2)] \\ & = \Pr[LR \leq 2n(E_*\|\hat{\theta} - \theta_0\|_*^2 - E_*\|\hat{\theta}_1 - \theta_{1,0}\|_*^2)], \end{aligned}$$

and in the case of the error term and the regressor being independent

$$\begin{aligned} (3.6) \qquad & \Pr[nAIC_1 < nAIC_0] \\ & = \Pr[2(\log q_n^0(\hat{\theta}) - \log q_n^1(\hat{\theta}_1)) \leq 2(k - k_1)] \\ & = \Pr[LR \leq 2(k - k_1)]. \end{aligned}$$

For n such that $\log(n) > 1/2$ we can clearly see that

$$4\log\left(\Pi_{i=k_1+1}^k s_i(n)\right) > 2(k - k_1)$$

so that

$$\Pr[GBIC(II)_0 < GBIC(II)_1] > \Pr[nAIC_1 < nAIC_0],$$

which implies that GBIC(II) is more parsimonious than AIC. The same is true for GBIC(I).

Now, for C_p and for $\bar{R}^2$

$$\Pr[\text{choose } m_1 \text{ over } m_0] = \Pr[(C_p - k_1) < 0]; \tag{3.7}$$

$$\Pr[\text{choose } m_1 \text{ over } m_0] = \Pr[\bar{R}_1^2 > \bar{R}_0^2]. \tag{3.8}$$

But, since

$$(C_p - k_1) = (n - k_1)\left(\frac{(1 - \bar{R}_1^2)}{(1 - \bar{R}_0^2)} - 1\right)$$

from (2.23), we have

$$\Pr[(C_p - k_1) < 0] = \Pr[\bar{R}_1^2 > \bar{R}_0^2]. \tag{3.9}$$

Therefore, $\bar{R}^2$ and C_p have the same level of parsimony.

Now, in the case $\varepsilon_t \sim N(0, \sigma^2)$ in (2.19) the $\bar{R}^2$ criterion is such that

$$\begin{aligned} LR &= 2(\log q_n^0(\hat{\theta}) - \log q_n^1(\hat{\theta}_1)) \\ &= n \log \frac{(T - k_1)(1 - \bar{R}_1^2)}{(T - k)(1 - \bar{R}_0^2)}, \end{aligned} \tag{3.10}$$

so that for $\rho(\bar{R}_0^2, \bar{R}_1^2) = (1 - \bar{R}_1^2)/(1 - \bar{R}_0^2)$

$$n \log \rho(\bar{R}_0^2, \bar{R}_1^2) = LR - n \log \frac{n - k_1}{n - k}$$

from which we have

$$\begin{aligned} \Pr[\text{choose } m_1 \text{ over } m_0] &= \Pr[n \log \rho(\bar{R}_0^2, \bar{R}_1^2) < 0] \\ &= \Pr[LR \leq n \log \frac{n - k_1}{n - k}] \\ &\approx \Pr[LR \leq (k - k_1)] \end{aligned} \tag{3.11}$$

where the approximation is through Taylor expansion for a large n: By Taylor expansion $\log \frac{n-k_1}{n-k} \approx \frac{k-k_1}{n-k}$, so that for a large n $n \log \frac{n-k_1}{n-k} \approx (k - k_1)$. Then, from (3.6) and (3.11)

$$\Pr[nAIC_1 < nAIC_0] > \Pr[n \log \rho(\bar{R}_0^2, \bar{R}_1^2) < 0]$$

which implies that AIC has a higher level parsimony than $\bar{R}^2$ or C_p in the case $\varepsilon_t \sim N(0, \sigma^2)$ in (2.19).

3.2 Power

We study how to compare power of the criteria after size adjustment. To adjust the size we need to get a critical value of each of the criteria. To get the critical value of a criterion for a given significance level we use the relationship between the criterion and LR studied above.

For GBIC(I)

$$\begin{aligned}(3.12)\qquad &\Pr[LR \le x] \\ &= \Pr[2(GBIC(I)_0 - GBIC(I)_1) - \log(|\Sigma_n^0|/|\Sigma_n^1|) \le x] \\ &= \Pr\left[(GBIC(I)_0 - GBIC(I)_1) \le \frac{1}{2}(x + \log(|\Sigma_n^0|/|\Sigma_n^1|))\right],\end{aligned}$$

and for GBIC(II)

$$\begin{aligned}(3.13)\qquad &= \Pr[LR \le x] \\ &= \Pr\left[(GBIC(II)_0 - GBIC(II)_1) \le \frac{1}{2}(x - 2\log\left(\Pi_{i=k_1+1}^{k} s_i(n)\right))\right].\end{aligned}$$

Therefore, for a $100(1-\alpha)\%$ level critical value of LR statistic x the corresponding critical value of $(GBIC(I)_0 - GBIC(I)_1)$ is $(x + \log(|\Sigma_n^0|/|\Sigma_n^1|))/2$ and that of $(GBIC(II)_0 - GBIC(II)_1)$ is $(x - 2\log\left(\Pi_{i=k_1+1}^{k} s_i(n)\right))/2)$.

Likewise, for AIC

$$\begin{aligned}(3.14)\qquad &\Pr[LR \le x] = \Pr[\frac{n}{2}(AIC_1 - AIC_0) + 2(k - k_1) \le x] \\ &= \Pr[n(AIC_1 - AIC_0) \le \frac{1}{2}(x - 2(k - k_1))]\end{aligned}$$

where $(n/2)AIC$ is used to have the same scale as BIC, so that for a $100(1-\alpha)\%$ level critical value of LR, the corresponding critical value of $(n/2)(AIC_1 - AIC_0)$ is $(x - 2(k-k_1))/2$ or $(x - 2n(E_*\|\hat{\theta} - \theta_0\|_*^2 - E_*\|\hat{\theta}_1 - \theta_{1,0}\|_*^2))/2$.

Also, from (3.11) we have

$$(3.15)\qquad \Pr[LR \le x] \approx \Pr[\frac{n}{2}\log\rho(\bar{R}_0^2, \bar{R}_1^2) \le \frac{1}{2}(x - (k - k_1))],$$

so that for a $100(1-\alpha)\%$ level critical value of LR statistic x the corresponding critical value of $n\log\rho(\bar{R}_0^2, \bar{R}_1^2)$ is $(x - (k-k_1))/2$.

Because the distribution of the LR statistic is usually not known under m_0, analytical comparison of power of the criteria would not be possible. However, we can perform simulation study to compare power under m_0, for which we can use the critical values of the criteria obtained above.

From the results (3.12)-(3.15) we can guess that GBIC would have the highest power eventually as n gets larger. This is because

$$
\begin{aligned}
(3.16) \qquad & \text{Power}(m_0) = \Pr[\text{choose } m_0 \text{ over } m_1 \text{—} m_0] \\
& = \Pr[GBIC(I)_0 - GBIC(I)_1 > \frac{1}{2}(x + \log(|\Sigma_n^0|/|\Sigma_n^1|)); \\
& \Pr[GBIC(II)_0 - GBIC(II)_1 > \frac{1}{2}(x - 2\log\left(\Pi_{i=k_1+1}^{k} s_i(n)\right)); \\
& \Pr[nAIC_1 - nAIC_0 > \frac{1}{2}(x - 2(k - k_1))]; \\
& \Pr[n\log\rho(\bar{R}_0^2, \bar{R}_1^2) > \frac{1}{2}(x - 2(k - k_1))].
\end{aligned}
$$

and $(-\log(|\Sigma_n^0|/|\Sigma_n^1|))$, $\log\left(\Pi_{i=k_1+1}^{k} s_i(n)\right) \to \infty$.

3.3 Consistency

From (3.3) and (3.4) we can see that GBIC(I) and GBIC(II) are consistent criteria if $\Pr[LR < \infty] = 1$ since $(-\log(|\Sigma_n^0|/|\Sigma_n^1|)$ and $4\log\left(\Pi_{i=k_1+1}^{k} s_i(n)\right)$ are positive and of $O_p(\log(n))$. However, AIC, $\bar{R}^2$ and C_p are not consistent, which is clear from (3.5)-(3.6) and (3.11).

3.4 Spurious Regressions and Model Selection

4 Examples and Simulation Study

In this section we study several examples of models for which some or all of the above criteria can be applied. We examine performance of the those criteria based on Monte Carlo simulation. Four examples are considered: *(i)* decision between $I(1)$ and $I(0)$; *(ii)* determination of the number of structural breaks in a model with

trend breaks; *(iii)* a vector error correction model and determination of the rank of cointegrating relations; *(iv)* order determination in an autoregression.

4.1 Decision Between I(1) and I(0) in an AR model

Suppose that a stochastic process y_t follows an $AR(p)$:

$$y_t = \sum_{i=1}^{p} \phi_i y_{t-i} + \varepsilon_t,$$

where $\varepsilon_t \sim N(0, \sigma^2)$. Consider an alternative representation of the process y_t:

$$y_t = \rho y_{t-1} + \sum_{i=1}^{p-1} \psi_{1,i} \Delta y_{t-i} + \varepsilon_t, \tag{4.1}$$

where $\rho = \sum_{i=1}^{p} \phi_i$ is the parameter of long-run autoregressive impact, and $\psi_{1,i} = -\sum_{j=i+1}^{p} \phi_j$ represents transient dynamics. If the characteristic equation $1 - \sum_{i=1}^{p} \phi_i z^i = 0$ has a unit root, then $\rho = 1$, in which case y_t is an $I(1)$ process. We can further transform the model (4.1) to

$$\Delta y_t = \beta_1 y_{t-1} + \sum_{i=1}^{p-1} \psi_{1,i} \Delta y_{t-i} + \varepsilon_t, \tag{4.1$'$}$$

where $\beta_1 = \rho - 1$. Then, the decision beween $I(1)$ and $I(0)$ is to decide between

$$\begin{aligned} &H_0 : \textit{ the model } (4.1') \textit{ with } \beta_1 = 0 \textit{ and} \\ &H_1 : \textit{ the model } (4.1') \textit{ with } \beta_1 < 0. \end{aligned} \tag{4.2}$$

The decision criterion of GBIC1 (2.9) for the hypotheses in (4.2) is

$$\begin{aligned} &\textit{choose the model } j \in \{0, 1\} \textit{ that minimizes} \\ &(n - k_j) \log(\hat{\sigma}_j^2)/n + \log(|X_j' X_j|)/n - k_j \log(2\pi)/n, \end{aligned} \tag{4.3}$$

where $\hat{\sigma}_j^2 = Y_n' M Y_n / n$, $Y_n = (y_1, ..., y_n)$, $M = I - X(X'X)^{-1}X'$, and $X = (x_1, ..., x_n)'$ where $x_t = (\Delta y_{t-1}, ..., \Delta y_{t-p+1})'$ for $j = 0$ and $x_t = (y_{t-1}, \Delta y_{t-1}, ..., \Delta y_{t-p+1})'$ for $j = 1$.

To evaluate performance of GBIC (4.3) we do two Monte Carlo simulations. First, we evaluate the probability of selecting the true model H_ρ against $H_0 : \rho = 1$ for various possible values of ρ from 0 to 1. The simulated model is an AR(2) with $\phi_2 = .3$ and with ϕ_1 running from .5 through .7, so that the range of the root ρ is from .8 to 1. The sample size is 100 with 5000 replications. The result is reported in Table 4.1. We find that, uniformly in ρ, the GBIC has higher frequency of selecting the true model than Schwarz criterion. Especially, for larger values of ρ, difference between the GBIC and Schwarz criterion becomes substantial. Schwarz criterion becomes heavily biased toward selecting H_0 for larger values of ρ.

Second we compare power of GBIC with two other decision criteria, Schwarz criterion and the augmented Dickey-Fuller test. As usual, we define the power as the probability of selecting a model when it is true. The critical values for GBIC and SBIC are obtained by the method studied in Section 3. As above, the simulated model is an AR(2) with $\phi_2 = .3$ and with ϕ_1 running from .5 through .7, so that the range of the root ρ is from .8 to 1. The sample size is 100 with 5000 replications. The result is displayed in Table 4.2 and Figure 4.1. As is shown in Figure 4.1, GBIC (4.3) has much better power property than Schwarz criterion or the augmented Dickey-Fuller t-test.

4.2 Determination of the Number of Trend Breaks

The second example of a model for which we want to examine the performance of our Bayesian criterion is a model with structural breaks. Suppose that there are several historical events, each of which brings permanent change in the trend of a linear time series. Thus, consider the following model:

$$y_t = \alpha + \beta t + \sum_{i=1}^{q} \gamma_i (t - k_i^0) I(t \geq k_i^0) + Z_t \tag{4.4}$$

where $I(\cdot)$ is the indicator function and $k_1^0, \ldots, k_q^0$ are break dates that are unknown. At a break date k_i^0, the trend shifts in the amount of γ_i.

For simplicity, we assume that $Z_t \sim N(0, \sigma_0^2)$. For $\theta = (\alpha, \beta, \gamma_1, ..., \gamma_q)'$, we have $s_1 = 1/2,\ s_2 = 3/2,\ s_3 = 3/2, ..., s_k = 3/2$ as s_i is the rate of convergence of $\hat{\theta}_i$. Therefore, GBIC2 (2.10) in this case is

(4.5) *choose the model with j that minimizes*

$$(n - k_j)\log(\hat{\sigma}_j^2)/n + (4 + 3q_j)\log(n)/n - k_j \log(2\pi)/n,$$

where q_j is the number of structural breaks in the model j, while GBIC1 is as in (2.9) with $x_t = (1, t, (t - k_1)I(t \geq k_1), ..., (t - k_q)I(t \geq k_q)$.

Monte Carlo simulation was performed to evaluate the performance of GBICs and AIC for selecting the true model. We consider the true values of q running from 0 to 2. The simulated model is the model (4.4) with $\alpha = .5$, $\beta = .2$, $\gamma_1 = -.15$ and $\gamma_2 = .1$. For the model with $q = 1$ $\tau_1 = .5$, and for the model with $q = 2$ $\tau_1 = .25$ and $\tau_2 = .5$. The sample size is 200 with 2000 replications. The result is reported in Table 4.2. We find that Schwarz criterion is heavily biased toward selecting a model with more breaks than the true model.[3] On the other hand, GBIC2 chooses the true model with relative frequency one for all three cases of true q from 0 to 2. Performance of GBIC1 is not as good as that of GBIC2 but better than that of Schwarz criterion. The performance of AIC is so bad, as is shown in Table 4.2. This might be an evidence that AIC would not provide a reliable result for model selection when the 'distance' between the true model and an assumed model is large. The model (4.4) is a model with nonstationary components (the trend and breaks of the trend) with the dimension of the nonstationarity, the number of breaks, being unknown. In this case, the 'distance' between the true model and an alternative model is much larger than that assumed in Akaike (1973), order $n^{-3/8}$.

[3]Bai and Perron (1995) also found that SBIC usually results in a biased inference with upward bias in choosing the number of structural breaks. The bias is due to its ignorance of the fact that $s_i \neq 1/2$ for nonstationary components.

4.3 VECM and Cointegration Rank Determination

Suppose that an m-vector $I(1)$ process $\mathbf{y}_t$ can be characterized by a vector autoregression of order p in levels. Then, we can write it in the following *error correction representation* form,

$$\Delta \mathbf{y}_t = \sum_{i=1}^{p-1} \zeta_i \Delta \mathbf{y}_{t-i} + \alpha + \zeta_0 \mathbf{y}_{t-1} + \varepsilon_t, \tag{4.6}$$

where $\varepsilon_t \sim N(\mathbf{0}, \Sigma_{\varepsilon\varepsilon})$, Under the hypothesis H_r that there are exactly $r(\geq 1)$ cointegrating relations in $\mathbf{y}_t$, the matrix ζ_0 is restricted as $\zeta_0 = -\mathbf{B}\mathbf{A}'$, for $\mathbf{B}$ an $(m \times r)$ matrix and $\mathbf{A}'$ an $(r \times m)$ matrix. Using the matrices $\mathbf{A}'$ and $\mathbf{B}$, we can rewrite the above expression (4.6) by

$$\Delta \mathbf{y}_t = \sum_{i=1}^{p-1} \zeta_i \Delta \mathbf{y}_{t-i} + \alpha + \mathbf{B}\mathbf{z}_{t-1} + \varepsilon_t, \tag{4.7}$$

where $\mathbf{z}_t = \mathbf{A}'\mathbf{y}_t$.

Let $\theta = vec(\zeta_1, ..., \zeta_{p-1}, \alpha, \mathbf{B})$, where $vec(\cdot)$ denotes the $((m^2p + mr) \times 1)$ vector formed by stacking all the parameters in $(\cdot)$. Now, under H_r that there exist exactly r cointegration relations the second derivative of log-likelihood is $L_n''(\hat{\theta}_n) = -\hat{\Sigma}_{\varepsilon\varepsilon}(r)^{-1} \otimes (X_r'X_r)$, so that the covariance matrix of θ, Σ_n, is

$$\Sigma_n \equiv [-L_n''(\hat{\theta}_n)]^{-1} = \hat{\Sigma}_{\varepsilon\varepsilon}(r) \otimes (X_r'X_r)^{-1} \tag{4.8}$$

where X_r is the matrix X for the model H_r and $\hat{\Sigma}_{\varepsilon\varepsilon}(r)$ is the MLE of $\Sigma_{\varepsilon\varepsilon}$ under H_r.

From the full information maximum likelihood (FIML) estimate of the system (4.7) studied in Johansen (1991) we get the following expression for the maximized likelihood of (4.7) under H_r:

$$\begin{aligned} &\log p_n(Y_n|\hat{\theta}_n) \\ &\quad = -(nm/2)\log(2\pi) - (nm/2) - (n/2)\log|\hat{\Sigma}_{uu}| - (n/2)\sum_{i=0}^{r}\log(1-\hat{\lambda}_i), \end{aligned} \tag{4.9}$$

where $\hat{\lambda}_0 = 0$ and $\hat{\lambda}_i$ for $i = 1, ..., r$ is the i^{th} largest eigenvalue of the matrix

$$\hat{\Sigma}_{vv}^{-1}\hat{\Sigma}_{vu}\hat{\Sigma}_{uu}^{-1}\hat{\Sigma}_{uv} \tag{4.10}$$

where $\hat{\Sigma}_{uu} = T^{-1}\sum \hat{u}_t\hat{u}_t'$, $\hat{\Sigma}_{vv} = T^{-1}\sum \hat{v}_t\hat{v}_t'$, $\hat{\Sigma}_{uv} = T^{-1}\sum \hat{u}_t\hat{v}_t'$ and $\hat{\Sigma}_{vu} = \hat{\Sigma}_{uv}'$, where $\hat{u}$ and $\hat{v}$ are residuals from the following regressions

$$\Delta \mathbf{y}_t = \hat{\pi}_0 + \hat{\Pi}_1 \Delta \mathbf{y}_{t-1} + \cdots + \hat{\Pi}_{p-1}\Delta \mathbf{y}_{t-p+1} + \hat{u}_t \quad \text{for } t = 1, ..., n$$
$$\mathbf{y}_{t-1} = \hat{\pi}_0 + \hat{\Pi}_1 \Delta \mathbf{y}_{t-1} + \cdots + \hat{\Pi}_{p-1}\Delta \mathbf{y}_{t-p+1} + \hat{v}_t \quad \text{for } t = 1, ..., n$$

where 'hat' denotes the least square estimate.

Notice that the first three terms of $\log p_n(Y_n|\hat{\theta}_n)$ in (4.9) do not depend on the cointegration dimension r. Then, from (4.8) and (4.9), the GBIC(I) for determining the cointegration rank r is

(4.11) *choose the model H_r that minimizes*
$$\sum_{i=0}^{r} \log(1-\hat{\lambda}_i) - \log(|\hat{\Sigma}_{\varepsilon\varepsilon}(r) \otimes (X_r'X_r)^{-1}|)/n - (mr)\log(2\pi)/n$$

while GBIC(II) is

(4.12) *choose the model H_r that minimizes*
$$\sum_{i=0}^{r} \log(1-\hat{\lambda}_i) - k\log(n)/n - (mr)\log(2\pi)/n$$

where k is the number of elements in θ.

We compare performance of our Bayesian procedures (4.11)-(4.12) with that of Jobansen (1991)'s LR_1 and LR_2 by simulation.[4] The simulation is based on AR(2) models with the true $r = 0, 1, ..., 4$, sample size 200 and 2000 replications:

(4.13) $$\mathbf{y}_t = \alpha + \Phi_1 \mathbf{y}_{t-1} + \Phi_2 \mathbf{y}_{t-2} + \varepsilon_t,$$

where the parameter values are set to be $\alpha = 0$,

$$A' = \begin{pmatrix} 1 & -0.3 & -0.4 & 0.2 & 0.1 \\ 0 & 1 & -0.7 & 0.2 & 0.3 \\ 0 & 0 & 1 & 0.5 & 0.1 \\ 0 & 0 & 0 & 1 & -0.4 \end{pmatrix}; \quad B = \begin{pmatrix} \mathrm{I}_r \\ 0_{(n-r)\times r} \end{pmatrix}$$

[4] LR_1: a likelihood-ratio test for testing H_r against H_m, LR_2: a likelihood-ratio test for testing H_r against H_{r+1}.

for $r = 4$, and for $r = 1, 2, 3$ the first r rows of the above A' are taken;

$$\Phi_2 = \begin{pmatrix} 0.2 & -0.2 & 0.1 & -0.1 & 0.4 \\ 0.2 & 0.1 & 0.1 & -0.2 & -0.1 \\ 0.3 & 0.1 & 0.3 & -0.4 & 0.3 \\ -0.2 & -0.1 & 0.2 & 0.3 & -0.1 \\ 0.1 & 0.2 & -0.1 & 0.2 & 0.9 \end{pmatrix},$$

and $\Phi_1 = -BA' + I_n - \Phi_2$; $\Sigma_{\varepsilon\varepsilon} = 0.01 I_n$.

The results are presented in Table 4.3. For all values of true r's, GBICs (4.11) and (4.12) show better performance than LR_1 and LR_2. The difference in the performance becomes larger as the true r gets larger. For $r = 1, 2$ GBIC2 shows better performance than GBIC1 while for $r = 3, 4$ GBIC1 shows better performance. This pattern of the results was not sensitive to the parameter values adopted.

4.4 A Regression with Non-i.i.d. Error and AR Order Determination

Consider the regression (2.21) but with a serially correlated error.

$$y_t = \beta' x_t + \varepsilon_t \tag{2.21}$$

$$\varepsilon_t = \rho \varepsilon_{t-1} + v_t, \quad v_t \sim i.i.d.N(0, \sigma_v^2) \tag{4.14}$$

Let $\theta = vec(\beta, \rho, \sigma_v^2)$, a $k \times 1$ vector. The GBIC(I) for (2.21)-(4.14) is to choose the model that minimizes

$$\log\left(\sum_{t=1}^{n} \hat{v}_t^2/n\right) + \log(|-L_n''(\hat{\theta})|)/n - k\log(2\pi)/n \tag{4.15}$$

where $\hat{v}_t = \hat{\varepsilon}_t - \hat{\rho}\hat{\varepsilon}_{t-1}$ for $\hat{\varepsilon}_t = y_t - \hat{\beta}' x_t$. On the other hand, AIC (2.16) is

$$\log\left(\sum_{t=1}^{n} \hat{v}_t^2/n\right) + \frac{2k}{n}. \tag{4.16}$$

Bayesian information criterion is widely used for determining the order of an autoregression. Now, as an example of the model (2.21)-(4.14) we consider an autoregression of order p with a serially correlated error:

$$(4.17) \qquad \begin{aligned} y_t &= \phi_1 y_{t-1} + \ldots + \phi_p y_{t-p} + \varepsilon_t \\ & \varepsilon_t = \rho \varepsilon_{t-1} + v_t, \quad v_t \sim i.i.d.N(0, \sigma_v^2). \end{aligned}$$

We compare performance of Schwarz criterion, GBIC(I) (4.15) and AIC (4.16) in classical cases of stationary y_t. The simulation is based on $AR(p)$ models for $p = 1, .., 4$ with $\sum_{i=1}^{p} \phi_i = 0.9$, $\rho = 0.2$, and $\sigma_v^2 = 0.01$; sample size 200 and 2000 replications.

The results are presented in Table 4.4. For all values of true p's, GBIC (4.15) shows consistently the best performance in selecting the true p. On the other hand, the other criteria sometimes overestimate and sometimes underestimate the true p. Also, from an extended simulation we have found that the difference in performance between GBIC (4.15) and the other criteria became larger as ρ becomes bigger.

References

[1] Akaike, H. (1973) "Information Theory and an Extension of the Maximum Likelihood Principle," in B.N. Petrov and F. Csaki, eds., *Second International Symposium on Information Theory,* (Budapest: Akademiai Kaido.

[2] Bai, J. and P. Perron (1995). "Testing and Estimation of Multiple Structural Changes," Manuscript, Presenred in North American winter meetings of Econometric Society.

[3] Chib S. (1995) "Marginal Likelihood from the Gibbs Output," *Journal of the American Statistical Association*, 90, 432, 1313-1321.

[4] Chao J.C. and P.C.B. Phillips (1998) "Posterior Distributions in Limited Information Analysis of the Simultaneous Equations Model Using the Jeffreys Prior," *Journal of Econometrics*, 87, 49-86.

[5] Chao J.C. and P.C.B. Phillips (1999) "Model Selection in Partially Nonstationary Vector Autoregressive Processes with Reduced Rank Structure," *Journal of Econometrics*, 91, 227-271.

[6] Geweke, J. and R. Meese (1981) "Estimating Regression Models of Finite But Unknown Order," *International Economic Review*, 22, 1.

[7] Johansen, S. (1991) "Estimation and Hypothesis Testing of Cointegrating Vectors in Gaussian Vector Autoregression Models," *Econometrica*, 59, 1551-1580.

[8] Kim, J. Y. (1998) "Large Sample Properties of Posterior Densities, Schwarz Criterion and the Likelihood Principle in Nonstationary Time Series Models," *Econometrica*, 66, 359-380.

[9] Leamer, E. E. (1978) *Specification Searches*, Wiley, New York.

[10] Phillips P. C. B. (1996) "Econometric Model Determination," *Econometrica*, 64, 763-812

[11] Phillips, P. C. B. and W. Ploberger (1994) "Posterior Odds Testing for a Unit Root with Data-based Model Selection," *Econometric Theory*, 10, 774-808.

[12] Phillips, P. C. B. and W. Ploberger (1996) "An Asymptotic Theory of Bayesian Inference for Time Series," *Econometrica*, 64, 381-412

[13] Pötscher B. M. (1989) "Model Selection Under Nonstationarity: Autoregressive Models and Stochastic Linear Regression Models," *Annals of Statistics*, pp.1257-1274.

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

[15] Sims, C. A. (1988) "Bayesian Skepticism on Unit Root Econometrics," *Journal of Economic Dynamics and Control*, no.12, 463-474.

[16] Sweeting, T.J. and Adekola, A.O. (1987) "Asymptotic Posterior Normality for Stochastic Processes Revisited," *J. of Royal Stat. Soc., Series B.* 49, 215-222.

[17] Tsay (1984) "Order Selection in Nonstationary Autoregressive Models," *Annals of Statistics*, pp.1425-1433.

Mathematical Appendix

Proof of Lemma 2.1: First, we will show that

$$\{p_T(Y_T)^{-1}\pi(\theta_0)p_T(Y_T|\hat{\theta}_T)|\Sigma_T|^{1/2}\}(2\pi)^{k/2} = r_0(T)$$

where $r_0(T) \longrightarrow 1$. When the meaning is clear from the context, we omit the word "convergence in probability." Write

$$p_T(Y_T|\theta) = p_T(Y_T|\hat{\theta}_T)\exp[L_T(\theta) - L_T(\hat{\theta})].$$

By Taylor expansion,

$$L_T(\theta) = L_T(\hat{\theta}) + \frac{1}{2}(\theta - \hat{\theta}_T)'L_T''(\bar{\theta^1}, ..., \bar{\theta^k})(\theta - \hat{\theta}_T)$$

where $\bar{\theta^i} = \lambda^i\theta + (1-\lambda^i)\hat{\theta}_T$ for $\theta \in N(\hat{\theta}_T, \delta_T)$ for $\lambda^i \in [0,1]$, $\lambda^i \in [0,1]$ for each $i = 1, ..., k$, and

$$L_T''(\bar{\theta^1}, ..., \bar{\theta^k}) = \begin{pmatrix} \frac{\partial(\partial L_T/\partial\theta_1)}{\partial\theta}(\bar{\theta^1}) \\ \vdots \\ \frac{\partial(\partial L_T/\partial\theta_k)}{\partial\theta}(\bar{\theta^k}) \end{pmatrix}.$$

Letting

$$R_T \equiv R_T(\bar{\theta^1}, ..., \bar{\theta^k}) = [L_T''(\hat{\theta}_T)]^{-1}[L_T''(\bar{\theta^1}, ..., \bar{\theta^k}) - L_T''(\hat{\theta}_T)],$$

we have

$$\begin{aligned} p_T(Y_T|\theta) &= p_T(Y_T|\hat{\theta}_T)\exp[\frac{1}{2}(\theta - \hat{\theta}_T)'L_T''(\bar{\theta}_T)(\theta - \hat{\theta}_T)] \\ &= p_T(Y_T|\hat{\theta}_T)\exp[-\frac{1}{2}(\theta - \hat{\theta}_T)'(I_k + R_T)\Sigma_T^{-1}(\theta - \hat{\theta}_T)]. \end{aligned}$$

So,

$$\begin{aligned}\pi_T(\theta|Y_T) &= p_T(Y_T)^{-1}\pi(\theta)p_T(Y_T|\theta) \\ &= p_T(Y_T)^{-1}\pi(\theta)p_T(Y_T|\hat{\theta}) \\ &\cdot \exp[-\frac{1}{2}(\theta-\hat{\theta}_T)'(I_k+R_T)\Sigma_T^{-1}(\theta-\hat{\theta}_T)].\end{aligned}$$

Now by (C1)(a), $\|R_T\|$ has upper and lower bounds, say r_T^+ and r_T^-, respectively. Further, by the assumption on the prior given $\epsilon > 0$, we can choose $T_0 = T_0(\epsilon)$ such that for $T > T_0$

$$(A.1) \qquad 1-\epsilon < \inf_{\theta\in N(\hat{\theta}_T,\delta_T)} \pi(\theta)/\pi(\theta_0) < \sup_{\theta\in N(\hat{\theta}_T,\delta_T)} \pi(\theta)/\pi(\theta_0) < 1+\epsilon.$$

Now define an integral $I_T(a,b)$ by

$$(A.2) \qquad I(N(\hat{\theta}_T,\delta_T)) = \int_{N(\hat{\theta}_T,\delta_T)} \pi_T(\theta|Y_T)d\theta.$$

Then by $(C1)(a)$ and (A.1), the integral (A.2) is bounded above by

$$\begin{aligned}I^+(N(\hat{\theta}_T,\delta_T)) &= (1+\epsilon)p_T(Y_T)^{-1}\pi(\theta_0)p_T(Y_T|\hat{\theta}_T) \\ &\times \int_{J_T^+} \exp[-\frac{1}{2}(\theta-\hat{\theta}_T)'(I_k+R_T)\Sigma_T^{-1}(\theta-\hat{\theta}_T)]d\theta\end{aligned}$$

where $J_T^+ = \{z : z \in (-\Sigma_T^{-1/2}\delta_T(1+r_T^-)^{1/2}, \Sigma_T^{-1/2}\delta_T(1+r_T^-)^{1/2})\}$. But it becomes, by change of variable,

$$\begin{aligned}&= (1+\epsilon)p_T(Y_T)^{-1}\pi(\theta_0)p_T(Y_T|\hat{\theta}_T) \\ &\times|I_k(1+r_T^-)|^{-1/2}|\Sigma_T|^{1/2}\int_{J_T^+}\exp[-z'z/2]dz.\end{aligned}$$

Also, the integral (A.2) is bounded below by

$$\begin{aligned}I^-(N(\hat{\theta}_T,\delta_T)) &= (1+\epsilon)p_T(Y_T)^{-1}\pi(\theta_0)p_T(Y_T|\hat{\theta}_T) \\ &\times|I_k(1+r_T^+)|^{-1/2}|\Sigma_T|^{1/2}\int_{J_T^-}\exp[-z'z/2]dz\end{aligned}$$

where $J_T^- = \{z : z \in (-\Sigma_T^{-1/2}\delta_T(1+r_T^+)^{1/2}, \Sigma_T^{-1/2}\delta_T(1+r_T^+)^{1/2})\}$. But by (C1)(b), each component of $\Sigma_T^{-1/2}\delta_T$ tends to infinity as $T \nearrow \infty$, so that for $J_T = J_T^+$ or J_T^-

we get

$$\int_{J_T} \exp[-z'z/2]dz \xrightarrow{p} (2\pi)^{k/2}.$$

Hence, for large T, we have

$$\begin{aligned}(1+\epsilon)^{-1}(1+r_T^-)^{1/2} & I(N(\hat{\theta}_T, \delta_T)) \\ & \leq p_T(Y_T)^{-1}\pi(\theta_0)p_T(Y_T|\hat{\theta}_T)|\Sigma_T|^{1/2}(2\pi)^{k/2} \\ & \leq (1-\epsilon)^{-1}(1+r_T^+)^{1/2} I(N(\hat{\theta}_T, \delta_T)).\end{aligned}$$

But since ϵ is arbitrary and r_T^+ and $r_T^- \searrow 0$, it implies that

$$(A.3) \qquad \lim_{T\nearrow\infty} \{p_T(Y_T)^{-1}\pi(\theta_0)p_T(Y_T|\hat{\theta}_T)|\Sigma_T|^{1/2}\} \leq (2\pi)^{-k/2}$$

where equality holds if and only if $I(N(\hat{\theta}_T, \delta_T)) \to 1$. But,

$$(C2) \Longleftrightarrow I(N(\hat{\theta}_T, \delta_T)) \longrightarrow 1,$$

which is equivalent to

$$\{p_T(Y_T)^{-1}\pi(\theta_0)p_T(Y_T|\hat{\theta}_T)|\Sigma_T|^{1/2}\}(2\pi)^{k/2} = r_0(T)$$

with $r_0(T) \longrightarrow 1$. Then it follows that

$$\log p_T(Y_T) = \log(p_T(Y_T|\hat{\theta}_T)) + (1/2)\log(|\Sigma_T|) + (k/2)\log(2\pi) + \log(\pi(\theta_0)) + R_0(T),$$

where $R_0(T) = -\log(r_0(T))$. But we know that

$$p_T(Y_T) = \int_\Theta p_T(Y_T|\theta)\pi(\theta)d\theta.$$

Proof of Lemma 2.2: From condition (2.6), we have

$$\begin{aligned}|\mathcal{D}(T)||\Sigma_T|^{-1}|\mathcal{D}(T)| &= |\mathcal{D}(T)|^2|\Sigma_T|^{-1} = |\Sigma_T \prod_{i=1}^{k} s_i(T)^{-2} \\ &= O_p(1)\end{aligned}$$

Taking log and rearranging terms we get the conclusion.

Table 4.1.1 Relative Frequency of Selecting the True Model: Unit Root v.s. Stationarity

ρ	SBIC	GBIC
0.80	0.9027	0.9526
0.82	0.8578	0.9216
0.84	0.7928	0.8680
0.86	0.6868	0.7802
0.88	0.5826	0.6758
0.90	0.4370	0.5204
0.92	0.3328	0.4020
0.94	0.2138	0.2640
0.96	0.1328	0.1582
0.98	0.0696	0.0766
1.00	0.9584	0.9726

SBIC: Schwarz criterion.
GBIC: Generalized Bayesian information Criterion

Table 4.1.2 Power of SBIC, ADF(t) and GBIC for Decision between I(1) and I(0)

ρ	SBIC	ADF(t)	GBIC
0.80	0.9396	0.9636	0.9943
0.82	0.9006	0.9366	0.9870
0.84	0.8430	0.8933	0.9660
0.86	0.7546	0.8286	0.9343
0.88	0.6503	0.7313	0.8686
0.90	0.5273	0.6063	0.7796
0.92	0.4023	0.4830	0.6390
0.94	0.2886	0.3463	0.4780
0.96	0.1863	0.2323	0.3200
0.98	0.1036	0.1383	0.1723
1.00	0.0523	0.0530	0.0520*

ADF(t): Augmented Dickey-Fuller t-test
*Type I error is set to be 0.05. However, the exact 0.05 was not obtained for each of these three decision criteria, probably due to simulation bias or small sample bias

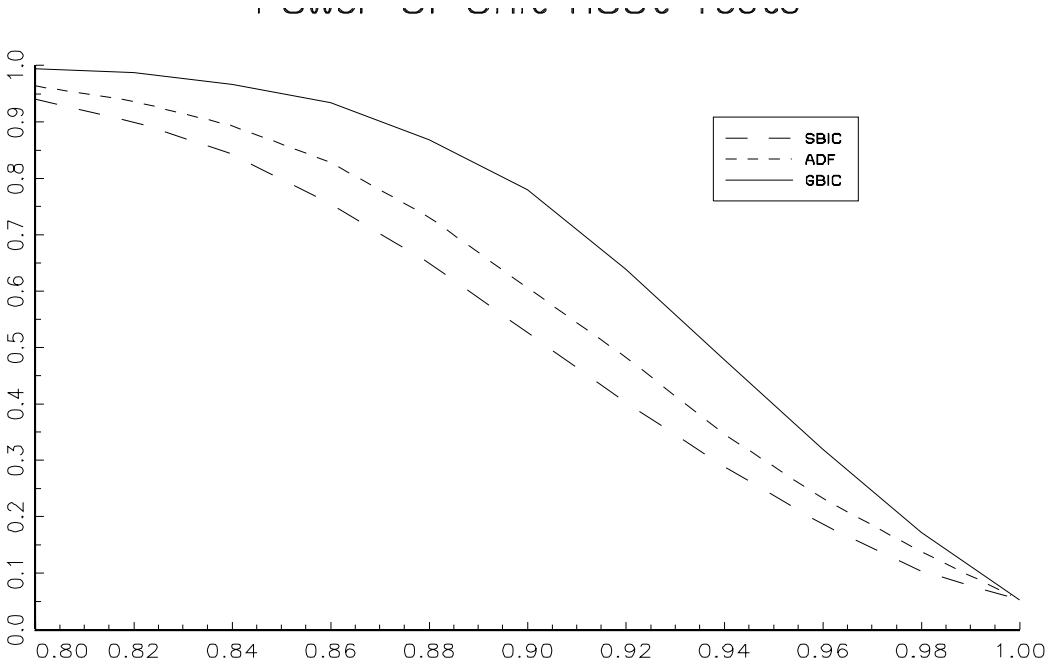

Table 4.2: Relative Frequency of Selecting the Number of Structural Breaks
(Sample Size =200, 2000 Replications)

True q=0

q	SBIC	GBIC1	GBIC2	AIC
0	0.799	0.888	1.000	0.006
1	0.136	0.085	0.000	0.011
2	0.050	0.020	0.000	0.036
3	0.015	0.007	0.000	0.947

True q=1

q	SBIC	GBIC1	GBIC2	AIC
0	0.000	0.000	0.000	0.000
1	0.724	0.846	1.000	0.001
2	0.196	0.119	0.000	0.006
3	0.080	0.035	0.000	0.993

True q=2

q	SBIC	GBIC1	GBIC2	AIC
0	0.000	0.000	0.000	0.000
1	0.000	0.000	0.000	0.000
2	0.670	0.783	1.000	0.000
3	0.330	0.217	0.000	1.000

Table 4.3 Determination of Cointegration Rank

True r=1

r	LR1	LR2	GBIC1	GBIC2
0	0.0000	0.0000	0.0000	0.0010
1	0.9370	0.9235	0.9590	0.9990
2	0.0630	0.0720	0.0405	0.0000
3	0.0000	0.0035	0.0005	0.0000
4	0.0000	0.0005	0.0000	0.0000

True r=2

r	LR1	LR2	GBIC1	GBIC2
0	0.0000	0.0000	0.0000	0.0000
1	0.0000	0.0000	0.0000	0.0080
2	0.9400	0.9285	0.9835	0.9920
3	0.0590	0.0665	0.0160	0.0000
4	0.0010	0.0045	0.0005	0.0000

True r=3

r	LR1	LR2	GBIC1	GBIC2
0	0.0000	0.0000	0.0000	0.0000
1	0.0000	0.0000	0.0000	0.0000
2	0.0000	0.0000	0.0000	0.0310
3	0.9365	0.9415	0.9975	0.9690
4	0.0610	0.0550	0.0025	0.0000

True r=4

r	LR1	LR2	GBIC1	GBIC2
0	0.0000	0.0000	0.0000	0.0000
1	0.0000	0.0000	0.0000	0.0000
2	0.0000	0.0000	0.0000	0.0010
3	0.0000	0.0000	0.0000	0.0250
4	0.9480	0.9480	1.0000	0.9740

LR1: Johansen (1991)’s test for H0 (h=r) against H1(h=n)
LR2: Johansen (1991)’s test for H0 (h=r) against H1(h=r+1)

Table 4.4 Determination of the Order of an Autoregression

True p=1

p	SBIC	AIC	GBIC
1	0.3677	0.0993	0.6763
2	0.5900	0.5600	0.1363
3	0.0333	0.1463	0.0743
4	0.0070	0.0880	0.0430
5	0.0020	0.1063	0.0700
6	0.0000	0.0000	0.0000

True p=2

p	SBIC	AIC	GBIC
1	0.1383	0.0250	0.0000
2	0.7280	0.4493	0.6807
3	0.1243	0.2683	0.1507
4	0.0077	0.1143	0.0723
5	0.0017	0.1430	0.0963
6	0.0000	0.0000	0.0000

True p=3

p	SBIC	AIC	GBIC
1	0.0343	0.0040	0.0000
2	0.6897	0.2983	0.1130
3	0.2377	0.3533	0.5997
4	0.0293	0.1810	0.1363
5	0.0090	0.1633	0.1510
6	0.0000	0.0000	0.0000

True p=4

p	SBIC	AIC	GBIC
1	0.0090	0.0000	0.0000
2	0.1593	0.0203	0.0057
3	0.4983	0.2553	0.1357
4	0.2970	0.4310	0.6127
5	0.0363	0.2933	0.2460
6	0.0000	0.0000	0.0000