Journal of Econometrics 169 (2012) 247–257

Contents lists available at SciVerse ScienceDirect

Journal of Econometrics

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

Model selection in the presence of nonstationarity

Jae-Young Kim

School of Economics, Seoul National University, Seoul, 151-742, South Korea

ARTICLE INFO

Article history:
Available online 23 January 2012

JEL classification:
C1
C2
C3
C5

Keywords:
Model selection
Nonstationarity
Bayesian rule
Parsimony
Power

ABSTRACT

This paper studies model selection methods in the presence of nonstationarity. We focus on the Bayesian model selection rule and compare it with other criteria that are frequently used in econometric practice. First, we derive each of these criteria in the presence of nonstationarity. In particular, we study the Bayesian model selection rule in detail and derive three alternative forms of it in the presence of nonstationarity. One important feature of the Bayesian model selection criterion (BMSC) is that BMSC gives different weights to the stationary and nonstationary components of a model while the other criteria do not. This feature of BMSC is a desirable property for a model selection rule in the presence of possible nonstationarity. Second, we compare these criteria with regard to parsimony and power. We have found that BMSC shows the highest parsimony, AIC is the second, and C_p and $\bar{R}^2$, having the same level of parsimony, are the third. With regard to power, the order is not clearly established. However, for the size adjusted power BMSC becomes dominant as the sample size increases. Without size adjustment the order in the power is exactly the opposite to that in parsimony. Also, we find that BMSC is a consistent selection rule while the other criteria are not. Third, we consider four different cases of practical interest for which BMSC with some of the other criteria is applicable. We discuss how our BMSC can be used in these cases of practical interest. Results of an extensive Monte Carlo simulation for models in these four cases show that overall the BMSC outperforms other criteria.

1. Introduction

The model selection problem has been an important subject in econometrics as well as in many other areas of science. Aside from theoretical *a priori* considerations in model building, the subject of how well a model fits the data is an important guide to model selection. Sims (1988), Phillips (1996) and Phillips and Ploberger (1996) have noted that econometric model selection strategy needs to be reconsidered in the presence of nonstationarity. This paper reexamines and redevelops model selection criteria in the presence of nonstationarity.

Several model selection methods have been studied in the literature. Theil (1961) proposed using the adjusted R^2; Akaike (1973) provided an information criterion (AIC); Schwarz (1978) proposed a Bayesian information criterion; Mallows (1973) proposed a prediction criterion (C_p). Other criteria in the literature include Hannan and Quinn (1979)'s criterion, Geweke and Meese (1981)'s criterion. Cavanaugh (1999)'s Kullback Information criterion, and the deviance information criterion of Spiegelhalter et al. (2002). Also, Tsay (1984), Hurvich and Tsai (1989) and Pötscher (1989) have studied model selection methods in times series models. Recently, the model selection problem has been extended to moment selection as in Andrews (1999), Andrews and Lu (2001) and Hong et al. (2003). These model selection methods are concerned with parsimony, as was stressed in Zellner et al. (2001), as well as accuracy or power in choosing models.

In this paper we consider four different approaches to model selection that are frequently used in econometric practice including the Bayesian approach, AIC, Mallow's C_p, and $\bar{R}^2$. We derive each of these criteria in the presence of nonstationarity. In particular, we study the Bayesian model selection rule in detail and derive three alternative forms of it in the presence of nonstationarity. One of the three forms of the Bayesian model selection method is the same as PIC in Phillips (1996). One important feature of the Bayesian model selection criterion (BMSC) is that BMSC gives different weights to the stationary and nonstationary components of a model while the other criteria do not. This feature of BMSC is a desirable property for a model selection rule in the presence of possible nonstationarity. It implies among other things that different levels of parsimony have to be applied to the stationary and nonstationary components of a model. The fact that a nonstationary component has to be given a different weight from the one given to a stationary component is noted by Sims (1988), Phillips (1996), Phillips and Ploberger (1996) and Kim (1998).

In this paper we also compare the derived criteria with regard to parsimony and power. A model selection method implicitly or explicitly employs parsimony. That is, if two models fit the

E-mail address: jykim017@snu.ac.kr.

0304-4076/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.jeconom.2012.01.029

data equally well, a simpler model is preferred, as is noted by Zellner et al. (2001). Based on some theoretical analysis, we have found that BMSC shows the highest parsimony, AIC the second, and C_p and $\bar{R}^2$, having the same level of parsimony, are third. This theoretical finding is confirmed by our Monte Carlo study. On the other hand, accuracy of a model selection criterion is investigated by evaluating power values of the criterion, where the power is defined as the probability of selecting a model when that model is true. With regard to the power of the criteria the order is not clearly established. However, for the size adjusted power, the BMSC becomes dominant as the sample size increases. Without size adjustment, on the other hand, the order in the power is exactly opposite to that in parsimony. These two findings imply that the relatively higher power values of *AIC*, C_p and $\bar{R}^2$ criteria are obtained by sacrificing parsimony. They further imply that the *AIC*, C_p and $\bar{R}^2$ criteria may overfit the model by allowing excessive levels of the Type I error. We also find that BMSC is a consistent selection rule while the other criteria are not. Monte Carlo study of a few interesting cases reveals that the BMSC has better power properties than the other competing criteria.

We consider four different cases of practical interest for which BMSC with some or all of the other criteria under study is applicable: (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. We discuss how BMSC can be applied for these problems. Also, we note that the traditional Schwarz BIC is either an incorrect or an inappropriate BIC in the presence of nonstationarity.

The paper is organized as follows. Section 2 examines each of the model selection criteria of practical use in the presence of nonstationarity. Section 3 compares the model selection criteria in parsimony and power. Section 4 studies four different cases of practical interest for which the model selection methods discussed in the paper are applicable. We perform Monte Carlo simulation to examine finite sample performance of the model selection methods in those cases.

2. Model selection rules in general frameworks

Let $(\Omega, \mathcal{F}, P)$ be a probability space and $\{\mathcal{F}_t\}_{t\geq 0}$ 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 $\theta \in \mathbb{R}^p$. Let $(\mathcal{V}, \mathcal{G}, P_v)$ 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 regarding the true model. A model $m_i \in \mathcal{M}$ is associated with a parameter space Θ^i of dimension k_i for $i \in \mathcal{I}$ where $\mathcal{I} = \{1, \ldots, I\}$ for a finite positive integer I. Assume that for each m_i a family $P_n^i(\theta^i, Y_n)$ of distribution functions with a family of densities $p_n^i(\theta^i, Y_n)$ is defined on the measurable space $(\mathcal{V}, \mathcal{G}) \times (\Omega, \mathcal{F})$.

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

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

where $\Pr(m_i)$ is the prior probability that m_i is true. Also, $p_n(Y_n|m_j)$ is the marginalized likelihood obtained from marginalization with respect to θ^j of the likelihood $p_n^j(\theta^j, Y_n) = p_n(Y_n|\theta^j, m_j)$ for model m_j:

$$\begin{aligned} p_n(Y_n|m_j) &= \int p_n(Y_n|\theta^j, m_j)\pi(\theta^j|m_j)d\theta^j \\ &= E_j[p_n(Y_n|\theta^j)], \end{aligned} \tag{2.2}$$

where $\pi(\theta^j|m_j)$ 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 $E_i[p_n(Y_n|\theta^i)]$ is the largest. Phillips (1996) provides another dimension of justification for the Bayesian approach for model selection based on the notion of a Bayesian model measure.

2.1.2. Approximations

Criterion (2.2) involves computation of an integral of $p_n \times \pi$ with respect to θ^j in $\mathbb{R}^{p_j}$. Certainly this computation is not an easy task even with a very fast computer. Also, the choice of the range of θ in the computation is a non-trivial problem. In the following we provide an approximation to the marginal likelihood $p_n(Y_n|m_j)$ in (2.2) that is valid for a large sample. The approximation allows us to derive Bayesian criteria that are computationally simple to handle and yet have sound theoretical justification.

Let $N(\hat{\theta}_n, \delta_n)$, $n = 1, \ldots, \infty$, be such that

$$N(\hat{\theta}_n, \delta_n) = \{\theta: |\theta_1 - \hat{\theta}_{n1}|^2/\delta_{n1}^2 + \cdots + |\theta_k - \hat{\theta}_{nk}|^2/\delta_{nk}^2 < 1\} \tag{2.3}$$

where $\hat{\theta}_{ni}$ is the ith element of $\hat{\theta}_n$, the maximum likelihood estimator (MLE) of θ based on Y_n; $\delta_n = (\delta_{n1}, \ldots, \delta_{nk})'$ is a k-vector of real numbers; $|\cdot|$ denotes the usual Euclidean norm. Thus, $N(\hat{\theta}_n, \delta_n)$ is a neighborhood around $\hat{\theta}_n$ whose area is determined by δ_n. We consider a sequence $\{\delta_n\}$ such that δ_n becomes smaller and smaller as $n \nearrow \infty$, so that the neighborhood $N(\hat{\theta}_n, \delta_n)$ shrinks as n gets larger. Also, note that the rate of shrinkage of δ_{ni} can be different across different i's.

Assume that the log-likelihood $L_n(\theta) = \log p_n(\theta)$, for $p_n(\theta) = p_n(\theta, Y_n)$, 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* $M_n(\hat{\theta}_n(\omega), \delta_n) = \sup_{\theta\in N(\hat{\theta}_n,\delta_n)} \|[L_n''(\theta) - L_n''(\hat{\theta}_n)][L_n''(\hat{\theta}_n)]^{-1}\|$. *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 in P-probability as n goes to infinity.*

(C2) *Let* $\pi_n(\theta|Y_n)$ *be the posterior formed from the likelihood* p_n *and a prior* π. *For* δ_n *satisfying* (C1),

$$\int_{\Theta\setminus N(\hat{\theta}_n,\delta_n)} \pi_n(\theta|Y_n)d\theta \longrightarrow 0 \tag{2.4}$$

in P-probability as n goes to infinity i.e., θ *concentrates in* $N(\hat{\theta}_n, \delta_n)$ *in P-probability as n goes to infinity.*

Conditions (C1) and (C2) cover a very wide variety of models containing nonstationary components. The two conditions (C1) and (C2) combined with a continuity condition for a prior π are sufficient for the posterior $\pi_n(\theta|Y_n)$ to be asymptotically normal in the presence of possible nonstationarity, as is studied in Kim (1998). The shrinking neighborhood $N(\hat{\theta}_n, \delta_n)$ above is the key device handling the problem of possible nonstationarity in