Model Selection for Nested and Overlapping Non-Linear, Dynamic and Possibly Misspecified Models

Massimiliano Marcellino — Bocconi University

Barbara Rossi — Duke University

January 2008

Abstract.

This paper develops tests for the selection of competing non-linear dynamic models, focusing on the nested and overlapping cases. The null hypothesis is that the models are equally close to the Data Generating Process (DGP), according to a certain measure of closeness. The alternative is that one model is closer to the DGP. The models can be correctly specified or not. Their parameters can be estimated by a variety of methods, including (pseudo) Maximum Likelihood and OLS. The tests are symmetric and directional. Their asymptotic distribution under the null is either normal or a weighted sum of chi-square distributions, depending on the nesting characteristics of the competing models. The comparison of nested AR models, and of nested ARMAX-GARCH models are discussed as examples.

Keywords: Model Selection, Nested Models, Misspecified Models, Non-linear Models, MLE.
JEL Classification: C52, C53

We are extremely grateful to Silvia Goncalves for many detailed comments. We would also like to thank Clive Granger, Soren Johansen, Hal White and seminar participants at Harvard and UCSD for valuable comments.

Corresponding author: Massimiliano Marcellino, IGIER - Bocconi, Via Salasco 3/5, 20136 Milan, Italy. E-mail: massimiliano.marcellino@uni-bocconi.it

1. Introduction

The comparison of competing models has attracted considerable attention in the literature, generating three main approaches to the problem: information criteria, non-nested tests, and encompassing statistics, see e.g. Gourieroux and Monfort (1994), Lavergne (1998), Hendry and Richard (1982), Mizon and Richard (1986) and Hendry and Mizon (1990) for a detailed discussion of their relative merits. In all these methods the underlying idea is that one of the models under comparison is closer to the Data Generating Process (DGP), and in this sense it can be preferred. Instead, Vuong (1989) proposed a statistic to test for the null hypothesis that the models are equally close to the DGP, in the sense of having the same KLIC, and therefore they are equivalent. He studied the properties of such a test in the case of models for i.i.d. processes, estimated by Maximum Likelihood (MLE).

In this paper we suggest a statistic similar to Vuong's test, but with a much broader range of applicability. Actually, it can be used to compare models for dependent and heterogenous processes. The models can be linear or non-linear; correctly specified or misspecified. Unlike Vuong (1989), we focus on M-estimators (with MLE and OLS being special cases).[1] This provides a powerful tool for model selection under general conditions. Since Rivers and Vuong (2002, RV henceforth) developed independently a similar analysis for non-nested models, here we focus on the compari-

[1] As it is well-known, in the presence of possibly misspecified models it is not meaningful to compare models according to a GMM objective function. See Hall and Pelletier (2007) for a discussion, and a critique to the approach in Rivers and Vuong (2002). We therefore do not consider GMM in this paper.

son of nested and overlapping models only. Our results could be applied, for example, in model selection problems in macroeconomic and financial data (e.g. selecting a model between two competing nested DSGE models or selecting a model between two competing nested GARCH models), which involve time series data for which the i.i.d. assumption in Vuong (1989) is way too restrictive.

Several other results are obtained as a by-product. In particular, we obtain the asymptotic distribution of the LR test under even weaker conditions than those in Vuong (1989), by allowing for dependence and heterogeneity. Moreover, under certain conditions, it is also possible to test for equality of the information criteria (such as Akaike's (1973) AIC) associated with the models. In fact, a common approach to model selection is to adopt the model that optimizes a penalized likelihood function (see Sin and White, 1996); if the penalty function diverges to infinity at a sufficiently slow rate, then the asymptotic distribution of the information criteria follows directly from our results. Finally, conditions for weak consistency of information criteria based on MLE estimators are also derived.

The paper builds upon the econometric literature on estimation, specification and inference for possibly misspecified non-linear models of dependent and heterogenous processes, see in particular White (1984), Gallant (1987), Gallant and White (1988), White (1994), and the references therein. Section 2 provides a heuristic discussion of the main point of the paper, Section 3 introduces assumptions and definitions, and derives a set of preliminary results. The main results are presented in Section

4, where the asymptotic distribution of the statistic is derived, and our approach is compared with model selection on the basis of information criteria. In Section 5, we use Monte Carlo simulations to assess the finite sample performance of the statistic when applied for the comparison of two AR models of different lag length, a common case in practice. In Section 6, as additional analytical examples, we apply the method for the comparison of two nested ARMAX-GARCH models or overlapping ARMAX-STAR models. Section 7 concludes. All the proofs are gathered in the Appendix, while the notation is summarized in Table 1.

2. Heuristics: the covariance stationary case

For simplicity and to gain intuition, in this section only, *let the data be stationary and ergodic* (an assumption that will be relaxed later); this also provides a useful benchmark for economic applications. We are interested in comparing two parametric models (indexed by j, $j = 1, 2$) with parameters θ_j^* estimated by extremum estimators based on the loss functions $Q_{jn}(\theta_j) = n^{-1}\sum_{t=1}^{n} q_{jt}(\theta_j)$; also, let θ^* denote $[\theta_1^*, \theta_2^*]'$ and $\widehat{\theta}_n$ denote their estimates $[\widehat{\theta}_{1n}, \widehat{\theta}_{2n}]'$. Let $Q_{jn}(\theta_j)$ be the log-likelihood function, and $Eq_{jt}(\theta_j)$ denote the expected value of $q_{jt}(\theta_j)$. Whether the models are correctly specified or not, one would often like to select the model that minimizes the probability limit of the loss function, $Q_j(\theta_j) = E\left(n^{-1}\sum_{t=1}^{n} q_{jt}(\theta_j)\right)$. We thus focus on testing whether:

$$SC(\theta_1^*, \theta_2^*) \equiv E\left[Q_{1n}(\theta_1^*) - Q_{2n}(\theta_2^*)\right] = 0 \tag{1}$$

which means that the two models are equally good. Otherwise, we propose the following rule for model selection: Choose Model 1 if $SC(\theta_1^*, \theta_2^*) < 0$ and Model 2 if $SC(\theta_1^*, \theta_2^*) > 0$. The test statistic is based on the sample counterpart of $SC(\theta_1^*, \theta_2^*)$, $SC_n(\widehat{\theta}_1, \widehat{\theta}_2)$, where $\widehat{\theta}_j$ is an estimate of θ_j^*. The first contribution of this paper is to provide a test statistic that can be used to test (1) and which is robust to serial correlation as well as heterogeneous data. For covariance stationary data, our main result is that, for nested models, under $H_0^* : nSC(\theta_1^*, \theta_2^*) = 0$,

$$-2n\, SC_n(\widehat{\theta}_1, \widehat{\theta}_2) \xrightarrow{D} M_{p+q}(\cdot; \lambda),$$

where $M_{p+q}(., \lambda)$ is a mixture of chi-squares distribution, λ is the vector of eigenvalues of $W\Sigma$, W is a block diagonal matrix with $A_1^* \equiv \nabla_\theta^2 EQ_{1n}(\theta_1^*)$ and $-A_2^* \equiv -\nabla_\theta^2 EQ_{2n}(\theta_2^*)$ on the main diagonal, Σ is the asymptotic variance of $\widehat{\theta}$, and p and q are, respectively, the number of parameters in the models 1 and 2, where $p \geq q$. Σ and W can be consistently estimated by Newey and West's (1987) HAC estimator.

As we will show, by using Mean Value expansions, $SC_n(\widehat{\theta}_1, \widehat{\theta}_2)$ can be written as the sum of components that have different rates of convergence:

$$\begin{aligned} SC_n(\widehat{\theta}_1, \widehat{\theta}_2) \quad = \quad & SC(\theta_1^*, \theta_2^*) + \\ & + n^{-1} \sum_{t=1}^{n} \left[\{ q_{1t}(\theta_1^*) - E(q_{1t}(\theta_1^*)) \} - \{ q_{2t}(\theta_2^*) - E(q_{2t}(\theta_2^*)) \} \right] \quad (2) \\ & + \frac{1}{2} \left(\widehat{\theta} - \theta^* \right)' W \left(\widehat{\theta} - \theta^* \right) + o_p\left(\frac{1}{n} \right). \end{aligned}$$

Under the null hypothesis $SC(\theta_1^*, \theta_2^*) = 0$, only the second and third terms will be asymptotically relevant. The distribution of $\sqrt{n}SC_n(\widehat{\theta}_1, \widehat{\theta}_2)$ will crucially depend on the asymptotic variance of the second term, $Var\left(n^{-1/2}\sum_{t=1}^{n}\left(q_{1t}(\theta_1^*) - q_{2t}(\theta_2^*)\right)\right)$, call it σ_n^2. If $\lim_{n\to\infty} n\sigma_n^2 = 0$ is zero,[2] then the asymptotic distribution will be driven by the last term in (2), and the asymptotic distribution is a mixture of chi-square distributions (this happens, for example, when the models are nested); otherwise it will be driven by the second term, and the asymptotic distribution is normal, as in Rivers and Vuong (2002).[3] To discern between the two possible limiting distribution, we propose a test for $\lim_{n\to\infty} n\sigma_n^2 = 0$ that is robust to serially correlated and heterogeneous data.

Our test procedure can be interpreted as an alternative way to perform tests based on Information Criteria (IC). IC are subject to sampling variability, and thus a model with a smaller IC loss value may not significantly outperform its competitor. Our procedure allows us to test whether two models yield IC values that are not statistically different from each other.

3. Assumptions, definitions and preliminary results

In this section we briefly state the main assumptions. They are similar to those in Gallant and White (1988), to whom we refer for further details. We also introduce some definitions and derive preliminary results to be used in the subsequent analysis.

---

[2] When data are stationary, as in this example, the condition $\lim_{n\to\infty} n\sigma_n^2 = 0$ is equivalent to $\sigma_n^2 = 0$, as the variance does not depend on $n : \sigma_n^2 = \sigma^2$. However, when data are heterogeneous, we will need a stronger condition, $\lim_{n\to\infty} n\sigma_n^2 = 0$, which is the one we will consider throughout the main paper.

[3] When the data are not stationary, then $\lim_{n\to\infty} n\sigma_n^2 \neq 0$ does not necessarily imply that the distribution in RV should be used. See Section 3 for more details.

Assumption 1 (Data Generation)

Let (Ω, F, P) be a complete probability space. The stochastic process $\{V_t\}$, with generic element $V_t : \Omega \rightarrow R^v$, $v \in N$, is a uniform mixing sequence with ϕ_m of size $-r/(r-1)$, $r \geq 2$, or a strong mixing sequence with α_m of size $-2r/(r-2)$, $r > 2$. The observed data are realizations of the stochastic process $\{X_t\}$, with generic element $X_t : \Omega \rightarrow R^{w_t}$, $w_t \in N$, and

$$X_t(\omega) = W_t(..., V_{t-1}(\omega), V_t(\omega), V_{t+1}(\omega), ...), \quad \omega \in \Omega, \tag{3}$$

where $W_t : \times_{\tau=-\infty}^{\infty} R^v \rightarrow R^{w_t}$ are such that X_t is measurable-$F/B(R^{w_t})$, $t = 0, \pm1, \pm2, ...$. ■

Mixing conditions weakly limit the memory of a process, allowing for considerable dependence and heterogeneity (e.g., White (1984)). In particular, the autocovariance function, γ_k, decreases as a power of k (White and Domowitz (1984, Lemma 2.2)), i.e., more slowly than for a finite ARMA process, whose autocovariance function decays exponentially. Yet, trending or explosive behavior is ruled out.

Measurable functions of a finite number of elements of a mixing process are still mixing, and of the same size as the argument, so that in this case $X_t(\omega)$ is also $\phi-$mixing of size $-r/(r-1)$, or $\alpha-$mixing of size $-2r/(r-2)$. When instead X_t can depend on an infinite number of elements of a mixing process, as in (3), it is necessary to restrict somewhat the dependence on distant elements. Gallant and

White (1988, Lemma 3.18) formulate proper near epoch dependence conditions that guarantee that X_t is a mixingale (McLeish (1975)). The general formulation for X_t in (3) allows us to deal with several types of data, including time series, cross-sections, and limited dependent variables, possibly aggregated, seasonally adjusted or subject to other kinds of transformations.

Definition 1 (Model)

A model is a family of parametric probability distributions $D = \{D_\theta : \theta \in \Theta\}$*, defined on the measurable space* $(R^w_\infty, B(R^w_\infty))$*, with* $R^w_\infty = \times_{\tau=-\infty}^{\infty} R^{w_\tau}$*.* ■

More commonly, models are specified by making a parametric hypothesis on the process generating X_t, such as $X_t = W_t(\cdot) = S_t(\cdot;\beta)$, $\beta \in B$. In this case, we have $D_\beta(A) = P\{\omega : \{S_t(..., V_{t-1}(\omega), V_t(\omega), V_{t+1}(\omega), ...; \beta)\} \in A\}$, for $\forall A \in \mathcal{B}(R^w_\infty)$. Often, a model is also defined by means of a finite dimensional density function, $d_n(X_1, ..., X_n; \theta) = \partial D_{n\theta}/\partial \mu_n$, where $D_{n\theta}(A) = D_\theta((X_1, ..., X_n) \in A)$ and $D_{n\theta}$ is absolutely continuous with respect to the $\sigma-$finite measure μ_n for all θ in Θ. A model is correctly specified when there exists $\theta_0 \in \Theta$ such that $D_{\theta_0}(A) = P\{\omega : \{X_t\} \in A\}$, $\forall A \in \mathcal{B}(R^w_\infty)$.

Once D is specified, MLE techniques can be adopted to estimate the parameters of the model, θ. Also, the investigator may be unwilling to make an explicit distributional assumption, and prefer to adopt an estimation method such as OLS or NLS. These three alternative estimation methods, and others, all belong to the class of M-estimators, i.e., estimators that are obtained as the optimand of an objective function. To establish the asymptotic properties of these estimators and related testing procedures, we have to impose suitable regularity conditions on the objective function that let laws of large numbers and central limit theorems for mixing processes or mixingales to be applied, e.g. White (1984).

Assumption 2 (Objective function)

The objective function to be minimized, $Q_n : \Omega \times \Theta \rightarrow \mathbb{R}$, is defined as

$$Q_n(\omega, \theta) \equiv n^{-1} \sum_{t=1}^{n} q_t(\omega, \theta), \quad n = 1, 2, ..., \tag{4}$$

and

i) Θ *is a compact subset of* $\mathbb{R}^k$.

ii) $q_t : \Omega \times \Theta \rightarrow \mathbb{R}^l$ *is a random function continuously differentiable of order 2 on* Θ *a.s. (almost surely-P),* $t = 1, 2, ...$.

iii) *(a)* $\{q_t(\theta)\}$, *(b)* $\{\nabla_\theta q_t(\theta)\}$, *(c)* $\{\nabla^2_\theta q_t(\theta)\}$ *are a.s. Lipschitz-*L_1.

iv) *The elements of (a)* $\{q_t(\theta)\}$, *(b)* $\{\nabla_\theta q_t(\theta)\}$, *(c)* $\{\nabla^2_\theta q_t(\theta)\}$ *are uniformly near epoch dependent on* $\{V_t\}$ *of size* -1 *on* (Θ, ρ), *where* ρ *is any convenient norm on* $\mathbb{R}^k$.

v) *The elements of (a)* $\{q_t(\theta)\}$, *(b)* $\{\nabla_\theta q_t(\theta)\}$, *(c)* $\{\nabla^2_\theta q_t(\theta)\}$ *are* $r-$*dominated on* Θ *uniformly in* $t = 1, 2, ...,$ $r > 2$.

vi) *The sequence* $\{\overline{Q}_n(\theta)\}=\{n^{-1}\sum_{t=1}^{n} E(q_t(\theta))\}$ *has identifiably unique minimizers* $\{\theta_n^*\}$ *on* Θ*, interior to* Θ *uniformly in* n*.*

vii) *(a)* $\{B_n^*\} = \{Var[n^{1/2}\nabla_\theta Q_n(\theta_n^*)]\}$*, (b)* $\{A_n^*\} = \{\nabla_\theta^2 \overline{Q}_n(\theta_n^*)\}$ *is* $O(1)$ *and uniformly non singular.* ■

The conditions i) and ii) are sufficient (not necessary) to ensure the existence of $\{\widehat{\theta}_n\}$, with $\widehat{\theta}_n : \Omega \to \Theta$, such that

$$Q_n(\omega, \widehat{\theta}_n(\omega)) = \inf_{\theta \in \Theta} Q_n(\omega, \theta), \quad a.s.$$

The conditions iii)-a), iv)-a), and v)-a) impose, respectively, smoothness, memory, and moment conditions on $q_t(\theta)$ to ensure that $Q_n(\theta) - \overline{Q}_n(\theta) \to 0$ a.s. uniformly in Θ (Gallant and White (1988, Theorem 3.18)). From i) it also follows that $Q_n(\theta) - \overline{Q}_n(\theta) \to 0$ a.s. uniformly in Θ and, under the additional condition vi), $\widehat{\theta}_n - \theta_n^* \to 0$ a.s. (Gallant and White (1988, Theorem 3.19)), i.e., the estimator is consistent for θ_n^*.

Under the additional conditions iii)-b), iv)-b), v)-b), and vii)-a), the asymptotic distribution of $B_n^{*-1/2} n^{1/2} \nabla_\theta Q_n(\theta_n^*)'$ is $N(0, I_k)$ (Gallant and White (1988, Corollary 5.5)).

With the further conditions on the matrix of second derivatives of the objective function in iii)-c), iv)-c, v)-c and vii)-b), from a mean value expansion of $\nabla_\theta Q_n(\widehat{\theta}_n)$ around θ_n^*, it follows that the asymptotic distribution of $B_n^{*-1/2} A_n^* n^{1/2}(\widehat{\theta}_n - \theta_n^*)$ is

$N(0, I_k)$ (Gallant and White (1988, Theorem 5.7)). The conditions in vii)-c) and d) will be used in the next section.

Finally, note that the evaluation function coincides with the objective function used for estimating the parameters. This assumption allows for a variety of in-sample model comparisons, but may prevent out-of-sample forecast model comparisons.

Definition 2 (Competing models)

The two competing models are $D_1 = \{D_{\theta_1} : \theta_1 \in \Theta_1\}$ *and* $D_2 = \{D_{\theta_2} : \theta_2 \in \Theta_2\}$, *where* Θ_1 *and* Θ_2 *are compact subsets of* $\mathbb{R}^p$ *and* $\mathbb{R}^q$*, respectively, with* $p \geq q$*.* D_1 *and* D_2 *are defined on the same measurable space (*$\mathbb{R}^w_\infty$*,*$B(\mathbb{R}^w_\infty)$*). Following Vuong (1989) and RV, we say that*

i) D_2 *is nested in* D_1 *if there exists a sequence of continuously differentiable functions,* $\phi_n : \Theta_2 \to \Theta_1$*, such that, for* n *sufficiently large,*

$$D_2(A; \theta_2) = D_1(A; \phi_n(\theta_2), \forall \theta_2 \in \Theta_2, \forall A \in \mathcal{B}(\mathbb{R}^w_\infty)). \tag{5}$$

ii) D_1 *and* D_2 *are overlapping if (5) holds for some* $\theta_2 \in \Theta_2$*.*

iii) D_1 *and* D_2 *are strictly non-nested if (5) never holds.* ■

Note that i) implies that D_2 is nested in D_1 when, for n sufficiently large, any distribution in D_2 is also in D_1, and the function ϕ_n provides the mapping from Θ_2 to Θ_1. The models overlap when there exist common distributions, but neither model is nested in the other. And the models are strictly non-nested when there is no distribution common to D_1 and D_2.

To simplify notation, in what follows we will ignore the dependence on ω, and distinguish the two models by using subscripts $j = 1, 2$.

Assumption 3 (Selection Criterion)

The criterion to evaluate the model D_1 against the model D_2 is

$$\begin{aligned} SC_n(\widehat{\theta}_{1n}, \widehat{\theta}_{2n}) &\equiv Q_{1n}(\widehat{\theta}_{1n}) - Q_{2n}(\widehat{\theta}_{2n}) \\ &= \left(n^{-1} \sum_{t=1}^{n} q_{1t}(\widehat{\theta}_{1n}) \right) - \left(n^{-1} \sum_{t=1}^{n} q_{2t}(\widehat{\theta}_{2n}) \right), \end{aligned} \tag{6}$$

where $q_{1t} : \Omega \times \Theta_1 \to \mathbb{R}^{l_1}$, $q_{2t} : \Omega \times \Theta_2 \to \mathbb{R}^{l_2}$. The conditions in Assumption 2 are satisfied for Q_{1t} and Q_{2t}. ■

Lemma 1 (Convergence of $SC_n(\widehat{\theta}_{1n}, \widehat{\theta}_{2n})$)

Given Assumptions 1, 2 i), ii), iii)-a), iv)-a), v)-a) and vi),

$$SC_n(\widehat{\theta}_{1n}, \widehat{\theta}_{2n}) - (\overline{Q}_{1n}(\theta^*_{1n}) - \overline{Q}_{2n}(\theta^*_{2n})) \xrightarrow{a.s.} 0,$$

where $\overline{Q}_{in}(\theta^*_{in}) = n^{-1} \sum_{t=1}^{n} E(q_{it}(\theta^*_{in}))$, $i = 1, 2$. ■

Next, we present a set of results on estimators, to be used in the derivation of the tests for model selection in the next section.

Lemma 2 (Joint Distribution of $\widehat{\theta}_{1n}$ and $\widehat{\theta}_{2n}$)

Given Assumptions 1 and 2 for D_1 and D_2,

$$n^{1/2}\begin{pmatrix}\widehat{\theta}_{1n}-\theta^*_{1n}\\ \\ \widehat{\theta}_{2n}-\theta^*_{2n}\end{pmatrix}\xrightarrow{D} N(0,\Sigma),$$

where $\Sigma \equiv \lim_{n\to\infty}\Sigma_n$ exists, and

$$\Sigma_n=\begin{pmatrix} A^{-1}_{1n}(\theta^*_{1n})B_{1n}(\theta^*_{1n})A^{-1}_{1n}(\theta^*_{1n}) & A^{-1}_{1n}(\theta^*_{1n})B_{12n}(\theta^*_{1n},\theta^*_{2n})A^{-1}_{2n}(\theta^*_{2n}) \\ \\ A^{-1}_{2n}(\theta^*_{2n})B_{21n}(\theta^*_{2n},\theta^*_{1n})A^{-1}_{1n}(\theta^*_{1n}) & A^{-1}_{1n}(\theta^*_{2n})B_{2n}(\theta^*_{2n})A^{-1}_{2n}(\theta^*_{2n})\end{pmatrix}, \tag{7}$$

$$\begin{aligned}A_{in}(\theta^*_{in}) &= \nabla^2_\theta\overline{Q}_{in}(\theta^*_{in}),\\ B_{in}(\theta^*_{in}) &= Var[n^{1/2}\nabla_\theta Q_{in}(\theta^*_{in})],\\ B_{ijn}(\theta^*_{in},\theta^*_{jn}) &= B_{jin}(\theta^*_{jn},\theta^*_{in})' = cov[n^{1/2}\nabla_\theta Q_{in}(\theta^*_{in}), n^{1/2}\nabla_\theta Q_{jn}(\theta^*_{jn})],\ for\ i\neq j.\ \blacksquare\end{aligned}$$

Lemma 3 (Distribution of quadratic forms in $\widehat{\theta}_{1n}$ and $\widehat{\theta}_{2n}$)

Let Q be a $(p+q)\times(p+q)$ real symmetric matrix and $Y=\sqrt{n}\left[(\widehat{\theta}_{1n}-\theta^*_{1n})',(\widehat{\theta}_{2n}-\theta^*_{2n})'\right]'$. Then, the asymptotic distribution function of $Y'QY$ is $M_{p+q}(\cdot;\lambda)$, where $M_{p+q}(\cdot;\lambda)$ indicates the distribution function of the weighted sum of $p+q$ central chi-square random variables, with weights λ given by the eigenvalues of $Q\Sigma$. $\blacksquare$

Define

$$\sigma^2_n \equiv R^*\Omega^*_n R^{*\prime}, \tag{8}$$

where $R^* \equiv [1, -1]$, and $\Omega_n^* \equiv var\left(\sqrt{n}\left[\left(Q_{1n}(\theta_{1n}^*) - \overline{Q}_{1n}(\theta_{1n}^*)\right), \left(Q_{2n}(\theta_{2n}^*) - \overline{Q}_{2n}(\theta_{2n}^*)\right)\right]'\right)$.

Then:

Lemma 4 (Distribution of $Q_{1n}(\theta_{1n}^*) - Q_{2n}(\theta_{2n}^*)$)

Given Assumptions 1 and 2 for D_1 and D_2 and $\liminf_{n\to\infty} \sigma_n^2 > 0$,

$$\sqrt{n}\,\sigma_n^{-1}\left[Q_{1n}(\theta_{1n}^*) - Q_{2n}(\theta_{2n}^*) - \left(\overline{Q}_{1n}(\theta_{1n}^*) - \overline{Q}_{2n}(\theta_{2n}^*)\right)\right] \xrightarrow{D} N(0,1). \blacksquare \qquad (9)$$

Finally, we have to deal with estimation of σ_n^2 in (8), of the variance covariance matrix Σ_n in (7), and of A_n^* and B_n^* in Assumption 2)-vii) that will be used in the next section. We have,

Lemma 5 (Estimation of $A_{in}(\theta_{in}^*)$)

Given Assumptions 1 and 2 for D_1 and D_2, we have $\widehat{A}_{in} - A_{in}(\theta_{in}^*) \xrightarrow{a.s.} 0$, for $i = 1, 2$, where $\widehat{A}_{in} = \nabla_\theta^2 Q_{in}(\widehat{\theta}_{in})$. $\blacksquare$

Estimation of $B_{in}(\theta_{in}^*)$, $B_{ijn}(\theta_{in}^*, \theta_{jn}^*)$, and σ_n^2 is more complex. Let us focus on $B_{in}(\theta_{in}^*)$. Following Gallant and White (1988, Theorem 5.4), we have

$$\{B_{in}^*\} = \{Var[n^{1/2}\nabla_\theta Q_{in}(\theta_{in}^*)]\} = \{Var[n^{-1/2}\sum_{t=1}^{n} M_{int}^*]\}, \qquad (10)$$

where

$$\begin{aligned} M^*_{int} &= S^*_{int} - E(S^*_{int}), \\ S^*_{int} &= \nabla_\theta Q_{it}(\theta^*_{in})'. \end{aligned}$$

If we expand the expression in (10), we get

$$\{B^*_{in}\} = n^{-1}\sum_{t=1}^{n} E(M^*_{int}M^{*'}_{int}) + n^{-1}\sum_{\tau=1}^{n-1}\sum_{t=\tau+1}^{n}\left[E(M^*_{int}M^{*'}_{in(t-\tau)}) + E(M^*_{in(t-\tau)}M^{*'}_{int})\right]. \tag{11}$$

The term $E(S^*_{int})$ is generally unknown, so that an estimator for B^*_{in} can be formulated as

$$\widehat{B}_{in} = w_{n0}n^{-1}\sum_{t=1}^{n}\widehat{S}_{int}\widehat{S}'_{int} + n^{-1}\sum_{\tau=1}^{m_n} w_{n\tau}\sum_{t=\tau+1}^{n}\left[\widehat{S}_{int}\widehat{S}'_{in(t-\tau)} + \widehat{S}_{in(t-\tau)}\widehat{S}'_{int}\right], \tag{12}$$

where $\widehat{S}_{int} = S_{int}(\widehat{\theta}_{in})$. In order to prove convergence of $\widehat{B}_{in}$, the memory and moment requirements in Assumption 2 have to be strengthened, and requirements on the truncation lag (m_n) and the weights ($w_{n\tau}$) added. We have,

Assumption 4 (Convergence of $\widehat{B}_{in}$)

i) *The elements of (a) $\{q_t(\theta)\}$ and (b) $\{\nabla_\theta q_t(\theta)\}$ are near epoch dependent on $\{V_t\}$ of size $-2(r-1)/(r-2)$ uniformly on (Θ,ρ). The elements of $\{\nabla^2_\theta q_t(\theta)\}$ are near epoch dependent on $\{V_t\}$ of size $-2r(r-1)/(r-2)$ uniformly on (Θ,ρ).*

ii) *The elements of (a)* $\{q_t(\theta)\}$*, (b)* $\{\nabla_\theta q_t(\theta)\}$*, (c)* $\{\nabla^2_\theta q_t(\theta)\}$ *are* $2r-$*dominated on* Θ *uniformly in* $t = 1, 2, ..., r > 2$.

iii) $\{m_n\}$ *is a sequence of integers such that* m_n *is* $o(n^{1/4})$.[4]

iv) *The weights are* $w_{n\tau} = \sum_{\lambda=\tau+1}^{m_n+1} a_{n\lambda}a_{n\lambda-\tau}$*, where* $\{a_{n\lambda}\}$*,* $n = 1, 2, ...,$ $\lambda = 1, 2, ..., m_n + 1$ *is any triangular array such that* $|w_{n\tau}| \leq \Delta < \infty$*,* $n = 1, 2, ...,$*,* $\tau = 0, 1, 2, ..., m_n$*, and for each* τ*,* $w_{n\tau} \to 1$ *as* $n \to \infty$.

v) $E(S^*_{nt}) = 0$.[5] ■

Defining

$$\begin{aligned} U^*_{in} &= w_{n0}n^{-1}\sum_{t=1}^{n} E(S^*_{int})E(S^{*'}_{int}) + \\ & n^{-1}\sum_{\tau=1}^{m_n} w_{n\tau} \sum_{t=\tau+1}^{n} \left[E(S^*_{int})E(S^{*'}_{in(t-\tau)}) + E(S^*_{in(t-\tau)})E(S^{*'}_{int})\right], \end{aligned}$$

we have:

Lemma 6 (Convergence of $\widehat{B}_{in}$)

Under the conditions in Assumptions 1, 2 and 4-i) to 4-iv) for D_1 *and* D_2*,* $\widehat{B}_{in}$ *and* U^*_{in} *are positive definite for all* n*, and* $\widehat{B}_{in} - (B^*_{in} + U^*_{in}) \xrightarrow{p} 0$*. When 4-v) also holds,* $\widehat{B}_{in} - B^*_{in} \xrightarrow{p} 0$. ■

---

[4] *These assumptions rely on Gallant and White (1988); such assumptions could be relaxed, see De Jong and Davidson (2000).*

[5] Assumption 4(v) is a high-level assumption that ensures consistent estimation of the asymptotic variance. It could be relaxed following RV, who exploit the fact that consistency is needed only under the null hypothesis. Such weaker assumptions, however, may influence the power of the test.

The additional condition $E(S^*_{nt}) = 0$ is usually imposed, e.g. Newey and West (1987a, Theorem 2), so that $\widehat{B}_{in}$ is consistent. Yet, for dynamic misspecified models it can be $E(S^*_{nt}) \neq 0$. As pointed out by Gallant and White (1988, p. 102), sufficient conditions for $E(S^*_{nt}) = 0$ are that either $\{X_t\}$ is strictly stationary or that the model is correctly specified. Notice also that when (M^*_{nt}, F^t) is a martingale difference sequence, $B^*_{in} = n^{-1} \sum_{t=1}^{n} E(M^*_{int} M^{*\prime}_{int})$, which simplifies estimation. Estimators for B_{ijn}, and for σ^2_n, are obtained along the same lines.[6]

4. Tests for model selection

Following RV, we formulate the hypothesis of interest when comparing D_1 and D_2 as

$$H_0 : \lim_{n\to\infty} \left(\overline{Q}_{1n}(\theta^*_{1n}) - \overline{Q}_{2n}(\theta^*_{2n})\right) = 0, \tag{13}$$

Under H_0 the difference of the loss functions from D_1 and D_2 vanishes when evaluated at the expected value of $Q_{jn}(\theta^*_{jn})$ and for n sufficiently large, and in this sense the two models are asymptotically equivalent. Under the additional hypothesis of i.i.d. observations, the null hypothesis simplifies to $E_P(q_1(\theta^*_1)) = E_P(q_2(\theta^*_2))$. With the further assumption of MLE estimation, H_0 is equivalent to the null hypothesis in Vuong (1989), and our results boil down to his.

[6] De Jong and Davidson (2000) further relaxed the conditions for consistency of the estimators.

The alternative hypothesis is either

$$H_1: \limsup_{n\to\infty} \left(\overline{Q}_{1n}(\theta^*_{1n}) - \overline{Q}_{2n}(\theta^*_{2n})\right) < 0, \tag{14}$$

or

$$H_2: \liminf_{n\to\infty} \left(\overline{Q}_{1n}(\theta^*_{1n}) - \overline{Q}_{2n}(\theta^*_{2n})\right) > 0. \tag{15}$$

Under H_1, D_1 is preferred to D_2, since $\overline{Q}_{1n}(\theta^*_{1n}) < \overline{Q}_{2n}(\theta^*_{2n})$ for n sufficiently large, and vice versa under H_2.

According to Lemma 1, the model selection criterion $SC_n(\widehat{\theta}_{1n}, \widehat{\theta}_{2n})$ in (6) converges to $\overline{Q}_{1n}(\theta^*_{1n}) - \overline{Q}_{2n}(\theta^*_{2n})$. Hence, it is a natural candidate as a statistic to test for H_0 against H_1 or H_2. Actually, it was used by RV to compare strictly non-nested models.

In the first two subsections we focus on the derivation of the asymptotic distribution of $SC_n(\widehat{\theta}_{1n}, \widehat{\theta}_{2n})$ for, respectively, nested and overlapping models. In the final subsection we compare our tests with model selection based on information criteria.

4.1. Nested models. To evaluate nested models, we have

Theorem 1 (Asymptotic Distribution of $SC_n(\widehat{\theta}_{1n}, \widehat{\theta}_{2n})$)

Given Assumptions 1-4 for D_1 and D_2, if $\lim_{n\to\infty} n\sigma_n^2 = 0$:

i) *Under H_0^* : $\lim_{n\to\infty}\ n\left(\overline{Q}_{1n}(\theta_{1n}^*) - \overline{Q}_{2n}(\theta_{2n}^*)\right) = 0$,*

$$-2n\, SC_n(\widehat{\theta}_{1n}, \widehat{\theta}_{2n}) \xrightarrow{D} M_{p+q}(\cdot;\lambda), \tag{16}$$

where λ is the vector of eigenvalues of $W\Sigma$, $W \equiv \lim_{n\to\infty} W_n$ exists, and

$$W_n = \begin{pmatrix} A_{1n}(\theta_{1n}^*) & 0 \\ 0 & -A_{2n}(\theta_{2n}^*) \end{pmatrix};$$

ii) *Under H_1^* : $\lim_{n\to\infty}\ n\left(\overline{Q}_{1n}(\theta_{1n}^*) - \overline{Q}_{2n}(\theta_{2n}^*)\right) = -\infty$, $-2nSC_n(\widehat{\theta}_{1n}, \widehat{\theta}_{2n}) \xrightarrow{p} \infty$.* ■

Remarks. 1) Condition $\lim_{n\to\infty} n\sigma_n^2 \to 0$ is implied by the models being nested, in which case $\sigma_n^2 = 0$ identically for all n. When such condition is not verified, the asymptotic distribution (appropriately rescaled) is asymptotically normal, which is the case considered by RV. 2) Note that, as in RV, H_0 does not necessarily imply H_0^*, so that the test has the correct asymptotic size on a subset of the null hypothesis H_0. Conversely, since H_1^* does not necessarily imply H_1, the test is consistent against a larger set of alternatives than H_1. Under strict stationarity of both models, $H_0 = H_0^*$

and $H_1 = H_1^*$. 3) The test is one sided because, when D_2 is nested in D_1, H_2 cannot hold. Moreover, the speed of convergence is higher than for non-nested models, n versus $\sqrt{n}$, see RV's Theorem 1 for comparison.

The eigenvalues of $\widehat{W}_n \widehat{\Sigma}_n$ provide consistent estimators for λ (Davies (1980)). When the models are estimated by MLE, following Vuong (1989, Theorem 7.2) it can be shown that the non-zero eigenvalues of $W_n \Sigma_n$ coincide with those of the lower dimensional matrix $B_{1n}(\theta_{1n}^*)[(\partial\phi(\theta_2)/\partial\theta_2^{'})A_{2n}^{-1}(\theta_{2n}^*)(\partial\phi^{'}(\theta_2)/\partial\theta_2) - A_{1n}^{-1}(\theta_{1n}^*)]$, which simplifies the calculations.

Theorem 1 nests and extends several previous results in the literature. In particular, under MLE estimation and when the models are defined as in Smith (1994, p.4), the distribution in (16) coincides with that in Smith (1994, Corollary 3.1). With the additional hypothesis that the information matrix equivalence holds for the nesting model D_1 $(A_{1n}(\theta_{1n}^*) = -B_{1n}(\theta_{in}^*))$, it can be shown that the limiting distribution reduces to χ_{p-q}^2, which coincides with the distribution in Gallant and White (1988, Theorem 7.8), see also Gallant (1989, Chapter 7).[7] Gourieroux and Monfort (1989, Property 3) show that a similar finding holds when testing for a null hypothesis in mixed form.

To conclude, it is worth pointing out that when Assumption 4-v) is not satisfied $(E(S_{nt}^*) \neq 0)$, $\widehat{\sigma}_n^2$ and $\widehat{\Sigma}_n$ overestimate σ_n^2 and Σ_n (Lemma 6). In this case the

[7] From Rao and Mitra (1971, Theorem 9.2.1) a necessary and sufficient condition for the asymptotic distribution of $-2nSC_n(\widehat{\theta}_{1n}, \widehat{\theta}_{2n})$ to be chi-square when the information matrix equivalence is not satisfied is $\lim_{n\to\infty}(\Sigma_n W_n \Sigma_n W_n \Sigma_n - \Sigma_n W_n \Sigma_n) = 0$.

asymptotic size of the test will be lower than the nominal size, i.e., the test is biased towards acceptance of the null hypothesis.

4.2. Overlapping models. For the comparison of overlapping models, we have:

Corollary 1 (Overlapping models)

Given Assumptions 1-4 for D_1 and D_2,

i) *If* $\lim_{n\to\infty}\sigma_n^2 > 0$, *then under* $H_0^{**}: \lim_{n\to\infty} \sqrt{n}\left(\overline{Q}_{1n}(\theta_{1n}^*) - \overline{Q}_{2n}(\theta_{2n}^*)\right) = 0$, $\sqrt{n}\widehat{\sigma}_n^{-1}SC_n(\widehat{\theta}_{1n}, \widehat{\theta}_{2n}) \xrightarrow{D} N(0,1)$; *under* $H_1^{**}: \lim_{n\to\infty} \sqrt{n}\left(\overline{Q}_{1n}(\theta_{1n}^*) - \overline{Q}_{2n}(\theta_{2n}^*)\right) = -\infty$, $\sqrt{n}\widehat{\sigma}_n^{-1}SC_n(\widehat{\theta}_{1n}, \widehat{\theta}_{2n}) \xrightarrow{p} -\infty$; *under* $H_2^{**}: \lim_{n\to\infty} \sqrt{n}\left(\overline{Q}_{1n}(\theta_{1n}^*) - \overline{Q}_{2n}(\theta_{2n}^*)\right) = +\infty$, $\sqrt{n}\widehat{\sigma}_n^{-1}SC_n(\widehat{\theta}_{1n}, \widehat{\theta}_{2n}) \xrightarrow{p} +\infty$.

ii) *If* $\lim_{n\to\infty} n\sigma_n^2 = 0$ *then, under* H_0^*, *for any* $z \geq 0$, $\lim_{n\to\infty}[Pr\{-2nSC_n(\widehat{\theta}_{1n}, \widehat{\theta}_{2n}) \leq z\}$
$-M_{p+q}(z; \widehat{\lambda})] \longrightarrow 0$; *under* H_1^*, $-2nSC_n(\widehat{\theta}_{1n}, \widehat{\theta}_{2n}) \xrightarrow{p} -\infty$. ■

A sequential testing procedure can be adopted to decide whether the statistic in ii) has to be used. In the first step the hypothesis

$$H_{0\sigma}: \lim_{n\to\infty} n\sigma_n^2 = 0 \tag{17}$$

is tested. If it is accepted, the test for H_0^* is conducted using the statistic in ii).[8]

[8]Note that if $H_{0\sigma}$ is not accepted, then it is not possible to conclude that Rivers and Vuong's (2002) statistic should be used unless the data generating process is stationary. In fact, if $n\sigma_n^2 > 0$ but $\sigma_n^2 = 0$ then, unlike the results in Rivers and Vuong (2002), the asymptotic distribution depends

Vuong (1989, p.321) shows that for the i.i.d. case an asymptotic upper bound for the significance level of this sequential procedure is given by the maximum of the asymptotic significance levels of the σ_n^2 test and the SC_n test, and the result remains true in our more general context.

To test the hypothesis $H_{0\sigma}$, we follow the discussion in Section 2 (compare in particular equation (12)), and define $\widehat{\sigma}_n^2$ as

$$\widehat{\sigma}_n^2 = w_{n0} n^{-1} \sum_{t=1}^{n} \widehat{S}_{nt}^2 + 2n^{-1} \sum_{\tau=1}^{m_n} w_{n\tau} \sum_{t=\tau+1}^{n} \widehat{S}_{nt} \widehat{S}_{n(t-\tau)} \tag{18}$$

where $\widehat{S}_{n(t-\tau)} = q_{1t-\tau}(\widehat{\theta}_{1n}) - q_{2t-\tau}(\widehat{\theta}_{2n})$.

We then have the following theorem:

Theorem 2 (Asymptotic Distribution of $\widehat{\sigma}_n^2$)

Given Assumptions 1-4 for D_1 and D_2, if $\lim_{n\to\infty} n\sigma_n^2 = 0$

i) $\left(\widehat{\sigma}_n^2 - \sigma_n^2\right) \xrightarrow{p} 0$,

ii) *Under* $H_{0\sigma} : \lim_{n\to\infty} n\sigma_n^2 = 0$ *and* H_0^*, *assuming that* $((q_{1t}(\theta_{1n}^*) - q_{2t}(\theta_{2n}^*), F^t)$ *is a martingale difference sequence so that* $\widehat{\sigma}_n^2$ *simplifies to* $\widehat{\sigma}_n^2 = n^{-1} \sum_{t=1}^{n} \widehat{S}_{nt}^2$,

$$n\,\widehat{\sigma}_n^2 \xrightarrow{D} M_{p+q}(\cdot;\eta), \tag{19}$$

also on the second order term in the Taylor expansion in the proof of part (ii). Note that if the data are covariance stationary, then $\sigma_n^2 = \sigma^2$ (does not depend on the sample size) and therefore it is not possible that $n\sigma_n^2 > 0$ and $\sigma_n^2 = 0$ hold simultaneously.

where η *is the vector of eigenvalues of* $V\Sigma$*, with* $V = \lim_{n\to\infty} V_n$ *and*

$$V_n = \begin{pmatrix} G_{1n}(\theta^*_{1n}) & G_{12n}(\theta^*_{1n}, \theta^*_{2n}) \\ G_{21n}(\theta^*_{2n}, \theta^*_{1n}) & G_{2n}(\theta^*_{2n}) \end{pmatrix}.$$

Under $H_{1\sigma} : \liminf_{n\to\infty} n\sigma^2_n > 0$*,* $n\,\widehat{\sigma}^2_n \to \infty$*.*

iii) *Under* $H_{0\sigma} : \lim_{n\to\infty} n\sigma^2_n = 0$ *and* H^*_0*,*

$$n\,\widehat{\sigma}^2_n \xrightarrow{D} M_{p+q}(\cdot;\delta_1), \tag{20}$$

where δ_1 *is the vector of eigenvalues of* $W\Sigma$*, with* $W = \lim_{n\to\infty} W_n$*,* $W_n = w_{n0}V_n + \sum_{\tau=1}^{m_n} 2w_{n\tau}V^\tau_n$*, and* V^τ_n *is defined in the Appendix. Under* $H_{1\sigma}$ *:* $\liminf_{n\to\infty} n\sigma^2_n > 0$*, it is* $n\,\widehat{\sigma}^2_n \to \infty$*.* ■

To conclude, notice that when $Q_{in}(\theta_{in})$ is equal to (minus) the likelihood function for model D_i, $i = 1, 2$,[9] the SC_n statistic coincides with the LR test. Hence, we also provide a characterization of the behavior of the LR test under more general conditions than those by Vuong (1989), who analyzed the i.i.d. case.

4.3. Comparison with Information Criteria. A common approach to model selection requires to adopt the model that optimizes a penalized likelihood criterion, e.g., Akaike's (1973) AIC, Schwarz's (1978) SIC, or Hannan and Quinn's (1979)

[9] In this case, conditions similar to those in Vuong (1989, Corollary 4.4) imply an asymptotic χ^2 distribution for the variance test.

HQ. For coherence with the previous analysis, we formulate the problem in terms of minimization of a loss function, so that Model 1 is selected if $IC_n < 0$, where

$$IC_n = \sum_{t=1}^{n} q_{1t}(\widehat{\theta}_{1n}) - \sum_{t=1}^{n} q_{2t}(\widehat{\theta}_{2n}) + \widehat{c}_n, \tag{21}$$

and $q_t(\theta)$ is equal to minus the likelihood function, and hence the term $-n^{-1}(\sum_{t=1}^{n} q_{1t}(\widehat{\theta}_{1n}) - \sum_{t=1}^{n} q_{2t}(\widehat{\theta}_{2n}))$ is an estimate of the average Kullback-Leibler (1951) Information Criterion (KLIC). The penalty function $\widehat{c}_n$ favors the selection of a parsimonious model, and it is typically (but not necessarily) a non stochastic sequence, e.g., $\widehat{c}_n = (p-q)$ for AIC, $\widehat{c}_n = ((p-q)\log n)/2$ for SIC, $\widehat{c}_n = (p-q)c\log(\log n)$ with $c > 1$ for HQ.

The following Corollary gives conditions on $\widehat{c}_n$ for IC_n to select with probability approaching one as the sample size increases either the model with lower average KLIC, or the more parsimonious model when the KLIC is the same (weak consistency of model selection).

Corollary 2 (Weak consistency of IC_n)

Given Assumptions 1-4 for D_1 and D_2,

i) *If* $\liminf_n(\overline{Q}_{1n}(\theta^*_{1n}) - \overline{Q}_{2n}(\theta^*_{2n})) > 0$ *and* $\widehat{c}_n$ *is* $o_p(n)$, *then* $\lim_{n\to\infty} P\{IC_n > 0\} = 1$.

ii) *If* $\limsup_n n^{1/2}(\overline{Q}_{1n}(\theta^*_{1n}) - \overline{Q}_{2n}(\theta^*_{2n})) < \infty$ *and* $P\{n^{-1/2}\widehat{c}_n \to \infty\} = 1$, *then* $\lim_{n\to\infty} P\{IC_n > 0\} = 1$.

iii) *If* $n(\overline{Q}_{1n}(\theta^*_{1n}) - \overline{Q}_{2n}(\theta^*_{1n}))$ *is* $O_p(1)$, $\lim_{n\to\infty} n\sigma^2_n = 0$, *and* $P\{\widehat{c}_n \to \infty\} = 1$, *then* $\lim_{n\to\infty} P\{IC_n > 0\} = 1.$ ■

The conditions in Corollary 2, turn out to be the same as those in Sin and White (1996, Proposition 4.2). For example, 2-i) indicates that when D_2 is preferred to D_1 on the basis of the likelihood function, it will be selected by the information criterion with probability approaching one if the penalty function diverges with a rate slower than n.[10] Note that weak consistency of selection is preserved even if the estimators in (21) are not MLE, as long as the conditions 1-4 are satisfied, and in this respect Corollary 2 extends the results by Sin and White (1996).

It can also be of interest to test for $IC_n = 0$, namely, whether two models yield values for the information criteria that are not statistically different from each other. For example, when IC_n is used to select the lag length of a dynamic model, the competing models are nested and, given that

$$IC_n = nSC_n\left(\widehat{\theta}_{1n}, \widehat{\theta}_{2n}\right) + \widehat{c}_n, \tag{22}$$

if $\widehat{c}_n$ is supposed to be $o_p(1)$, the asymptotic distribution for IC_n under H_0^* immediately follows from Theorem 1. On the other hand, as pointed out by Vuong (1989), IC_n can also be interpreted as a small sample corrected version of the SC_n statistic, where the correction is represented by the penalty function $\widehat{c}_n$.

[10] Sin and White (1996) also present stronger requirements that guarantee strong consistency of model selection, i.e., the model with lower KLIC is selected with probability one when the sample size increases.

5. A Monte Carlo evaluation

To provide a simple evaluation of the finite sample behaviour of our procedure, we assess the performance of the SC_n test for model selection when applied for the comparison of two AR models with different lag length, a common situation where researchers typically use AIC or BIC to select the number of lags. Formally, it is

$$
\begin{aligned}
Q_{in}(\theta_i) &\equiv n^{-1} \sum_{t=1}^{n} q_{it}(\theta_i), \\
q_{it}(\theta_i) &\equiv \frac{1}{2} \log(v_t^2) + \frac{1}{2} \eta_t^2(\theta_i) v_t^{-2}, \\
\eta_t(\theta_i) &\equiv Y_t - \sum_{i=1}^{k_i} b_i Y_{t-i},
\end{aligned}
\tag{23}
$$

$i = 1, 2$, and the SC_n statistic for AR(p) versus AR(q), $p \geq q$, is

$$
SC_n(\widehat{\theta}_{1n}, \widehat{\theta}_{2n}) = Q_{1n}(\widehat{\theta}_1) - Q_{2n}(\widehat{\theta}_2). \tag{24}
$$

For both models, we assume that $\{Y_t\}$ is a strictly stationary and ergodic process. For simplicity, we also maintain all the other assumptions in Sin and White (1996, Section 7), even if some of them could be relaxed. Basically, these assumptions guarantee that both $q_{it}(\theta_i)$ and $\nabla_{\theta_i} q_{it}(\theta_i)$ are near epoch dependent, and satisfy a Central Limit Theorem, for $i = 1, 2$. We can also define $\widehat{\sigma}_n^2$ as in (18). The models are nested. Hence, the asymptotic distribution of the SC_n statistic is that in Theorem 1, and we follow the two-step procedure described in Section 3.2.

First, we test for $\lim_{n\to\infty} n\sigma_n^2 = 0$ using the asymptotic distribution of $\widehat{\sigma}_n^2$ given in Theorem 2. Notice that λ can be estimated by $\widehat{\lambda}$, the vector of eigenvalues of $\widehat{V}_n\widehat{\Sigma}_n$ (Davies (1980)), and estimation of $\widehat{V}_n$ and $\widehat{\Sigma}_n$ is discussed in Section 2.[11]

Second, if the hypothesis $\lim_{n\to\infty} \sigma_n^2 = 0$ is accepted, it is

$$2nSC_n(\widehat{\theta}_{1n}, \widehat{\theta}_{2n}) \xrightarrow{D} M_{p+q}(\cdot; \lambda), \tag{25}$$

where λ is estimated by $\widehat{\lambda}$, the vector of eigenvalues of $\widehat{W}_n\widehat{\Sigma}_n$. Otherwise,

$$\sqrt{n}\widehat{\sigma}^{-1}SC_n(\widehat{\theta}_{1n}, \widehat{\theta}_{2n}) \xrightarrow[H_0]{D} N(0,1). \tag{26}$$

Rejection of the null hypothesis in the second step provides evidence in favor of the AR(p) (AR(q), $q > p$) model when SC_n is negative (positive).

For the simulation experiment, the DGP is an AR process such that $v_t = 1$ and $k = 1$ (for size evaluation) or $k = 2$ (for power evaluation). We consider different values of b_1 and b_2; size results correspond to rows where $b_2 = 0$ and power results to rows where $b_2 \neq 0$. In order to avoid the analysis to be contaminated by the poor small sample properties of heteroskedasticity and autocorrelation covariance estimators, we use $n = 1,000$.

Figure 1 shows the size and power of the tests (16) and (19), and Table 2 reports

[11] For MLE the expression for S^*_{int} in (10) simplifies to $S^*_{int} = -\nabla_\theta q_{it}(\theta^*_{in})'$, so that $\widehat{S}_{int}$ in (12) becomes $\widehat{S}_{int} = -\nabla_\theta q_{it}(\widehat{\theta}_{in})'$.

details. The results indicate that the empirical size of both tests is close to the nominal size for any value of b_1. The power is also fairly good, in particular when b_1 is fairly large. Overall, the simulation results are encouraging and suggest that the tests we propose have good size and power properties.

6. Analytical examples

We now theoretically discuss the application of the SC_n test for model selection in the case of two overlapping models, ARMAX versus STAR, and two nested ARMAX-GARCH models, estimated by MLE.

For the ARMAX model, it is

$$\begin{aligned} Q_{1n}(\theta_1) &\equiv n^{-1}\sum_{t=1}^{n} q_{1t}(\theta_1), \\ q_{1t}(\theta_1) &\equiv -\frac{1}{2}\log(v_t^2) + \frac{1}{2}\eta_t^2(\theta_1)v_t^{-2}(\theta_1), \\ \eta_t(\theta_1) &\equiv Y_t - \sum_{i=1}^{k} b_i Z_{t-i} - \sum_{i=1}^{l} a_i \eta_{t-i}(\theta_1). \end{aligned} \quad (27)$$

For the STAR model, it is

$$\begin{aligned} Q_{2n}(\theta_2) &\equiv n^{-1}\sum_{t=1}^{n} q_{2t}(\theta_2), \\ q_{2t}(\theta_2) &\equiv \left\{ Y_t - \sum_{i=1}^{k} \left[c_i + d_i F(\gamma_0 Z_{t-k-1} - \gamma_1)\right] Z_{t-i} \right\}^2. \end{aligned} \quad (28)$$

The SC_n statistic for ARMAX versus STAR is

$$SC_n(\widehat{\theta}_{1n}, \widehat{\theta}_{2n}) = Q_{1n}(\widehat{\theta}_1) - Q_{2n}(\widehat{\theta}_2). \tag{29}$$

Again, for both models, we assume that $\{Y_t\}$ and $\{Z_{t-i},\, i = 0, 1..., k\}$ are strictly stationary and ergodic processes; we also maintain all the other assumptions in Sin and White (1996, Section 7), that guarantee that both $q_{it}(\theta_i)$ and $\nabla_{\theta_i} q_{it}(\theta_i)$ are near epoch dependent, and satisfy a Central Limit Theorem, for $i = 1, 2$. We also define $\widehat{\sigma}_n^2$ as in (18).

It can be easily shown that the ARMAX and the STAR models are partially overlapping. Hence, the asymptotic distribution of the SC_n statistic is that in Corollary 1, and we follow the two-step procedure described in Section 3.2. First, one can test for $\lim_{n\to\infty} n\sigma_n^2 = 0$ using the asymptotic distribution of $\widehat{\sigma}_n^2$ given in Theorem 2, where η can be estimated by the vector of eigenvalues of $\widehat{V}_n \widehat{\Sigma}_n$ (Davies (1980)), and estimation of $\widehat{V}_n$ and $\widehat{\Sigma}_n$ is discussed in Section 3. Second, if the hypothesis $\lim_{n\to\infty} n\sigma_n^2 = 0$ is accepted, we have (25), otherwise we have (26). Rejection of the null hypothesis in the second step provides evidence in favor of the ARMAX (STAR) model when SC_n is negative (positive).

As far as the comparison of nested models is concerned, an ARMAX-GARCH model is specified as in (27) with the additional recursive relationship

$$v_t^2(\theta_1) \equiv \alpha_0 + \sum_{i=1}^{c} \beta_i v_{t-i}^2(\theta_1) - \sum_{i=1}^{d} \alpha_i \eta_{t-i}^2(\theta_1). \tag{30}$$

The nested model is obtained by imposing $(p-q) > 0$ zero restrictions in (27) and/or (30). The SC_n statistic can be written as

$$SC_n(\widehat{\theta}_{1n}, \widetilde{\theta}_{1n}) = Q_{1n}(\widehat{\theta}_1) - Q_{1n}(\widetilde{\theta}_1), \tag{31}$$

where $\widetilde{\theta}_1$ is the restricted estimator of θ_1. In this case, the distribution of the SC_n test follows from that in Theorem 1, (16). A similar approach can be followed for the comparison of nested STAR models.

7. Conclusions

In this paper we have proposed a statistic for model selection that can be applied in a broad range of cases because it only requires weak assumptions on the processes and models under analysis. Under the null hypothesis the models under comparison are at the same distance of the DGP, under the alternative hypothesis one model is closer. The test is symmetric and directional. Its asymptotic distribution under the null is either a weighted sum of chi-squares or it is normal, depending on whether the models are nested or not. A test for the latter hypothesis is also suggested.

The extension of the procedure for the comparison of several models (see e.g. Shimodaira (1998)), possibly for non-stationary processes, a more elaborate evaluation of its finite sample behavior, and a comparison with other model selection criteria, are left for future research.

References

[1] Akaike, H. (1973) "Information Theory and an Extension of the Likelihood Principle", in B.N. Petrov and F. Csaki (eds.), *Proceedings of the Second International Symposium of Information Theory*, Budapest: Akademiai Kiado, 257-81.

[2] Davies, R.B. (1980), "AS155 The Distribution of a Linear Combination of χ^2 Random Variables", *Applied Statistics, Series C*, 29, 323-333.

[3] De Jong, R.M. and J. Davidson (2000), "Consistency of Kernel Estimators of Heteroschedastic and Autocorrelated Covariance Matrices", *Econometrica*, 68, 407-423.

[4] Gallant, A.R. (1987), *Nonlinear Statistical Models*, New York: John Wiley and Sons.

[5] Gallant, A.R. and H. White (1988), *A Unified Theory of Estimation and Inference for Nonlinear Dynamic Models*, Oxford: Basil Blackwell.

[6] Gourieroux, C. and A. Monfort (1989), "A General Framework for Testing a Null Hypothesis in a Mixed Form", *Econometric Theory*, 5, 63-82.

[7] Gourieroux, C. and A. Monfort (1994), "Testing Non-nested Hypotheses", in R.F. Engle and D.L. McFadden (eds.), *Handbook of Econometrics*, vol. IV, 2583-2637.

[8] Hall, A. and D. Pelletier (2007). "Non-Nested Testing in Models Estimated via Generalized Method of Moments", *mimeo*, NCState University.

[9] Hannan, E.J. and B.G. Quinn (1979), "The Determination of the Order of an Autoregression", *Journal of the Royal Statistical Society*, Series B, 41, 190-95.

[10] Hendry, D.F. and J.-F. Richard (1982), "On the formulation of empirical models in dynamic econometrics", *Journal of Econometrics*, 20, 3-33

[11] Kullback, L. and R.A. Leibler (1951), "On Information and Sufficiency", *Annals of Mathematical Statistics*, 22, 79-86.

[12] Lavergne, P. (1998), "Selection of Regressors in Econometrics: Parametric and Nonparametric Methods", *Econometric Reviews*, 17, 227-273.

[13] Hendry, D.F. and G.E. Mizon (1990), "Evaluating Dynamic Econometric Models by Encompassing the VAR", *Applied Economics* 102.

[14] Mizon, G.E. and J-F.Richard (1986), "The encompassing principle and its application to testing non-nested hypotheses", *Econometrica*, 54, 657-678.

[15] McLeish, D.L. (1975), "A Maximal Inequality and Dependent Strong Laws", *Annals of Probability*, 3, 826-836.

[16] Newey, W. and K. West (1987a), "A Simple Positive Definite Heteroskedasticity and Correlation Consistent Covariance Matrix", *Econometrica*, 55, 703-08.

[17] Newey, W. and K. West (1987b), "Hypothesis Testing with Efficient Method of Moments Estimators", *International Economic Review*, 28, 777-787.

[18] Rao, C.R. and S.K. Mitra (1971), *Generalized Inverse of Matrices and Its Applications*, New York: John Wiley and Sons.

[19] Rivers, D. and Q. Vuong (2002), "Model Selection Tests for Nonlinear Dynamic Models", *Econometrics Journal* 5, 1-39.

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

[21] Shimodaira, H. (1998), "An Application of Multiple Comparison Techniques to Model Selection", *Annals of the Institute of Statistical Mathematics*, 50, 1-15.

[22] Sin, C.-Y. and H. White (1996), "Information Criteria for Selecting Possibly Misspecified Parametric Models", *Journal of Econometrics*, 71, 207-225.

[23] Smith, R. (1994), "Consistent Tests for the Encompassing Hypothesis", w.p. 9403, INSEE, Paris.

[24] Vuong, Q. (1989), "Likelihood Ratio Tests for Model Selection and Non-nested Hypotheses", *Econometrica*, 57, 307-33.

[25] White, H. (1984), *Asymptotic Theory for Econometricians*, New York: Academic Press.

[26] White, H. (1994), *Estimation, Inference and Specification Analysis*, New York: Cambridge University Press.

[27] White, H. and I. Domowitz (1984), "Nonlinear Regressions with Dependent Observations", *Econometrica*, 52, 143-62.

APPENDIX

Proof of Lemma 2. From a first order Taylor expansion of the normal equations for D_1 and D_2, we obtain

$$0 = n^{1/2}\nabla_\theta Q_{1n}(\theta^*_{1n}) + A_{1n}(\theta^*_{1n})n^{1/2}(\widehat{\theta}_{1n} - \theta^*_{1n}) + o_p(1),$$

$$0 = n^{1/2}\nabla_\theta Q_{2n}(\theta^*_{2n}) + A_{2n}(\theta^*_{2n})n^{1/2}(\widehat{\theta}_{2n} - \theta^*_{2n}) + o_p(1).$$

It immediately follows from Gallant and White (1988, Corollary 5.5) that

$$n^{1/2}\begin{pmatrix} B_{1n}(\theta^*_{1n}) & B_{12n}(\theta^*_{1n}, \theta^*_{2n}) \\ B_{21n}(\theta^*_{2n}, \theta^*_{1n}) & B_{2n}(\theta^*_{2n}) \end{pmatrix}^{-1/2} \begin{pmatrix} \nabla_\theta Q_{1n}(\theta^*_{1n}) \\ \nabla_\theta Q_{2n}(\theta^*_{2n}) \end{pmatrix} \xrightarrow{D} N\left(0, I\right).$$

Finally, from A2-vii)-b), $A_{1n}(\theta^*_{1n})$ and $A_{2n}(\theta^*_{2n})$ are nonsingular, and the result follows. ∎

Proof of Lemma 3. Follows from Lemma 2 and Vuong (1989, Lemma 3.2). ∎

Proof of Lemma 4. Given Assumptions 1 and 2 for D_1 and D_2, from Gallant and White (1988, Theorem 5.3), $\sqrt{n}(Q_{1n}(\theta^*_{1n}) - \overline{Q}_{1n}(\theta^*_{1n}))$ and $\sqrt{n}(Q_{2n}(\theta^*_{2n}) - \overline{Q}_{2n}(\theta^*_{2n}))$ have a joint asymptotically normal distribution with mean zero and variance

$$\Omega^*_n \equiv var\left(\sqrt{n}\left(\left[\left(Q_{1n}(\theta^*_{1n}) - \overline{Q}_{1n}(\theta^*_{1n})\right), \left(Q_{2n}(\theta^*_{2n}) - \overline{Q}_{2n}(\theta^*_{2n})\right)\right]'\right)\right).$$

Then,

$$
\begin{aligned}
& \left[Q_{1n}(\theta^*_{1n}) - Q_{2n}(\theta^*_{2n}) - \left(\overline{Q}_{1n}(\theta^*_{1n}) - \overline{Q}_{2n}(\theta^*_{2n})\right)\right] \\
= \;& R^* \left(\left[\left(Q_{1n}(\theta^*_{1n}) - \overline{Q}_{1n}(\theta^*_{1n})\right), \left(Q_{2n}(\theta^*_{2n}) - \overline{Q}_{2n}(\theta^*_{2n})\right)\right]' \right) \\
= \;& n^{-1} \sum_{t=1}^{n} R^* \begin{bmatrix} q_{1t}(\theta^*_{1n}) - E(q_{1t}(\theta^*_{1n})) \\ q_{2t}(\theta^*_{2n}) - E(q_{2t}(\theta^*_{2n})) \end{bmatrix},
\end{aligned}
$$

where $R^* \equiv [1, -1]$, $\sigma_n^{-1} R^* \sqrt{n} \left(\left[\left(Q_{1n}(\theta^*_{1n}) - \overline{Q}_{1n}(\theta^*_{1n})\right), \left(Q_{2n}(\theta^*_{2n}) - \overline{Q}_{2n}(\theta^*_{2n})\right)\right]' \right) \xrightarrow{D} N(0,1)$, and $\sigma_n^2 \equiv R^* \Omega^*_n R^{*\prime}$ satisfies $\liminf_{n\to\infty} \sigma_n^2 > 0$, which yields the result. ■

Proof of Lemma 5. Follows from Gallant and White (1988, Theorem 6.1). ■

Proof of Lemma 6. Follows from Gallant and White (1988, Theorem 6.8). ■

Proof of Theorem 1. (i) From a first order Taylor expansion of $Q_{jn}(\theta^*_{jn})$ around $\widehat{\theta}_{jn}$, we obtain

$$
Q_{1n}(\theta^*_{1n}) = Q_{1n}(\widehat{\theta}_{1n}) + \frac{1}{2}(\widehat{\theta}_{1n} - \theta^*_{1n})' A_{1n}(\theta^*_{1n})(\widehat{\theta}_{1n} - \theta^*_{1n}) + o_p(1/n),
$$

$$
Q_{2n}(\theta^*_{2n}) = Q_{2n}(\widehat{\theta}_{2n}) + \frac{1}{2}(\widehat{\theta}_{2n} - \theta^*_{2n})' A_{2n}(\theta^*_{2n})(\widehat{\theta}_{2n} - \theta^*_{2n}) + o_p(1/n),
$$

so that

$$SC_n(\widehat{\theta}_{1n}, \widehat{\theta}_{2n}) = SC_n(\theta^*_{1n}, \theta^*_{2n}) - \frac{1}{2}(\widehat{\theta}_{1n} - \theta^*_{1n})' A_{1n}(\theta^*_{1n})(\widehat{\theta}_{1n} - \theta^*_{1n}) \tag{32}$$
$$+\frac{1}{2}(\widehat{\theta}_{2n} - \theta^*_{2n})' A_{2n}(\theta^*_{2n})(\widehat{\theta}_{2n} - \theta^*_{2n}) + o_p(1/n)$$
$$= SC_n(\theta^*_{1n}, \theta^*_{2n}) - \frac{1}{2}(\widehat{\theta}_n - \theta^*_n)' A^*_n(\widehat{\theta}_n - \theta^*_n),$$

where

$$A^*_n \equiv \begin{bmatrix} A_{1n}(\theta^*_{1n}) & 0 \\ 0 & -A_{2n}(\theta^*_{2n}) \end{bmatrix}.$$

By using (32), we have:

$$SC_n(\widehat{\theta}_{1n}, \widehat{\theta}_{2n}) = \overline{Q}_{1n}(\theta^*_{1n}) - \overline{Q}_{2n}(\theta^*_{2n})$$
$$+ \left(Q_{1n}(\theta^*_{1n}) - \overline{Q}_{1n}(\theta^*_{1n})\right) - \left(Q_{2n}(\theta^*_{2n}) - \overline{Q}_{2n}(\theta^*_{2n})\right)$$
$$-\frac{1}{2}(\widehat{\theta}_n - \theta^*_n)' A^*_n(\widehat{\theta}_n - \theta^*_n) + o_p(1/n),$$

Under H^*_0, $n\left(\overline{Q}_{1n}(\theta^*_{1n}) - \overline{Q}_{2n}(\theta^*_{2n})\right) = 0$. Also, $n\left(Q_{1n}(\theta^*_{1n}) - \overline{Q}_{1n}(\theta^*_{1n})\right) - \left(Q_{2n}(\theta^*_{2n}) - \overline{Q}_{2n}(\theta^*_{2n})\right) \xrightarrow{p} 0$ provided $\lim_{n\to\infty} n\sigma^2_n = 0$. Therefore, in this case, the leading terms in the expansion of $SC_n(\widehat{\theta}_{1n}, \widehat{\theta}_{2n})$ when n diverges is the quadratic form. Its asymptotic distribution follows directly Lemmas 2 and 3.

(ii) To prove consistency, under H^*_1: $\lim_{n\to\infty} n\left(\overline{Q}_{1n}(\theta^*_{1n}) - \overline{Q}_{2n}(\theta^*_{2n})\right) = -\infty$, and the leading term in the expansion of $SC_n(\widehat{\theta}_{1n}, \widehat{\theta}_{2n})$ becomes $n\left(\overline{Q}_{1n}(\theta^*_{1n}) - \overline{Q}_{2n}(\theta^*_{2n})\right)$.

Proof of Corollary 1. i) follows from RV (Theorem 1); ii) from Theorem 1 above. ■

Proof of Theorem 2. i) Convergence to zero of $\left(\widehat{\sigma}_n^2 - \sigma_n^2\right)$ follows from Lemma 6.

ii) Along the lines of Vuong (1989, p. 328-329), let us consider a Taylor series expansion of $n\widehat{\sigma}_n^2$ around $(\theta_{1n}^*, \theta_{2n}^*)$.

(a) The term of order zero in $(\theta_{1n}^*, \theta_{2n}^*)$ in the expansion is $n\sigma_n^2$, which is negligible under H_0^*.

(b) The terms of order one in $(\theta_{1n}^*, \theta_{2n}^*)$ in the expansion are

$$2n\left(n^{-1}\sum_{t=1}^{n}(q_{1t}(\theta_{1n}^*) - q_{2t}(\theta_{2n}^*))\nabla_\theta q_{1t}(\theta_{1n}^*)\right)(\widehat{\theta}_{1n} - \theta_{1n}^*)$$
$$-2n\left(n^{-1}\sum_{t=1}^{n}(q_{1t}(\theta_{1n}^*) - q_{2t}(\theta_{2n}^*))\nabla_\theta q_{2t}(\theta_{2n}^*)\right)(\widehat{\theta}_{2n} - \theta_{2n}^*),$$

Since $\lim\limits_{n\to\infty} n\sigma_n^2 = 0$, the heterogeneity asymptotically disappears, and $\mathop{lim}\limits_{n\to\infty} n^{-1}\sum_{t=1}^{n}(q_{1t}(\theta_{1n}^*) - q_{2t}(\theta_{2n}^*))$ is constant. Further, under H_0, $n^{-1}\sum_{t=1}^{n}(q_{1t}(\theta_{1n}^*) - q_{2t}(\theta_{2n}^*)) = 0$ for n sufficiently large. Therefore, for n sufficiently large, $q_{1t}(\theta_{1n}^*) - q_{2t}(\theta_{2n}^*) = 0$, and the terms of order one in the expansion become asymptotically negligible.

(c) The second order terms are

$$n(\widehat{\theta}_{1n} - \theta_{1n}^*, \widehat{\theta}_{2n} - \theta_{2n}^*)\widetilde{V}_n(\widehat{\theta}_{1n} - \theta_{1n}^*, \widehat{\theta}_{2n} - \theta_{2n}^*)' + o_p(1),$$

with,

$$
\begin{aligned}
\widetilde{V}_n &= \begin{pmatrix} G_{1n}(\widetilde{\theta}_{1n}) & G_{12n}(\widetilde{\theta}_{1n}, \widetilde{\theta}_{2n}) \\ G_{21n}(\widetilde{\theta}_{2n}, \widetilde{\theta}_{1n}) & G_{2n}(\widetilde{\theta}_{2n}) \end{pmatrix} \\
G_{1n}(\widetilde{\theta}_{1n}) &= n^{-1} \sum_{t=1}^{n} \nabla_\theta q_{1t}(\widetilde{\theta}_{1n}) \nabla_\theta q_{1t}(\widetilde{\theta}_{1n})' + n^{-1} \sum_{t=1}^{n} (q_{1t}(\widetilde{\theta}_{1n}) - q_{2t}(\widetilde{\theta}_{2n})) \nabla_\theta^2 q_{1t}(\widetilde{\theta}_{1n}), \\
G_{12n}(\widetilde{\theta}_{1n}, \widetilde{\theta}_{2n}) &= n^{-1} \sum_{t=1}^{n} \nabla_\theta q_{1t}(\widetilde{\theta}_{1n}) \nabla_\theta q_{2t}(\widetilde{\theta}_{2n})', \\
G_{2n}(\widetilde{\theta}_{2n}) &= n^{-1} \sum_{t=1}^{n} \nabla_\theta q_{2t}(\widetilde{\theta}_{2n}) \nabla_\theta q_{2t}(\widetilde{\theta}_{2n})' + n^{-1} \sum_{t=1}^{n} (q_{1t}(\widetilde{\theta}_{1n}) - q_{2t}(\widetilde{\theta}_{2n})) \nabla_\theta^2 q_{2t}(\widetilde{\theta}_{2n}),
\end{aligned}
$$

for some $\widetilde{\theta}_{1n}$ and $\widetilde{\theta}_{2n}$ in the segments $[\theta_{1n}^*, \widehat{\theta}_{1n}]$ and $[\theta_{2n}^*, \widehat{\theta}_{2n}]$, respectively. Given assumptions A2-A4, the second terms in $\widetilde{G}_{1n}$ and $\widetilde{G}_{2n}$ converge to zero in probability under H_0 and $\lim_{n\to\infty} n\sigma_n^2 = 0$, as in (b). From Lemma 6, $\widetilde{V}_n - V_n \xrightarrow{p} 0$ uniformly over the parameter space. Hence, the distribution for $\widehat{\sigma}_n^2$ in (19) follows from Lemma 3. Divergence of $\widehat{\sigma}_n^2$ under $H_1^* : \liminf_{n\to\infty} n\sigma_n^2 > 0$ follows from (ii), part (a), since the term of order zero does not vanish asymptotically.

iii) (a) The additional terms of order zero in $(\theta_{1n}^*, \theta_{2n}^*)$ in the expansion are of the type (constant multiplied by)

$$
\begin{aligned}
& n\left(n^{-1}\sum_{t=1}^{n}(q_{1t}(\theta_{1n}^*) - q_{2t}(\theta_{2n}^*))(q_{1t-k}(\theta_{1n}^*) - q_{2t-k}(\theta_{2n}^*))\right) \qquad (33) \\
= \; & n\left(n^{-1}\sum_{t=1}^{n}(q_{1t-k}(\theta_{1n}^*) - q_{2t-k}(\theta_{2n}^*))(q_{1t-k}(\theta_{1n}^*) - q_{2t-k}(\theta_{2n}^*))\right) \\
& + n\left(n^{-1}\sum_{t=1}^{n}(q_{1t}(\theta_{1n}^*) - q_{2t}(\theta_{2n}^*) - q_{1t-k}(\theta_{1n}^*) + q_{2t-k}(\theta_{2n}^*))(q_{1t-k}(\theta_{1n}^*) - q_{2t-k}(\theta_{2n}^*))\right),
\end{aligned}
$$

(for $k = 1, 2, ..., m_n$) and, under H_0, H_0^* and Assumptions 1-4, are $o_p(1)$.

(b) The additional terms of order one in $(\theta_{1n}^*, \theta_{2n}^*)$ in the expansion are also negligible, as in ii) above.

(c) The additional second order terms are of the type

$$2nw_{kn}(\widehat{\theta}_{1n} - \theta_{1n}^*, \widehat{\theta}_{2n} - \theta_{2n}^*)\widetilde{V}_n^k(\widehat{\theta}_{1n} - \theta_{1n}^*, \widehat{\theta}_{2n} - \theta_{2n}^*)' + o_p(1),$$

with

$$\begin{aligned}
\widetilde{V}_n^k &= \begin{pmatrix} G_{1n}^k(\widetilde{\theta}_{1n}) & G_{12n}^k(\widetilde{\theta}_{1n}, \widetilde{\theta}_{2n}) \\ G_{21n}^k(\widetilde{\theta}_{2n}, \widetilde{\theta}_{1n}) & G_{2n}^k(\widetilde{\theta}_{2n}) \end{pmatrix} \\
G_{1n}^k(\widetilde{\theta}_{1n}) &= n^{-1}\sum_{t=1}^n \nabla_\theta^2 q_{1t}(\widetilde{\theta}_{1n})(q_{1t-k}(\widetilde{\theta}_{1n}) - q_{2t-k}(\widetilde{\theta}_{2n})) + 2n^{-1}\sum_{t=1}^n \nabla_\theta q_{1t}(\widetilde{\theta}_{1n})\nabla_\theta q_{1t-k}(\widetilde{\theta}_{1n})' \\
&\quad + n^{-1}\sum_{t=1}^n (q_{1t}(\widetilde{\theta}_{1n}) - q_{2t}(\widetilde{\theta}_{2n}))\nabla_\theta q_{1t-k}(\widetilde{\theta}_{1n}), \\
G_{12n}^k(\widetilde{\theta}_{1n}, \widetilde{\theta}_{2n}) &= n^{-1}\sum_{t=1}^n \nabla_\theta q_{1t}(\widetilde{\theta}_{1n})\nabla_\theta q_{2t-k}(\widetilde{\theta}_{2n})' + n^{-1}\sum_{t=1}^n \nabla_\theta q_{1t-k}(\widetilde{\theta}_{1n})\nabla_\theta q_{2t}(\widetilde{\theta}_{2n})', \\
G_{2n}^k(\widetilde{\theta}_{2n}) &= n^{-1}\sum_{t=1}^n \nabla_\theta^2 q_{2t}(\widetilde{\theta}_{1n})(q_{1t-k}(\widetilde{\theta}_{1n}) - q_{2t-k}(\widetilde{\theta}_{2n})) + 2n^{-1}\sum_{t=1}^n \nabla_\theta q_{2t}(\widetilde{\theta}_{1n})\nabla_\theta q_{2t-k}(\widetilde{\theta}_{1n})' \\
&\quad + n^{-1}\sum_{t=1}^n (q_{1t}(\widetilde{\theta}_{1n}) - q_{2t}(\widetilde{\theta}_{2n}))\nabla_\theta q_{2t-k}(\widetilde{\theta}_{1n}).
\end{aligned}$$

for $k = 1, ..., m_n$ and for some $\widetilde{\theta}_{1n}$ and $\widetilde{\theta}_{2n}$ in the segments $[\theta_{1n}^*, \widehat{\theta}_{1n}]$ and $[\theta_{2n}^*, \widehat{\theta}_{2n}]$, respectively. From Lemma 6, $\widetilde{V}_n^k - V_n^k \xrightarrow{p} 0$. The distribution for $\widehat{\sigma}_n^2$ in (19) follows from Lemma 3. Divergence of $\widehat{\sigma}_n^2$ under H_1^* follows from the proof of (ii) part (a). ■

Proof of Corollary 2. The criterion IC_n in (21) can be written as

$$IC_n = n(Q_{1n}(\widehat{\theta}_{1n}) - Q_{2n}(\widehat{\theta}_{2n})) + \widehat{c}_n$$

or, using the expansions in Theorem 1 and Lemma 4,

$$\begin{aligned} IC_n &= n\left[\overline{Q}_{1n}(\theta^*_{1n}) - \overline{Q}_{2n}(\theta^*_{2n})\right] \\ &\quad + n\left[n^{-1}\sum_{t=1}^{n}(q_{1t}(\theta^*_{1n}) - q_{2n}(\theta^*_{2n})) - E\left(n^{-1}\sum_{t=1}^{n}(q_{1t}(\theta^*_{1n}) - q_{2t}(\theta^*_{2n}))\right)\right] \\ &\quad -\frac{1}{2}n(\widehat{\theta}_{1n} - \theta^*_{1n})' A_{1n}(\theta^*_{1n})(\widehat{\theta}_{1n} - \theta^*_{1n}) + \frac{1}{2}n(\widehat{\theta}_{2n} - \theta^*_{2n})' A_{2n}(\theta^*_{2n})(\widehat{\theta}_{2n} - \theta^*_{2n}) \\ &\quad + \widehat{c}_n + o_p(1) \end{aligned}$$

The conditions in i) imply that the leading term in the expansion is $n\left[\overline{Q}_{1n}(\theta^*_{1n}) - \overline{Q}_{2n}(\theta^*_{2n})\right]$, in which case IC_n diverges to positive infinity; in fact:

$$\begin{aligned} IC_n &= n\{\left[\overline{Q}_{1n}(\theta^*_{1n}) - \overline{Q}_{2n}(\theta^*_{2n})\right] \\ &\quad + \left[n^{-1}\sum_{t=1}^{n}(q_{1t}(\theta^*_{1n}) - q_{2n}(\theta^*_{2n})) - E\left(n^{-1}\sum_{t=1}^{n}(q_{1t}(\theta^*_{1n}) - q_{2t}(\theta^*_{2n}))\right)\right] \\ &\quad -\frac{1}{2}(\widehat{\theta}_{1n} - \theta^*_{1n})' A_{1n}(\theta^*_{1n})(\widehat{\theta}_{1n} - \theta^*_{1n}) + \frac{1}{2}(\widehat{\theta}_{2n} - \theta^*_{2n})' A_{2n}(\theta^*_{2n})(\widehat{\theta}_{2n} - \theta^*_{2n}) \\ &\quad + \widehat{c}_n/n\} + o_p(1). \end{aligned}$$

Those in ii) that the leading term is $\widehat{c}_n$, since it grows faster than $\sqrt{n}$, the rate of growth of the first two terms below, and the third term is asymptotically negligible when divided by $\sqrt{n}$:

$$\begin{aligned}IC_n &= n^{1/2}\{n^{1/2}\left[\overline{Q}_{1n}(\theta^*_{1n}) - \overline{Q}_{2n}(\theta^*_{2n})\right] \\ &\quad + \left[n^{-1/2}\sum_{t=1}^{n}\left[(q_{1t}(\theta^*_{1n}) - q_{2n}(\theta^*_{2n})) - E\left(q_{1t}(\theta^*_{1n}) - q_{2t}(\theta^*_{2n})\right)\right]\right] \\ &\quad -\frac{1}{2}n^{1/2}(\widehat{\theta}_{1n} - \theta^*_{1n})' A_{1n}(\theta^*_{1n})(\widehat{\theta}_{1n} - \theta^*_{1n}) + \frac{1}{2}n^{1/2}(\widehat{\theta}_{2n} - \theta^*_{2n})' A_{2n}(\theta^*_{2n})(\widehat{\theta}_{2n} - \theta^*_{2n}) \\ &\quad +n^{-1/2}\widehat{c}_n\} + o_p(1).\end{aligned}$$

Those in iii) again that the leading term is $\widehat{c}_n$, since in this case the first and third terms below are bounded in probability, and the second term is zero because $\lim_{n\to\infty} n\sigma_n^2 = 0$ holds:

$$\begin{aligned}IC_n &= n\left[\overline{Q}_{1n}(\theta^*_{1n}) - \overline{Q}_{2n}(\theta^*_{2n})\right] \\ &\quad +n^{1/2}\left[n^{-1/2}\sum_{t=1}^{n}\left[(q_{1t}(\theta^*_{1n}) - q_{2n}(\theta^*_{2n})) - E\left(q_{1t}(\theta^*_{1n}) - q_{2t}(\theta^*_{2n})\right)\right]\right] \\ &\quad -\frac{1}{2}n(\widehat{\theta}_{1n} - \theta^*_{1n})' A_{1n}(\theta^*_{1n})(\widehat{\theta}_{1n} - \theta^*_{1n}) + \frac{1}{2}n(\widehat{\theta}_{2n} - \theta^*_{2n})' A_{2n}(\theta^*_{2n})(\widehat{\theta}_{2n} - \theta^*_{2n}) \\ &\quad +\widehat{c}_n + o_p(1). \blacksquare\end{aligned}$$

Table 1. Notation

$Q_{jn}(\theta) \equiv n^{-1}\sum_{t=1}^{n} q_{jt}(\theta),\ j=1,2$

$\overline{Q}_{in}(\theta_{in}^{*}) = n^{-1}\sum_{t=1}^{n} E(q_{it}(\theta_{in}^{*}))$

$SC_n(\theta_{1n},\theta_{2n}) \equiv Q_{1n}(\theta_{1n}) - Q_{2n}(\theta_{2n}),$

$SC_n(\widehat{\theta}_{1n},\widehat{\theta}_{2n}) \equiv Q_{1n}(\widehat{\theta}_{1n}) - Q_{2n}(\widehat{\theta}_{2n})$

$\overline{SC}_n(\theta_n) \equiv \overline{Q}_{1n}(\theta_{1n}) - \overline{Q}_2(\theta_{2n})$

$\Omega_n^{*} \equiv var\left(\sqrt{n}\left[Q_{1n}(\theta_{1n}) - Q_{2n}(\theta_{2n}) - \left(\overline{Q}_{1n}(\theta_{1n}) - \overline{Q}_2(\theta_{2n})\right)\right]\right)$

$\sigma_n^2 \equiv R^{*}\Omega_n^{*}R^{*\prime},\ \ R^{*} \equiv [1,-1],$

$$V_n = \begin{pmatrix} G_{1n}(\theta_{1n}^{*}) & G_{12n}(\theta_{1n}^{*},\theta_{2n}^{*}) \\ G_{21n}(\theta_{2n}^{*},\theta_{1n}^{*}) & G_{2n}(\theta_{2n}^{*}) \end{pmatrix}$$

$G_{in}(\widetilde{\theta}_{1n}) = n^{-1}\sum_{t=1}^{n}\nabla_\theta q_{it}(\theta_{in})\nabla_\theta q_{it}(\theta_{in})' + n^{-1}\sum_{t=1}^{n}(q_i(\theta_{in}) - q_{jt}(\theta_{jn}))\nabla_\theta^2 q_{it}(\theta_{in}),$

$G_{ijn}(\widetilde{\theta}_{in},\widetilde{\theta}_{jn}) = n^{-1}\sum_{t=1}^{n}\nabla_\theta q_{it}(\theta_{in})\nabla_\theta q_{jt}(\theta_{jn})',$

$$A_n^{*} \equiv \nabla_\theta^2\overline{Q}_n(\theta_n^{*}) = \begin{pmatrix} A_{1n}^{*} & 0 \\ 0 & A_{2n}^{*} \end{pmatrix},\quad W_n \equiv \begin{pmatrix} A_{1n}^{*} & 0 \\ 0 & -A_{2n}^{*} \end{pmatrix}$$

$\widehat{\theta}_n \equiv \left[\widehat{\theta}_{1n}',\widehat{\theta}_{2n}'\right]',\theta_n^{*} \equiv \left[\theta_{1n}^{*\prime},\theta_{2n}^{*\prime}\right]'$

$$\Sigma \equiv Var\begin{pmatrix} \widehat{\theta}_{1n} - \theta_{1n}^{*} \\ \widehat{\theta}_{2n} - \theta_{2n}^{*} \end{pmatrix} = \lim_{n\to\infty}\Sigma_n$$

$$\Sigma_n = \begin{pmatrix} A_{1n}^{-1}(\theta_{1n}^{*})B_{1n}(\theta_{1n}^{*})A_{1n}^{-1}(\theta_{1n}^{*}) & A_{1n}^{-1}(\theta_{1n}^{*})B_{12n}(\theta_{1n}^{*},\theta_{2n}^{*})A_{2n}^{-1}(\theta_{2n}^{*}) \\ A_{2n}^{-1}(\theta_{2n}^{*})B_{21n}(\theta_{2n}^{*},\theta_{1n}^{*})A_{1n}^{-1}(\theta_{1n}^{*}) & A_{1n}^{-1}(\theta_{2n}^{*})B_{2n}(\theta_{2n}^{*})A_{2n}^{-1}(\theta_{2n}^{*}) \end{pmatrix}$$

$B_{in}^{*} = Var[n^{1/2}\nabla_\theta Q_{in}(\theta_{in}^{*})],$

$B_{ijn}^{*} = B_{jin}^{*\prime} = Cov[n^{1/2}\nabla_\theta Q_{in}(\theta_{in}^{*}), n^{1/2}\nabla_\theta Q_{jn}(\theta_{jn}^{*})]$, for $i,j=1,2,\ i\neq j,$

∇_θ denotes the gradient function

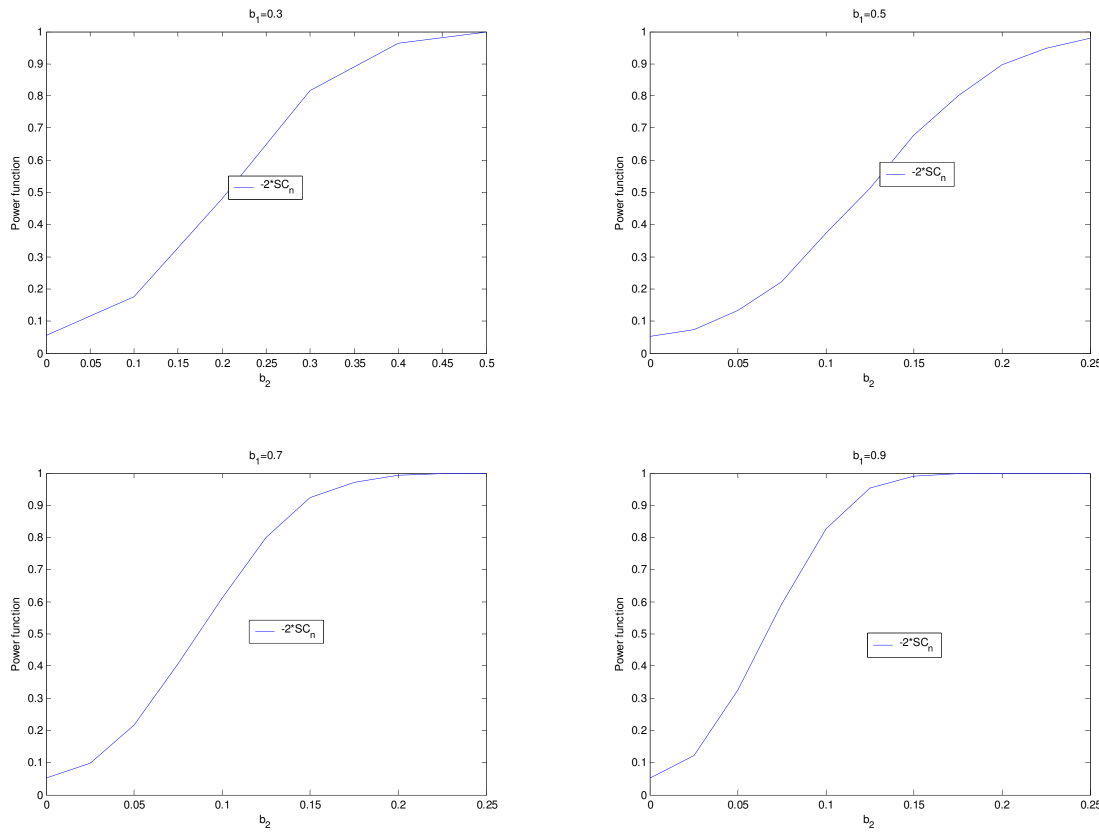
$\xrightarrow{p}$ is convergence in probability, $\xrightarrow{D}$ is convergence in distribution

Table 2. Monte Carlo results

b_1	b_2	$-2SC_n(\widehat{\theta})$	$n\widehat{\sigma}_n^2$	b_1	b_2	$-2SC_n(\widehat{\theta})$	$n\widehat{\sigma}_n^2$
0.3	0	0.0552	0.0548	0.7	0.05	0.215	0.2094
0.3	0.1	0.175	0.1766	0.7	0.075	0.4052	0.4058
0.3	0.2	0.4802	0.4814	0.7	0.1	0.6136	0.6134
0.3	0.3	0.8164	0.8162	0.7	0.125	0.8014	0.8004
0.3	0.4	0.9648	0.9666	0.7	0.15	0.9242	0.9228
0.3	0.5	0.9978	0.9972	0.7	0.175	0.9726	0.9716
0.5	0	0.0512	0.0516	0.7	0.2	0.9942	0.9938
0.5	0.025	0.0736	0.0726	0.7	0.225	0.9992	0.9988
0.5	0.05	0.133	0.1352	0.7	0.25	1	0.9998
0.5	0.075	0.2216	0.2248	0.9	0	0.0532	0.0532
0.5	0.1	0.3728	0.3764	0.9	0.025	0.1216	0.1188
0.5	0.125	0.5122	0.512	0.9	0.05	0.3248	0.322
0.5	0.15	0.6768	0.677	0.9	0.075	0.5914	0.5874
0.5	0.175	0.8004	0.8032	0.9	0.1	0.8278	0.8218
0.5	0.2	0.8964	0.896	0.9	0.125	0.9524	0.9524
0.5	0.225	0.9468	0.9444	0.9	0.15	0.9908	0.991
0.5	0.25	0.9802	0.9798	0.9	0.175	0.9982	0.9984
0.7	0	0.051	0.0516	0.9	0.2	0.9998	1
0.7	0.025	0.0966	0.0974	0.9	0.225	1	1

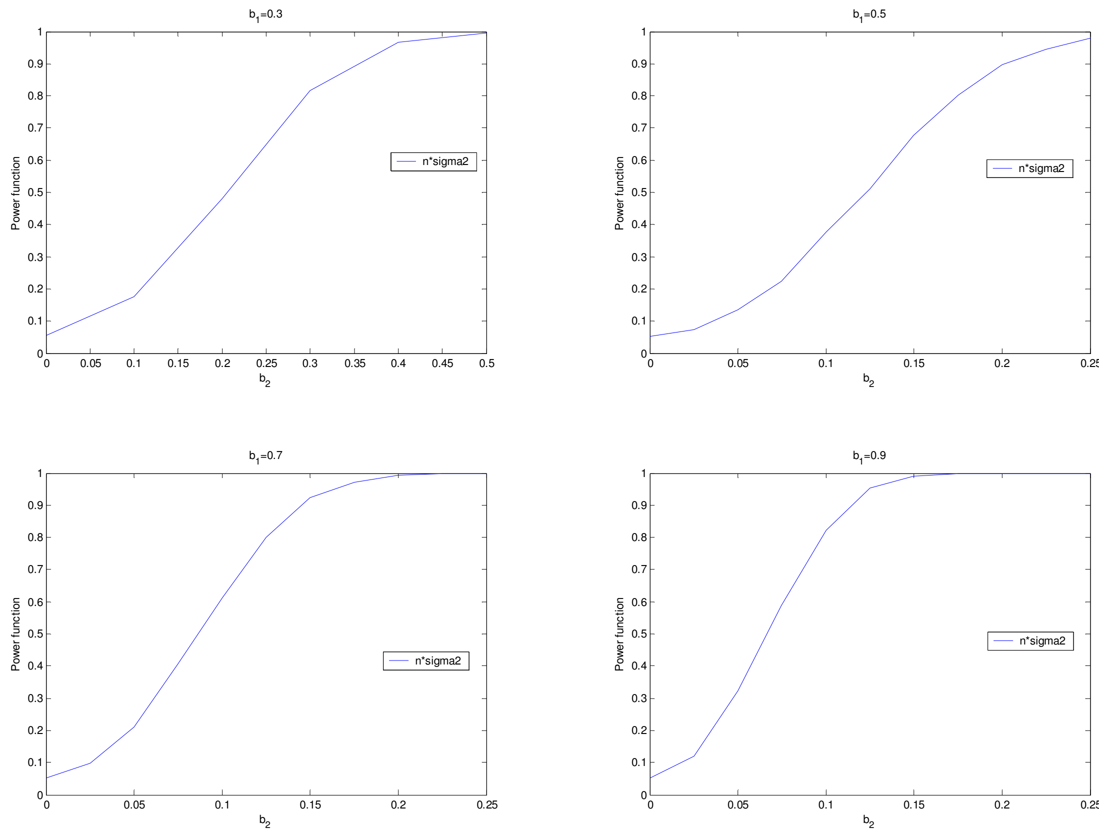
Note. The table reports size and power functions of (16) and (19).

Figure 1 (a)



Note. The figure reports size and power functions of (16) and (19).

Figure 1 (b)



Note. The figure reports size and power functions of (16) and (19).