Model Selection for Non-Linear Dynamic Models

Massimiliano Marcellino[*][†]

Istituto di Economia Politica, Università Bocconi,

IGIER and European University Institute

This version: October 1999

Abstract

This paper develops tests for selection of competing non-linear dynamic models. The null hypothesis is that the models are equally close 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 non-nested, overlapping, or nested. They can be correctly specified or not. Their parameters can be estimated by a variety of methods, including Maximum Likelihood, Non-Linear Least Squares, Method of Moments, where the choice depends on the selected measure of closeness to the DGP. 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 ARMAX and STAR models, and of nested ARMAX-GARCH models are discussed as examples.

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), Marcellino and Mizon (1999) 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, 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 (ML).

In this paper we suggest a statistic similar to Vuong's test, that can be applied to compare models for dependent and heterogenous processes. The models can be linear or non-linear; nested, non-nested, or overlapping; correctly specified or mis-specified. Their parameters are estimated by means of optimand estimators, that include ML, Non-Linear Least Squares (NLS), and Method of Moments (MM). This provides a powerful tool for model selection under general conditions.[1]

Several other results are obtained as a by-product. In particular, when ML estimation is adopted, our statistic boils down to a Likelihood Ratio (LR) test. We thus 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. Finally, conditions for weak consistency of information criteria based on optimand estimators (not necessarily MLE) are also derived.

The paper builds upon recent work on estimation, specification and inference for possibly misspecified non-linear models of dependent and heterogenous pro-

---

[1]After the paper was completed, Hal White brought to my attention a related unpublished manuscript by Rivers and Vuong (1999). They only deal with *non-nested* non-linear dynamic models, but allow for a different objective function for model estimation and selection, which may depend on nuisance parameters. Instead, I also consider the more common case where the models under comparison are nested or overlapping, which yields quite different results from the non-nested case, relate the results to weak consistency of information criteria, and analyze two important time-series examples. Hence, the papers are more complementary than substitutes.

cesses, see in particular White (1984), Gallant (1987), Gallant and White (1988), White (1994), and the references therein. Section 2 introduces assumptions and definitions, and derives a set of preliminary results. The main results are presented in Section 3, where the asymptotic distribution of the statistic is derived, and our approach is compared with the use of information criteria, non-nested tests and encompassing statistics. The comparison of ARMAX and STAR models, and of nested ARMAX-GARCH models are discussed as examples in Section 4. Section 5 concludes. All the proofs are gathered in the Appendix.

2. 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.

Assumption 1 (Data Generation)
Let (Ω, F, P) be a complete probability space. The stochastic process $\{V_t\}$, with generic element $V_t : \Omega \to R^v$, $v \in N$, 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 $X_t : \Omega \to R^{w_t}$, $w_t \in N$, and

$$(\omega) = W_t(..., V_{t-1}(\omega), V_t(\omega), V_{t+1}(\omega), ...), \quad \omega \in \Omega, \qquad (2.1)$$

where $W_t : \times_{\tau=-\infty}^{\infty} R^v \to R^{w_t}$ are such that X_t is measurable-$F/\mathcal{B}(R^{w_t})$, $t = 0, \pm 1, \pm 2, ...$. ∎

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 processes 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 (2.1), 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 (2.1) let us 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 probability distributions $D = \{D_\theta : \theta \in \Theta\}$, defined on the measurable space $(R_\infty^w, \mathcal{B}(R_\infty^w))$, with $R_\infty^w = \times_{\tau=-\infty}^{\infty} R^v$. ∎

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, it is $D_\beta(A) = P\{\omega : \{S_t(..., V_{t-1}(\omega), V_t(\omega), V_{t+1}(\omega), ...); \beta\} \in A\}$, for $\forall A \in \mathcal{B}(R_\infty^w)$. Often, a model is also defined by means of a finite dimensional density function, $d_n(X_1, ..., X_n; \theta) = dD_{n\theta}/d\mu_n$, 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 $D_{\theta_0}(A) = P\{\omega : \{X_t\} \in A\}$, $\forall A \in \mathcal{B}(R_\infty^w)$.

Once D is specified, ML techniques can be adopted to estimate the parameters of the model, θ. In practice, the specification of D can be so complex as to preclude analytical tractability. In these cases, it can still be feasible to derive a set of moment conditions that have to be satisfied, such as

$$E_\theta(m_t(X_1, ..., X_n; \theta)) = \int m_t(X_1, ..., X_n; \theta) dD_\theta = 0.$$

The moment conditions can then be exploited to derive an estimator of θ. Also, the investigator may be unwilling to make an explicit distributional assumption, and prefer to adopt an estimation method such as NLS. These three alternative estimation methods, and others, all belong to the class of optimand 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)

Θ is a compact subset of R^k. The objective function to be minimized, $Q_n : \Omega \times \Theta \to R$, is defined as

$$
\begin{aligned}
Q_n(\omega, \theta) &\equiv g_n(\psi_n(\omega, \theta)), &(2.2)\\
\psi_n(\omega, \theta) &\equiv n^{-1} \sum_{t=1}^{n} q_t(\omega, \theta), \quad n = 1, 2, ...,
\end{aligned}
$$

and

i) $\{g_n : R^l \to R\}$ is continuously differentiable of order 2 on compact subsets of R^l uniformly in n.

ii) $q_t : \Omega \times \Theta \to R^l$ 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) $\{\nabla_\theta^2 q_t(\theta)\}$ are a.s. Lipschitz-L_1.

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

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

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

vii) (a) $\{B_n^*\} = \{Var[n^{1/2} \nabla_\theta Q_n(\theta_n^*)]\}$, (b) $\{A_n^*\} = \{\nabla_\theta^2 \overline{Q}_n(\theta_n^*)\}$, (c) $\{G_n^*\} = \{Var[n^{1/2} \nabla_\theta \psi_n(\theta_n^*)]\}$, and (d) $\{H_n^*\} = \{\nabla_\theta^2 \overline{\psi}_n(\theta_n^*)\}$ are uniformly positive definite. ∎

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

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

4

Within this framework, ML estimators are obtained by setting $g_n(\psi_n) = \psi_n = -d_n(X_1, ..., X_n; \theta)$, considered as a function of θ. For NLS estimators, $g_n(\psi_n) = \psi_n$ and $q_t(\omega, \theta) = (x_t - s_t(\beta))^2$. For moment estimators, $g_n(\psi_n) = \psi_n^2$ and $q_t(\omega, \theta) = m_t(X_1, ..., X_n; \theta)$. For GMM estimators, $g_n(\psi_n) = \psi' P_n^* \psi$, where $\{P_n^*\}$ is a sequence of non stochastic matrices.[2]

The conditions iii)-a), iv)-a), and v)-a) impose, respectively, smoothness, memory, and moment conditions on $q_t(\theta)$ to ensure that $\psi_n(\theta) - \overline{\psi}_n(\theta) \to 0$ 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 next section.

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 R^p and R^q, respectively, with $p \geq q$. D_1 and D_2 are defined on the same measurable space $(R_\infty^w, \mathcal{B}(R_\infty^w))$. Following Vuong (1989), we say that

i) D_1 and D_2 are strictly non-nested if and only if $D_1 \frown D_2 = \emptyset$.

ii) D_2 is nested in D_1 if and only if $D_2 \subset D_1$. In this case, there exists a function continuously differentiable of order 2, $\phi : \Theta_2 \to \Theta_1$, such that $D_2(A; \theta_2) = D_1(A; \phi(\theta_2))$, $\forall \theta_2 \in \Theta_2$, $\forall A \in \mathcal{B}(R_\infty^w)$.

iii) D_1 and D_2 are overlapping if and only if (a) $D_1 \frown D_2 \neq \emptyset$, (b) $D_1 \nsubseteq D_2$ and $D_2 \nsubseteq D_1$. ∎

---

[2]See Gallant and White (1988, p. 12-13) for the additional complications that can arise when $\{P_n^*\}$ is a stochastic sequence.

The models are strictly non nested when there is no distribution common to D_1 and D_2. From ii), D_2 is nested in D_1 when any distribution in D_2 is also in D_1, and the function ϕ provides the mapping from D_2 to D_1. The models overlap when there exist common distributions, but neither model is nested in the other.

Assumption 3 (Selection Criterion)

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

$$
\begin{aligned}
SC_n(\widehat{\theta}_{1n}, \widehat{\theta}_{2n}) &\equiv Q_{1n}(\omega, \widehat{\theta}_{1n}) - Q_{2n}(\omega, \widehat{\theta}_{2n}) \\
&= g_{1n}(\psi_{1n}(\omega, \widehat{\theta}_{1n})) - g_{2n}(\psi_{2n}(\omega, \widehat{\theta}_{2n})) \\
&= g_{1n}\left(n^{-1}\sum_{t=1}^{n} q_{1t}(\omega, \widehat{\theta}_{1n})\right) - g_{2n}\left(n^{-1}\sum_{t=1}^{n} q_{2t}(\omega, \widehat{\theta}_{2n})\right),
\end{aligned}
\tag{2.3}
$$

where $q_{1t} : \Omega \times \Theta_1 \to R^l$, $q_{2t} : \Omega \times \Theta_2 \to R^l$. 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),

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

where $\overline{Q}_{in}(\omega, \theta_{in}^*) = g_{1n}(n^{-1}\sum_{t=1}^{n} E_P(q_{it}(\omega, \theta_{in}^*)))$, $i = 1, 2$. ∎

We now introduce some additional mild requirements on the objective function. They are all satisfied, e.g., when Q_n coincides with the likelihood function.

Assumption 4 (Nesting and Q_n)

i) If D_1 and D_2 are strictly non-nested, then $P\{Q_{1n}(\omega, \theta_{1n}^*) = Q_{2n}(\omega, \theta_{2n}^*)\} = 0$, uniformly in n.

ii) If D_1 and D_2 are nested and $\theta_{1n}^* = \phi(\theta_{2n}^*)$, then $P\{Q_{1n}(\omega, \theta_{1n}^*) = Q_{2n}(\omega, \theta_{2n}^*)\} = 1$, uniformly in n.

iii) If D_1 and D_2 are overlapping, then $\int Q_{1n}(\omega, \theta_{1n}^*)dP = \int Q_{2n}(\omega, \theta_{2n}^*)dP$, uniformly in n. ∎

Lemma 2 (Overlapping and Q_n)

If D_1 and D_2 are overlapping and Assumption 4-iii) holds, then $P\{Q_{1n}(\omega, \theta_{1n}^*) = Q_{2n}(\omega, \theta_{2n}^*)\} = 1$ uniformly in n if and only if $\sigma_n^2 = 0$ uniformly in n, where $\sigma_n^2 = n^2 var(Q_{1n}(\omega, \theta_{1n}^*) - Q_{2n}(\omega, \theta_{2n}^*))$. $\blacksquare$

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

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

Given Assumptions 1 and 2 for D_1 and D_2, it is

$$n^{1/2}\Sigma_n^{-1/2} \begin{pmatrix} \widehat{\theta}_{1n} - \theta_{1n}^* \\ \widehat{\theta}_{2n} - \theta_{2n}^* \end{pmatrix} \xrightarrow{D} N(0, I), \tag{2.4}$$

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

where

$$\begin{aligned} A_{in}(\theta_{in}^*) &= \nabla_\theta^2 \overline{Q}_n(\theta_{in}^*), \\ B_{in}(\theta_{in}^*) &= Var[n^{1/2}(\nabla_\theta Q_{in}(\theta_{in}^*)], \\ B_{ijn}(\theta_{in}^*, \theta_{jn}^*) &= B_{jin}(\theta_{jn}^*, \theta_{in}^*)' = cov[n^{1/2}(\nabla_\theta Q_{in}(\theta_{in}^*), n^{1/2}(\nabla_\theta Q_{jn}(\theta_{jn}^*)], \end{aligned}$$

for $i, j = 1, 2$, $i \neq j$. $\blacksquare$

Lemma 4 (Distribution of quadratic forms in $\widehat{\theta}_1$ and $\widehat{\theta}_2$)

Let Q be a $(p+q) \times (p+q)$ real symmetric matrix and $Y = (\widehat{\theta}_{1n} - \theta_{1n}^*, \widehat{\theta}_{2n} - \theta_{2n}^*)'$. 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 $Q\Sigma_n$. $\blacksquare$

Lemma 5 (Distribution of $Q_{1n}(\theta_{1n}^*) - Q_{2n}(\theta_{2n}^*)$)

Given Assumptions 1 and 2 for D_1 and D_2, if σ_n^2 is $O(n)$, it is

$$n\sigma_n^{-1}\left[Q_{1n}(\theta_{1n}^*) - Q_{2n}(\theta_{2n}^*) - \left(\overline{Q}_{1n}(\theta_{1n}^*) - \overline{Q}_{2n}(\theta_{2n}^*)\right)\right] \xrightarrow{D} N(0, 1). \blacksquare \tag{2.5}$$

Now, we have to deal with estimation of σ_n^2 in (2.5), of the variance covariance matrix Σ_n in (2.4), and of G_n and H_n in Assumption 2)-vii) that will be used in next section. We have,

Lemma 6 (Estimation of $A_{in}(\theta_i^*)$ and $H_{in}(\theta_i^*)$)

7

Given Assumptions 1 and 2 for D_1 and D_2, it is $\widehat{A}_{in} - A_{in}(\theta_i^*) \xrightarrow{a.s.} 0$ and $\widehat{H}_{in} - H_{in}(\theta_i^*) \xrightarrow{a.s.} 0$, for $i = 1, 2$, where $\widehat{A}_{in} = \nabla_\theta^2 Q_{in}(\widehat{\theta}_{in})$ and $\widehat{H}_{in} = \nabla_\theta^2 \psi_{in}(\widehat{\theta}_{in})$. $\blacksquare$

Estimation of $B_{in}(\theta_i^*)$, $B_{ijn}(\theta_i^*, \theta_j^*)$, $G_{in}(\theta_i^*)$ and σ_n^2 is more complex. Let us focus on $B_{in}(\theta_i^*)$. From Gallant and White (1988, Theorem 5.4), it follows that

$$\{B_{in}^*\} = \{Var[n^{1/2}\nabla_\theta Q_{in}(\theta_{in}^*)]\} = \{Var[n^{-1/2}\sum_{t=1}^{n} M_{int}^*]\}, \qquad (2.6)$$

where

$$\begin{aligned} M_{int}^* &= S_{int}^* - E(S_{int}^*), \\ S_{int}^* &= \nabla_\theta \psi_{in}(\theta_{in}^*)' \nabla_\psi^2 g_{in}(\overline{\psi}_{in}(\theta_{in}^*)) q_{it}(\theta_{in}^*) + \nabla_\theta q_{it}(\theta_{in}^*)' \nabla_\psi g_{in}(\overline{\psi}_{in}(\theta_{in}^*))'. \end{aligned}$$

If we expand the expression in (2.6), we get

$$\{B_{in}^*\} = n^{-1}\sum_{t=1}^{n} E(M_{int}^* M_{int}^{*'}) + n^{-1}\sum_{\tau=1}^{n-1}\sum_{t=\tau+1}^{n}\left[E(M_{int}^* M_{in(t-\tau)}^{*'}) + E(M_{in(t-\tau)}^* M_{int}^{*'})\right].$$
$$(2.7)$$

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

$$\widehat{B}_{in} = w_{n0}n^{-1}\sum_{t=1}^{n}\widehat{S}_{int}\widehat{S}_{int}' + n^{-1}\sum_{\tau=1}^{m_n} w_{n\tau}\sum_{t=\tau+1}^{n}\left[\widehat{S}_{int}\widehat{S}_{in(t-\tau)}' + \widehat{S}_{in(t-\tau)}\widehat{S}_{int}'\right], \quad (2.8)$$

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 strenghtened, and requirements on the truncation lag (m_n) and the weights $(w_{n\tau})$ added. We have,

Assumption 5 (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).

ii) The elements of (a) $\{q_t(\theta)\}$, (b) $\{\nabla_\theta q_t(\theta)\}$, (c) $\{\nabla_\theta^2 q_t(\theta)\}$ 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 $O(n^{1/4})$.

8

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

v) $E(S_{nt}^*) = 0$. ■

Defining

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

the following lemma follows from Gallant and White (1988, Theorem 6.8).

Lemma 7 (Convergence of $\widehat{B}_{in}$)
Under the conditions in Assumptions 1, 2 and 5-i) to 5-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 5-v) also holds, $\widehat{B}_{in} - B_{in}^* \xrightarrow{p} 0$. ■

The additional condition $E(S_{nt}^*) = 0$ is usually imposed, e.g. Newey and West (1987, 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 stationary and $q_t(\omega, \theta) = q(T^t\omega, \theta)$, or that the model is correctly specified. Notice also that when (M_{nt}^*, F^t) is a martingale difference sequence, it is $\{B_{in}^*\} = n^{-1} \sum_{t=1}^{n} E(M_{int}^* M_{int}^{*'})$, which simplifies estimation, even if the estimator $\widehat{B}_{in} = n^{-1} \sum_{t=1}^{n} \widetilde{S}_{int} \widehat{S}_{int}'$ remains in general non consistent. Estimators for the other parameters, $\widehat{B}_{ijn}$, $\widehat{G}_n$ and $\widehat{\sigma}_n^2$, are obtained along the same lines.

Finally, we need additional conditions to derive the asymptotic distribution of $\widehat{\sigma}_n^2$ in next section.

Assumption 6 (Distribution of $\widehat{\sigma}_n^2$)

i) The elements of $\{q_t(\theta)\nabla_\theta^2 q_t(\theta)\}$ are near epoch dependent on $\{V_t\}$ of size -1 on (Θ, ρ), where ρ is any convenient norm on R^k.

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

3. Tests for model selection

In the first subsection we derive the asymptotic distribution of the model selection criterion $SC_n(\widehat{\theta}_{1n}, \widehat{\theta}_{2n})$, that underlies the model selection tests. In the second subsection we present the tests for selection of non-nested, nested, and overlapping models. In the third subsection we compare the model selection tests with other approaches in the literature, i.e., with information criteria, non-nested tests, and the encompassing principle.

3.1. The SC_n statistic

Theorem 1 (Asymptotic Distribution of $SC_n(\widehat{\theta}_{1n}, \widehat{\theta}_{2n})$)
Given Assumptions 1 to 6 for D_1 and D_2,

i) if $\lim_{n \to \infty} P\{Q_{1n}(\omega, \theta_{1n}^*) = Q_{2n}(\omega, \theta_{2n}^*)\} = 1$, then

$$2nSC_n(\widehat{\theta}_{1n}, \widehat{\theta}_{2n}) \xrightarrow{D} M_{p+q}(\cdot; \lambda), \qquad (3.1)$$

where λ is the vector of eigenvalues of $W_n \Sigma_n$, with

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

ii) if $\lim_{n \to \infty} P\{Q_{1n}(\omega, \theta_{1n}^*) = Q_{2n}(\omega, \theta_{2n}^*)\} = 0$, then

$$\sigma_n^{-1} n \left[SC_n(\widehat{\theta}_{1n}, \widehat{\theta}_{2n}) - \left(\overline{Q}_{1n}(\theta_{1n}^*) - \overline{Q}_{2n}(\theta_{2n}^*) \right) \right] \xrightarrow{D} N(0,1). \blacksquare \qquad (3.2)$$

Theorem 1 indicates that the asymptotic distribution of $SC_n(\widehat{\theta}_{1n}, \widehat{\theta}_{2n})$ depends on whether, asymptotically, $Q_{1n}(\theta_{1n}^*) = Q_{2n}(\theta_{2n}^*)$ or $Q_{1n}(\theta_{1n}^*) \neq Q_{2n}(\theta_{2n}^*)$, i.e., on whether the models are asymptotically nested and $\theta_{1n}^* = \phi(\theta_{2n}^*)$, or D_1 and D_2 are non-nested. Notice, in particular, that the additional information provided by the condition $Q_{1n}(\theta_{1n}^*) = Q_{2n}(\theta_{2n}^*)$ increases the speed of convergence by $\sqrt{n}$. From Lemma 2, the crucial condition $\lim_{n \to \infty} P\{Q_{1n}(\omega, \theta_{1n}^*) = Q_{2n}(\omega, \theta_{2n}^*)\} = 1$ holds if and only if $\lim_{n \to \infty} \sigma_n^2 = 0$. Hence, a test for its validity can be based on whether an estimator of σ_n^2, $\widehat{\sigma}_n^2$, is significantly different from zero or not. We have

Theorem 2 (Asymptotic Distribution of $\widehat{\sigma}_n^2$)
Given Assumptions 1-6 for D_1 and D_2, if $Q_{in}(\omega, \theta_{in}) = \psi_{in}(\omega, \theta_{in})$, $i = 1, 2$, σ_n^2 is $O(n)$, and $((q_{1t}(\omega, \theta_{1n}^*) - q_{2t}(\omega, \theta_{2n}^*)), F^t)$ is a martingale difference sequence, then

i) $n^{-1}(\widehat{\sigma}_n^2 - \sigma_n^2) = n^{-1}\left(\sum_{t=1}^n (q_{1t}(\widehat{\theta}_{1n}) - q_{2t}(\widehat{\theta}_{2n}))^2 - \sigma_n^2\right) \xrightarrow{p} 0,$

ii) Under $H_{0\sigma} : \lim_{n\to\infty} \sigma_n^2 = 0$,

$$\widehat{\sigma}_n^2 \xrightarrow{D} M_{p+q}(\cdot; \eta), \tag{3.3}$$

where η is the vector of eigenvalues of $V_n \Sigma_n$, with

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

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

The hypothesis $Q_{in}(\omega, \theta_{in}) = \psi_{in}(\omega, \theta_{in})$ holds for several estimation methods such as MLE or NLS. The requirement that $((q_{1t}(\omega, \theta_{1n}^*) - q_{2t}(\omega, \theta_{2n}^*)), F^t)$ is a martingale difference sequence let us use the estimator for the variance in i). This assumption can be easily relaxed to deal with an m-dependent sequence, at the cost of additional complications in the notation. The derivation of the distribution of an heteroskedastic-autocorrelation consistent estimator, such as $\widehat{B}_{in}$ in (2.8), is left for future research.

The following Corollary presents conditions under which the statistics in (3.1) and (3.3) converge to a central chi-square distribution.

Corollary 1 (χ^2 Distribution of $SC_n(\widehat{\theta}_{1n}, \widehat{\theta}_{2n})$ and $\widehat{\sigma}_n^2$)
Given assumptions 1-6 for D_1 and D_2, if $Q_{in}(\omega, \theta_{in}) = \psi_{in}(\omega, \theta_{in})$, $A_{in}(\theta_{in}^*) = -B_{in}(\theta_{in}^*)$, $i = 1, 2$, $\lim_{n\to\infty} P\{\psi_{1n}(\omega, \theta_{1n}^*) = \psi_{2n}(\omega, \theta_{2n}^*)\} = 1$, and

$$\lim_{n\to\infty} (B_{2n}(\theta_{2n}^*) - B_{21n}(\theta_{2n}^*, \theta_{1n}^*)B_{1n}^{-1}(\theta_{1n}^*)B_{12n}(\theta_{1n}^*, \theta_{2n}^*)) = 0, \tag{3.4}$$

then

i) $2nSC_n(\widehat{\theta}_{1n}, \widehat{\theta}_{2n}) \xrightarrow{D} \chi_{p-q}^2$

ii) $\widehat{\sigma}_n^2 \xrightarrow{D} \chi_{p+q-r}^2$, where $r = 2rank[B_{12n}(\theta_{1n}^*, \theta_{2n}^*)]$. ∎

11

The condition in (3.4) requires $n^{1/2}\nabla_\theta Q_{2n}(\theta_{2n}^*)$ to lie, asymptotically, in the space spanned by $n^{1/2}\nabla_\theta Q_{1n}(\theta_{1n}^*)$. The hypothesis $Q_{in}(\psi_{in}) = \psi_{in}$ and the information matrix equivalence $(A_{in}(\theta_{in}^*) = -B_{in}(\theta_{in}^*))$ are not strictly necessary to obtain the required result, but they lead to a substantial simplification in the calculations. From Rao and Mitra (1971, Theorem 9.2.1) a necessary and sufficient condition for the asymptotic distribution of $2nSC_n(\widehat{\theta}_{1n}, \widehat{\theta}_{2n})$ $(\widehat{\sigma}_n^2)$ to be chi-square is $\lim_{n\to\infty}(\Sigma_n W_n \Sigma_n W_n \Sigma_n - \Sigma_n W_n \Sigma_n) = 0$ $(\lim_{n\to\infty}(\Sigma_n V_n \Sigma_n V_n \Sigma_n - \Sigma_n V_n \Sigma_n) = 0)$.

To conclude, notice that when $Q_{in}(\omega, \theta_{in})$ is equal to (minus) the likelihood function for model D_i, $i = 1, 2$, 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. In the i.i.d. case and for MLE it is $Q_{in}(\omega, \theta_{in}) = \psi_i(\omega, \theta_i)$, $i = 1, 2$, which substantially simplifies the derivations.

3.2. Model selection for non-nested, nested and overlapping models

We formulate the hypothesis of interest when comparing D_1 and D_2 as

$$H_0 : \lim_{n\to\infty} \sqrt{n}\left(\overline{Q}_{1n}(\theta_{1n}^*) - \overline{Q}_{2n}(\theta_{2n}^*)\right) = 0, \tag{3.5}$$

namely, the difference of the loss functions from D_1 and D_2 is $o_p(n^{-1/2})$ when evaluated at the expected value of $\psi_n(\theta_n^*)$, and in this sense the two models are asymptotically equivalent. When $Q_{in}(\omega, \theta_{in}) = \psi_{in}(\omega, \theta_{in})$, $i = 1, 2$, under H_0 the difference of the expected loss functions is $o_p(n^{-1/2})$. For MLE and i.i.d. observations H_0 is equivalent to the null hypothesis in Vuong (1989), i.e., $E_P(q_1(\omega, \theta_1^*)) = E_P(q_2(\omega, \theta_2^*))$, and our results boil down to his.

The alternative hypothesis is either

$$H_1 : \limsup_{n\to\infty} \sqrt{n}\left(\overline{Q}_{1n}(\theta_{1n}^*) - \overline{Q}_{2n}(\theta_{2n}^*)\right) = -\infty, \tag{3.6}$$

or

$$H_2 : \liminf_{n\to\infty} \sqrt{n}\left(\overline{Q}_{1n}(\theta_{1n}^*) - \overline{Q}_{2n}(\theta_{2n}^*)\right) = +\infty. \tag{3.7}$$

Under H_1, D_1 is preferred to D_2 $(\overline{Q}_{1n}(\theta_{1n}^*) < \overline{Q}_{2n}(\theta_{2n}^*))$, and viceversa under H_2.

¿From Lemma 1, $SC_n(\widehat{\theta}_{1n}, \widehat{\theta}_{2n})$ converges to $\left(\overline{Q}_{1n}(\theta_{1n}^*) - \overline{Q}_{2n}(\theta_{2n}^*)\right)$. Hence, it is a natural candidate as a statistic to test for H_0 against H_1 or H_2. For the comparison of non-nested models (Definition 2), we have

Corollary 2 (Non-nested models)
Given Assumptions 1-6 for D_1 and D_2, if D_1 and D_2 are asymptotically non-nested, then

   i) under H_0, $\widehat{\sigma}_n^{-1} n SC_n(\widehat{\theta}_{1n}, \widehat{\theta}_{2n}) \xrightarrow{D} N(0,1)$.

   ii) under H_1, $\widehat{\sigma}_n^{-1} n SC_n(\widehat{\theta}_{1n}, \widehat{\theta}_{2n}) \xrightarrow{p} -\infty$.

   iii) under H_2, $\widehat{\sigma}_n^{-1} n SC_n(\widehat{\theta}_{1n}, \widehat{\theta}_{2n}) \xrightarrow{p} +\infty$. ∎

Hence, the test is two-sided and directional. For a given significance level, a critical value, c, is determined from the standard normal distribution. If $\widehat{\sigma}_n^{-1} n SC_n(\widehat{\theta}_{1n}, \widehat{\theta}_{2n})$ falls within the acceptance region, D_1 and D_2 are asymptotically equivalent. Otherwise, $\widehat{\sigma}_n^{-1} n SC_n(\widehat{\theta}_{1n}, \widehat{\theta}_{2n}) < -c \ (> c)$ implies that $D_1 \ (D_2)$ is preferred.

To evaluate nested models, we have

Corollary 3 (Nested models)
Given Assumptions 1-6 for D_1 and D_2, if D_2 is asymptotically nested in D_1, then

   i) under H_0, for any $z \geq 0$, $Pr\{-2nSC_n(\widehat{\theta}_{1n}, \widehat{\theta}_{2n}) \leq z\} - M_{p+q}(z; \widehat{\lambda}) \xrightarrow{p} 0$,
      where $\widehat{\lambda}$ is the vector of eigenvalues of $\widehat{W}_n \widehat{\Sigma}_n$.

   ii) under H_1, $-2nSC_n(\widehat{\theta}_{1n}, \widehat{\theta}_{2n}) \xrightarrow{p} +\infty$. ∎

The test is one sided, because, when D_2 is nested in D_1, H_1 cannot hold. 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. Under MLE estimation and when the models are defined as in Smith (1994, p.4), the distribution in i) coincides with that in Smith (1994, Corollary 3.1). With the additional hypothesis that the information matrix equivalence $(A_{in}(\theta_{in}^*) = -B_{in}(\theta_{in}^*))$ holds, it can be shown that the condition in (3.4) is satisfied, and the limiting distribution reduces to χ_{p-q}^2, which coincides with the distribution in Gallant and White (1988, Theorem 7.8).

For the comparison of overlapping models, we have

Corollary 4 (Overlapping models)
Given Assumptions 1-6 for D_1 and D_2,

i) If $\lim_{n\to\infty} P\{Q_{1n}(\omega, \theta_{1n}^*) = Q_{2n}(\omega, \theta_{2n}^*)\} = 0$, then under H_0, $\widehat{\sigma}_n^{-1} nSC_n(\widehat{\theta}_{1n}, \widehat{\theta}_{2n}) \xrightarrow{D}$ $N(0,1)$, under H_1, $\widehat{\sigma}_n^{-1} nSC_n(\widehat{\theta}_{1n}, \widehat{\theta}_{2n}) \xrightarrow{p} -\infty$, under H_2, $\widehat{\sigma}_n^{-1} nSC_n(\widehat{\theta}_{1n}, \widehat{\theta}_{2n}) \xrightarrow{p}$ $+\infty$.

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

When $Q_{in}(\omega, \theta_{in}) = \psi_{in}(\omega, \theta_{in})$, $i = 1, 2$, a sequential testing procedure can be adopted to decide whether the statistic in i) or ii) has to be used. More generally, this procedure can be followed whenever it is not known whether for the two models under comparison $\lim_{n\to\infty} P\{\psi_{1n}(\omega, \theta_{1n}^*) = \psi_{2n}(\omega, \theta_{2n}^*)\}$ is equal to zero or to one. In the first step the hypothesis $H_{0\sigma}$ is tested with the statistic in (3.3), using $\widehat{\eta}$ instead of η, where $\widehat{\eta}$ is the vector of eigenvalues of $\widehat{V}_n \widehat{\Sigma}_n$. If $H_{0\sigma}$ is accepted, the test for H_0 is conducted using the statistic in ii); otherwise that in i). 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 this broader context.

Finally, it is worth pointing out that when Assumption 5-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 7). In this case the asymptotic size of the tests will be lower than the nominal size, i.e., the tests are biased towards acceptance of the null hypothesis.

3.3. Comparison with other approaches

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) HQ. Model 1 is selected if $IC_n > 0$, where

$$IC_n(\omega) = \sum_{t=1}^n q_{1t}(\omega, \widehat{\theta}_{1n}) - \sum_{t=1}^n q_{2t}(\omega, \widehat{\theta}_{2n}) - \widehat{c}_n(\omega), \qquad (3.8)$$

and $q_t(\omega, \theta)$ is equal to the likelihood function, so that the term $n^{-1}(\sum_{t=1}^n q_{1t}(\omega, \widehat{\theta}_{1n}) - \sum_{t=1}^n q_{2t}(\omega, \widehat{\theta}_{2n}))$ is an estimate of the average Kullback-Leibler (1951) Information Criterion (KLIC). The penalty function $\widehat{c}_n(\omega)$ favors the selection of a

parsimonious model, and it is typically (but not necessarily) a non stochastic sequence, e.g., $\widehat{c}_n(\omega) = (p - q)$ for AIC, $\widehat{c}_n(\omega) = ((p - q)\log n)/2$ for SIC, $\widehat{c}_n(\omega) = (p - q)c\log(\log n)$ 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 5 (Weak consistency of IC_n)

Given Assumptions 1-6 for D_1 and D_2,

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

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

   iii) If $n(\overline{Q}_{1n}(\omega, \theta_{1n}^*) - \overline{Q}_{2n}(\omega, \theta_{1n}^*))$ is $O_p(1)$ and $P\{\omega : \widehat{c}_n(\omega) \to \infty\} = 1$, then $\lim_{n\to\infty} P\{\omega : IC_n(\omega) \leq 0\} = 1$. ∎

The conditions in Corollary 5 turn out to be the same as those in Sin and White (1996, Proposition 4.2). For example, 5-i) indicates that when D_1 is preferred to D_2 on the basis of the likelihood function, it will be selected by the information criterion with probability approaching one if the penalty function converges to zero at the rate n. 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. Notice that weak consistency of selection is preserved even if the estimators in (3.8) are not ML, as long as the conditions in Theorem 1 are satisfied.

It can also be of interest to test whether two models yield values for the information criteria that are not statistically different from each other, i.e., whether $IC_n = 0$. Given that

$$IC_n(\omega) = nSC_n(\omega) - \widehat{c}_n(\omega), \tag{3.9}$$

if $\widehat{c}_n(\omega)$ is $o_p(1)$ when the models are non-nested, or $\widehat{c}_n(\omega)$ is $o_p(n^{1/2})$ when the models are nested, then the asymptotic distribution for IC_n 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 $\widehat{c}_n(\omega)$.

Another approach to model selection that has received considerable attention since the pioneering work of Cox (1961, 1962) is that of non-nested testing. The implicit null hypothesis of the Cox statistic for testing D_1 against D_2 can be written as (Aguirre-Torres and Gallant (1983), White (1982), Vuong (1989)),

$$H_{0C} : E_P\left(\psi_1(\omega, \theta_1^*) - \psi_2(\omega, \theta_2^*)\right) = E_{D_1}\left(\psi_1(\omega, \theta_1^*) - \psi_2(\omega, \theta_2^*)\right). \tag{3.10}$$

For models of i.i.d. processes estimated by MLE, Vuong (1989, p. 319) noticed that H_{0C} in (3.10) cannot hold under H_0 in (3.5), when the models are strictly non-nested. Actually, for (3.10) to hold under H_0 it must be $\psi_1(\omega, \theta_1^*) = \psi_2(\omega, \theta_2^*)$, that cannot be valid when the models are strictly non-nested (Assumption 4-i is automatically satisfied for ML estimation). Hence, H_{0C} and H_0 are mutually exclusive. This point remains valid after relaxing the hypothesis of i.i.d. variables as in Assumption 1, and for overlapping models when $\sigma_n^2 > 0$. The Cox test is not suited for comparison of D_1 and D_2 when they are nested, or overlapping with $\sigma_n^2 = 0$: the statistic degenerates (e.g., Gourieroux and Monfort (1994, p. 2605)).

A third approach to model comparison is that of encompassing, e.g., Mizon and Richard (1986), Gourieroux and Monfort (1995), Smith (1994). The underlying idea is that D_1 can be preferred to D_2 when the former can correctly predict a set of characteristics of interest of the latter, typically its parameters. In the formulation of Smith (1994) for models estimated by MLE, D_1 encompasses D_2 when

$$H_{0E} : E_P\left(\psi_{2n}(\omega, \theta_{2n}^*) - \psi_{2n}(\omega, b(\theta_{1n}^*))\right) = 0, \tag{3.11}$$

where $b(\theta_{1n})$ is the asymptotic binding function defined by

$$b(\theta_{1n}) = \arg\max_{\theta_{2n} \in \Theta_2} E_{D_1}(\psi_{2n}(\omega, \theta_{2n})).$$

A comparison of H_{0E} in (3.11) and H_0 in (3.5) suggests that the former is neither necessary nor sufficient for the latter. This statement remains valid in the case of mutual encompassing, i.e. when D_1 encompasses D_2 and D_2 encompasses D_1.

Further details on the three approaches to model selection discussed in this subsection can be found, e.g., in Marcellino and Mizon (1999).

4. Examples

To illustrate the theoretical results, we now apply the SC_n test for model selection in the case of two non-nested but possibly overlapping, models, ARMAX versus STAR, and two nested ARMAX-GARCH models, estimated by ML.

For the ARMAX model, it is

$$
Q_{1n}(\omega, \theta_1) \equiv n^{-1} \sum_{t=1}^{n} q_{1t}(\omega, \theta_1), \tag{4.1}
$$

$$
q_{1t}(\omega, \theta_1) \equiv -\frac{1}{2} \log(v_t^2) - \frac{1}{2} \eta_t^2(\omega, \theta_1) v_t^{-2}(\omega, \theta_1),
$$

$$
\eta_t(\omega, \theta_1) \equiv Y_t - \sum_{i=1}^{k} b_i Z_{t-i} - \sum_{i=1}^{l} a_i \eta_{t-i}(\omega, \theta_1).
$$

For the STAR model, it is

$$
Q_{2n}(\omega, \theta_2) \equiv n^{-1} \sum_{t=1}^{n} q_{2t}(\omega, \theta_2), \tag{4.2}
$$

$$
q_{2t}(\omega, \theta_2) \equiv - \left\{ Y_t - \sum_{i=1}^{k} [c_i + d_i F(\gamma_0 Z_{t-k-1} - \gamma_1)] Z_{t-i} \right\}^2.
$$

The SC_n statistic for ARMAX-GARCH versus STAR is

$$
SC_n(\widehat{\theta}_{1n}, \widehat{\theta}_{2n}) = Q_{1n}(\omega, \widehat{\theta}_1) - Q_{2n}(\omega, \widehat{\theta}_2). \tag{4.3}
$$

For both models, we assume that $\{Y_t\}$ and $\{Z_{t-i},\ i = 0, 1..., k\}$ are strictly stationary and ergodic processes. 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 $q_{it}(\omega, \theta_i)$ and $\nabla_{\theta_i} q_{it}(\omega, \theta_i)$ are near epoch dependent, and satisfy a central limit theorem, for $i = 1, 2$. Under the additional hypothesis that $((q_{1t}(\omega, \theta_{1n}^*) - q_{2t}(\omega, \theta_{2n}^*), F^t)$ is a martingale difference sequence, we can also define

$$
\widehat{\sigma}^2 = n^{-1} \sum_{t=1}^{n} (q_{1t}(\widehat{\theta}_1) - q_{2t}(\widehat{\theta}_2))^2. \tag{4.4}
$$

17

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 4, and it depends on whether $\lim_{n\to\infty} P\{Q_{1n}(\omega, \theta_{1n}^*) = Q_{2n}(\omega, \theta_{2n}^*)\}$ is equal to zero or to one. Thus, we follow the two-step procedure described in Section 3.2. First, we test for $\lim_{n\to\infty} \sigma^2 = 0$ using the asymptotic distribution of $\widehat{\sigma}^2$ given in Theorem 2. Notice that η can be estimated by $\widehat{\eta}$, 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.[3] Second, if the hypothesis $\lim_{n\to\infty} \sigma^2 = 0$ is accepted, it is

$$2n SC_n(\widehat{\theta}_{1n}, \widehat{\theta}_{2n}) \xrightarrow{D} M_{p+q}(\cdot; \widehat{\lambda}), \tag{4.5}$$

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

$$\widehat{\sigma}^{-1} n SC_n(\widehat{\theta}_{1n}, \widehat{\theta}_{2n}) \xrightarrow[H_0]{D} N(0, 1). \tag{4.6}$$

Rejection of the null hypothesis in the second step provides evidence in favor of the ARMAX (STAR) model when SC_n is positive (negative).

As far as the comparison of nested models is concerned, an ARMAX-GARCH model is specified as in (4.1) with the additional recursive relationship

$$v_t^2(\omega, \theta_1) \equiv \alpha_0 + \sum_{i=1}^{c} \beta_i v_{t-i}^2(\omega, \theta_1) - \sum_{i=1}^{d} \alpha_i \eta_{t-i}^2(\omega, \theta_1). \tag{4.7}$$

The nested model is obtained by imposing $(p - q) > 0$ zero restrictions in (4.1) and/or (4.7). The SC_n statistic can be written as

$$SC_n(\widehat{\theta}_{1n}, \widetilde{\theta}_{1n}) = Q_{1n}(\omega, \widehat{\theta}_1) - Q_{1n}(\omega, \widetilde{\theta}_1), \tag{4.8}$$

where $\widetilde{\theta}_1$ is the restricted estimator of θ_1. In this case, the distribution of the SC_n test is that in Corollary 3, namely,

$$2n SC_n(\widehat{\theta}_{1n}, \widetilde{\theta}_{1n}) \xrightarrow{D} M_{p+q}(\cdot; \widehat{\lambda}). \tag{4.9}$$

A similar approach can be followed for the comparison of nested STAR models.

---

[3]For MLE the expression for S_{int}^* in (2.6) simplifies to $S_{int}^* = -\nabla_\theta q_{it}(\theta_{in}^*)'$, so that $\widehat{S}_{int}$ in (2.8) becomes $\widehat{S}_{int} = -\nabla_\theta q_{it}(\widehat{\theta}_{in})'$, .

5. Conclusions

In this paper we have proposed a statistic for model selection that can be applied under weak assumptions on the processes and models under analysis. For i.i.d. processes and ML estimation, the statistic reduces to Vuong's (1989) test. Otherwise, it can be thought of as a generalization thereof. 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, possibly for non-stationary processes, and an evaluation of its finite sample behavior, together with a comparison with other model selection criteria, are left for future research.

APPENDIX

Proof of Lemma 1. Follows from Gallant and White (1988, Theorems 3.18, 3.19). ∎

Proof of Lemma 2. If $Q_{1n}(\theta_{1n}^*) = Q_{2n}(\theta_{2n}^*)$ *a.s.* then $\sigma_n^2 = 0$. If $\sigma_n^2 = 0$, then $Q_{1n}(\theta_{1n}^*) = Q_{2n}(\theta_{2n}^*) + C$. Under Assumption 4-iii), integrating both sides of this equality yields $C = 0$. ∎

Proof of Lemma 3. 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

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

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

Proof of Lemma 4. Follows from Lemma 3 and Vuong (1989, Lemma 3.2). ∎

Proof of Lemma 5. From a first order Taylor expansion of $Q_{in}(\theta_{in}^*)$ around $\overline{\psi}_{in}(\theta_{in}^*)$, $i = 1, 2$, it is

$$Q_{1n}(\theta_{1n}^*) - Q_{2n}(\theta_{2n}^*) = \overline{Q}_{1n}(\theta_{1n}^*) - \overline{Q}_{2n}(\theta_{2n}^*) + \nabla_\theta Q_{1n}(\overline{\psi}_{1n}(\theta_{1n}^*))(\psi_{1n}(\theta_{1n}^*) - \overline{\psi}_{1n}(\theta_{1n}^*)) - \nabla_\theta Q_{2n}(\overline{\psi}_{2n}(\theta_{2n}^*))(\psi_{2n}(\theta_{2n}^*) - \overline{\psi}_{2n}(\theta_{2n}^*)) + o_p(1).$$

Given Assumtpions 1 and 2 for D_1 and D_2, from Gallant and White (1988, Theorem 5.3), $\sqrt{n}(\psi_{1n}(\theta_{1n}^*) - \overline{\psi}_{1n}(\theta_{1n}^*))$ and $\sqrt{n}(\psi_{2n}(\theta_{2n}^*) - \overline{\psi}_{2n}(\theta_{2n}^*))$ have an asymptotically normal joint distribution, which yields the result. ∎

Proof of Lemma 6. Follows from Gallant and White (1988, Theorem 6.1). ∎

Proof of Lemma 7. Follows from Gallant and White (1988, Theorem 6.8). ∎

Proof of Theorem 1. From a first order Taylor expansion of $Q_n(\omega, \theta_n^*)$ around $\widehat{\theta}_n$, we obtain

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

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

so that

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

i) When $\lim_{n \to \infty} P\{Q_{1n}(\omega, \theta_{1n}^*) = \psi_{2n}(\omega, \theta_{2n}^*)\} = 1$, $SC_n(\theta_{1n}^*, \theta_{2n}^*)$ is $o_p(1)$. The asymptotic distribution of $2nSC_n(\widehat{\theta}_{1n}, \widehat{\theta}_{2n})$ follows from Lemmas 3 and 4.

ii) When $\lim_{n \to \infty} P\{Q_{1n}(\omega, \theta_{1n}^*) = \psi_{2n}(\omega, \theta_{2n}^*)\} = 0$, $\sigma_n^{-1} n(\widehat{\theta}_{in} - \theta_{in}^*) H_{in}(\theta_{in}^*)(\widehat{\theta}_{in} - \theta_{in}^*)$ are $o_p(1)$, $i = 1, 2$, and the result follows from Lemma 5. ∎

Proof of Theorem 2. i) Convergence to zero of $n^{-1}(\widehat{\sigma}_n^2 - \sigma_n^2)$ follows from Lemma 7. ii) As in Vuong (1989, p. 328-329), a Taylor series expansion of $n^{-1}\widehat{\sigma}_n^2$ around $(\theta_{1n}^*, \theta_{2n}^*)$ yields, under $H_{0\sigma} : \lim_{n \to \infty} \sigma_n^2 = 0$,

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

with,

$$\widetilde{V}_n = \begin{pmatrix} G_{1n}(\widetilde{\theta}_{1n}) & G_{12n}(\widetilde{\theta}_{1n}, \widetilde{\theta}_{2n}) \\ G_{21n}(\widetilde{\theta}_{2n}, \widetilde{\theta}_{1n}) & G_{2n}(\widetilde{\theta}_{2n}) \end{pmatrix}$$

$$G_{1n}(\widetilde{\theta}_{1n}) = n^{-1} \sum_{t=1}^{n} \nabla_\theta q_{1t}(\widetilde{\theta}_{1n}) \nabla_\theta q_{1t}(\widetilde{\theta}_{1n})' + n^{-1} \sum_{t=1}^{n} (q_{1t}(\widetilde{\theta}_{1n}) - q_{2t}(\widetilde{\theta}_{2n})) \nabla_\theta^2 q_{1t}(\widetilde{\theta}_{1n}),$$

$$G_{12n}(\widetilde{\theta}_{1n}, \widetilde{\theta}_{2n}) = n^{-1} \sum_{t=1}^{n} \nabla_\theta q_{1t}(\widetilde{\theta}_{1n}) \nabla_\theta q_{2t}(\widetilde{\theta}_{2n})',$$

$$G_{2n}(\widetilde{\theta}_{2n}) = n^{-1} \sum_{t=1}^{n} \nabla_\theta q_{2t}(\widetilde{\theta}_{2n}) \nabla_\theta q_{2t}(\widetilde{\theta}_{2n})' + n^{-1} \sum_{t=1}^{n} (q_{1t}(\widetilde{\theta}_{1n}) - q_{2t}(\widetilde{\theta}_{2n})) \nabla_\theta^2 q_{2t}(\widetilde{\theta}_{2n}),$$

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-A6, the second terms in $\widetilde{G}_{1n}$ and $\widetilde{G}_{2n}$ converge to zero in probability under $\lim_{n\to\infty} \sigma_n^2 = 0$, and, from Lemma 7, $\widetilde{V}_n \xrightarrow{p} V_n$. The distribution in (3.3) follows from Lemma 4. Divergence of $\widehat{\sigma}_n^2$ under $H_{1\sigma} : \lim_{n\to\infty} \sigma_n^2 > 0$ follows from i). ∎

Proof of Corollary 1. Follows from Rao and Mitra (1971, Theorem 9.2.1) and Vuong (1989, Corollary 3.4 and 4.4). ∎

Proof of Corollary 2. Follows from Assumption 4-i) and Theorem 1. ∎

Proof of Corollary 3. From a simple modification of Vuong (1989, Lemma 7.1), under H_0, it is $lim_{n\to\infty}\theta_{1n}^* - \phi(\theta_{2n}^*) = 0$. The latter, from Assumption 4-ii) implies that $lim_{n\to\infty} P\{Q_{1n}(\omega, \theta_{1n}^*) = Q_{2n}(\omega, \theta_{2n}^*)\} = 1$. The result then follows from Theorem 1. ∎

Proof of Corollary 4. Follows from Corollary 2 and 3. ∎

Proof of Corollary 5. i) follows from Corollary 2-ii) and 2)-iii). ii) follows from Corollary 2-i). iii) follows from Corollary 3-i). ∎

References

[1] Aguirre-Torres, V. and A.R. Gallant (1983), "The Null and Non-null Asymptotic Distribution of the Cox Test for Multivariate Non-Linear Regression", *Journal of Econometrics*, 21, 5-33.

[2] 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.

[3] Cox, D.R. (1961), "Test of Separate Families of Hypotheses", in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistic and Probability*, vol. I, Berkeley: University of California Press, 105-23.

[4] Cox, D.R. (1962), "Further Results on Tests of Separate Families of Hypotheses", *Journal of the Royal Statistical Society*, Series B, 24, 406-24.

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

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

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

[8] Gourieoroux, 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.

[9] Gourieoroux, C. and A. Monfort (1995), "Testing, Encompassing and Simulating Dynamic Econometric Models", *Econometric Theory*, 11, 195-228.

[10] 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.

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

[12] Marcellino, M. and G.E. Mizon (1999), *Progressive Modelling: Encompassing and Non-nested Testing*, Oxford: Oxford University Press (*forthcoming*).

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

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

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

[17] Rivers, D. and Q. Vuong (1999), "Model Selection Tests for Nonlinear Dynamic Models", *mimeo.*

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

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

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

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

[22] White, H. (1982), "Regularity Conditions for Cox's Test of Non-nested Hypotheses", *Journal of Econometrics*, 19, 301-18.

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

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

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