Lag Length Estimation in Large Dimensional Systems

Jesus Gonzalo
Universidad Carlos III de Madrid
jgonzalo@elrond.uc3m.es

Jean-Yves Pitarakis
University of Reading
J.Pitarakis@reading.ac.uk

November 2000

Abstract

We study the impact of the system dimension on commonly used model selection criteria (AIC,BIC, HQ) and LR based general to specific testing strategies for lag length estimation in VAR's. We show that AIC's well known overparameterization feature becomes quickly irrelevant as we move away from univariate models, with the criterion leading to consistent estimates under sufficiently large system dimensions. Unless the sample size is unrealistically small, all model selection criteria will tend to point towards low orders as the system dimension increases, with the AIC remaining by far the best performing criterion. This latter point is also illustrated via the use of an analytical power function for model selection criteria. The comparison between the model selection and general to specific testing strategy is discussed within the context of a new penalty term leading to the same choice of lag length under both approaches.

Keywords: Dimensionality, Information Criteria, Lag Length Selection, VAR.

JEL: C32, C52.

[1] Financial support from the Spanish Secretary of Education (PB980026) is gratefully acknowledged. Address for Correspondence: Jean-Yves Pitarakis, Department of Economics, University of Reading, PO Box 218, Whiteknights, Reading RG6 6AA, United-Kingdom. Tel: (0118) 9875 123 Ext. 7470, Fax: (0118) 9750 236

1 Introduction

The specification of a proper dynamic structure is a crucial preliminary step in univariate or multivariate ARMA type time series models. Although the determination of a proper lag structure is seldom of individual interest or the final objective of an empirical investigation, it has a great impact on subsequent inferences whether they are about causality, cointegration, impulse response analysis or forecasting. Typically the most common way of selecting an appropriate lag structure for a VAR involves first assuming that the true but unknown lag length is bounded by some finite constant and subsequently using information theoretic criteria such as the AIC (Akaike (1974)), BIC (Schwarz (1978)) or HQ (Hannan & Quinn (1979), Quinn (1980)) to determine an optimal lag length. This is clearly the most frequently used approach in the time series literature which abunds in studies that evaluated the asymptotic and finite sample properties of the above mentioned methods. On the theoretical side it has been shown that criteria such as the BIC and HQ lead to consistent estimates in both stationary and nonstationary systems (Hannan (1980), Quinn (1980), Tsay (1984), Paulsen (1984), Pötscher (1987) among others) while the AIC is characterized by a positive limiting probability of overfitting.

Focusing on the finite sample properties of lag length selection methods, Lütkepohl (1985) conducted an extensive Monte-Carlo study analyzing the properties of a large number of methods in bivariate and trivariate stationary VAR's. The overall conclusion of the study supported the view that the BIC and the HQ lead to the most accurate results. In the context of a cointegrated system, Cheung & Lai (1993) found that both the AIC and BIC perform well in finite samples provided that the true error structure has a finite and parsimonious autoregressive representation. If the system contains moving average components however, then both criteria displayed poor performance. In this latter case, the AIC led to lag length estimates as distorted as the ones obtained by the BIC in the sense of generating truncation lags that are too short for the finite autoregressive approximation to be reliable. This confirms a recent point by Ng and Perron (1995) who showed in the context of a univariate framework that despite its well known overfitting feature the AIC abandons information at long lag lengths and is therefore also unreliable under moving average components. Their analysis further suggests that a sequential testing strategy could be preferable under moving average errors, leading to a better size-power trade off in the subsequent inferences about the presence of unit roots. Overall however, our reading of the literature is that the AIC and BIC, still remain the favorite tools for specifying the lag structure in both univariate and multivariate models. More recently Ho and Sorensen (1996) analyzed the impact of the system dimension on the performance of LR based cointegration tests, and as a byproduct of their study concluded that the BIC is more reliable than the AIC in such a setting. The fact that the negative consequences of an underparameterized model are much more serious than in an overparameterized case (wrong inferences versus loss of efficiency for instance) however often led practitioners to argue in favor of the AIC criterion. These mixed and often contradictory conclusions, clearly highlight the point that it is difficult to come up with a universally accepted typology of methods ranked in terms of their performance. Indeed the number of factors influencing the behavior of these procedures is such that conclusions can only be DGP specific, with different parameterizations possibly leading to contradictory features for the same criterion. It is however possible to explain why most studies reached conflicting results by focusing mainly on the system dimension and available sample size together with the rates of convergence of the various model selection criteria . This can then allow us to better classify the circumstances under which a specific method will perform better than the others.

In this paper our objectives are twofold. First, we focus on a series of factors (system dimension, sample size, preset upper bound etc.) that influence the performance of alternative lag length selection methods in both small and large samples with the aim of explaining and clarifying the often conflicting results obtained in the literature. Our second objective is then to provide a set of practical guidelines for the choice of the lag order determination method. The plan of the paper is as follows. Section 2 will present the competing information theoretic methods and evaluate their theoretical features in relation to the dimensionality aspect. Section 3 focuses on the general to specific testing strategy and its connection with the model selection approach. Section 4 concludes. All proofs are relegated to the appendix.

2 Features of Commonly Used Lag Length Selection Criteria

In this section, we focus on some theoretical features of the penalized likelihood based methods for selecting the lag order in the following vector autoregression

$$X_t = \Phi_1 X_{t-1} + \ldots + \Phi_{p_0} X_{t-p_0} + \epsilon_t, \tag{1}$$

where X_t is a $K \times 1$ vector, p_0 denotes the unknown true lag length and

Assumptions: *(A1)* $\{\epsilon_t\}$ *is a gaussian i.i.d. vector sequence with mean zero and* $E(\epsilon_t \epsilon_t') = \Omega_\epsilon > 0$ $\forall t$, *(A2) The determinant of the autoregressive polynomial* $|\Phi(z)| = |I_K - \Phi_1 z - \ldots - \Phi_{p_0} z^{p_0}|$ *has all its roots outside the unit circle or at most* K *roots at* $z = 1$ *and the lag length* p_0 *is such that* $p_0 \leq p_{max}$ *with* p_{max} *denoting a known finite constant.*

Using Engle and Granger's (1987) terminology the above assumptions allow the vector autoregressive process in (1) to be purely stationary ($I(0)$), purely non-stationary ($I(1)$) or cointegrated ($CI(1,1)$). Given the above specification the primary objective of any investigation involving VAR models is the selection of an optimal value for p the unknown lag length. The general expression of the objective function of penalty based methods is given by

$$IC(p) = \log|\hat{\Omega}(p)| + \frac{c_T}{T} m_p \tag{2}$$

where $\hat{\Omega}(p)$ denotes the estimated residual covariance matrix when p lags have been fitted to (1), m_p the number of freely estimated parameters ($m_p = K^2 p$) and c_T a deterministic penalty term. When $c_T = 2$ we have the well known AIC criterion, $c_T = \log T$ corresponds to the BIC and $c_T = 2 \log\log T$ is commonly referred to as the HQ. The optimal lag length, say $\hat{p}$ is then selected as follows

$$\hat{p} = \arg \min_{0 \leq p \leq p_{max}} IC(p). \tag{3}$$

Regarding the asymptotic properties of $\hat{p}$ obtained from (3), Tsay (1984) and Paulsen (1984) showed that provided that $c_T \rightarrow \infty$ and $c_T/T \rightarrow 0$ as $T \rightarrow \infty$, $\hat{p}$ is consistent in both stationary and I(1) systems. Clearly the AIC criterion violates the first of the above two conditions leading to a non zero limiting probability of overfitting. It is worth pointing out however that even for the AIC the probability of underestimation vanishes asymptotically. These limiting results however, provide little guidance for the choice of a reliable criterion in finite samples.

2.1 Overfitting in Large Samples

The impact of the system dimension on the probability of overfitting of criteria such as the AIC can be analyzed by focusing on $P[IC(p_0 + h) < IC(p_0)]$ which represents the probability of overparameterizing the model with $h \geq 1$ extra variates. Numerous studies have shown that this probability does not vanish asymptotically for constant penalty criteria such as the AIC since the requirement that $c_T \to \infty$ is violated. This has often been used as a strong argument against the practice of model selection via the AIC. However, an important point established in Paulsen and Tjostheim (1985) in the context of a purely stationary VAR is that the AIC's nonzero asymptotic probability of overfitting is also a decreasing function of the system dimension. This feature of the AIC criterion seems to have often been overlooked in applied work. The following proposition will allow us to formally quantify the behaviour of the overfitting probability across different system dimensions for purely stationary, nonstationary and cointegrated systems and will illustrate the fact that even for a criterion such as the AIC the probability becomes rapidly negligible as we move from a univariate to a larger dimensional system.

Proposition 2.1 *Under assumptions (A1)-(A2) and letting $\hat{p}$ denote the lag length estimate obtained via the model selection approach using a constant penalty $c_T = c$, the probability of fitting h spurious variates beyond p_0 converges to $P[\chi^2(K^2\ h) > K^2\ h\ c]$ as $T \to \infty$ and $\forall p_0 \in [1, p_{max}]$ if the polynomial in (A2) has at least one root on the unit circle and $\forall p_0 \in [0, p_{max}]$ if it has all its roots outside the unit circle.*

The requirement that $p_0 \geq 1$ under the presence of $I(1)$ components ensures that lag length restrictions on the VAR in levels can be reformulated as restrictions on coefficient matrices of stationary regressors only, thus validating the use of standard asymptotics. Differently put when the polynomial in (A2) has at least one root on the unit circle, the quantity $T(\log|\hat{\Omega}(p)| - \log|\hat{\Omega}(p+h)|)$ will be asymptotically distributed as $\chi^2(K^2 h)$ only if $p \geq 1$. The above result highlights the crucial importance that the system dimension will have on the performance of model selection criteria and illustrates the fact that the probability of overfitting is an exponentially decreasing function of K in both stationary and nonstationary systems. For the AIC criterion for instance it is clear that one does not need an extremely large system dimension for the above probability to be close to zero and for practical purposes it can be considered as negligible even in the trivariate case. Indeed, under $K = 3$ for instance the limiting probabilities of selecting h extra variates beyond p_0 are given by 3.52%, 0.71% and 0.15% for $h = 1, 2$ and 3 respectively while when K=1 (i.e. univariate model) the corresponding figures increase to 17.73%, 13.43% and 11.16%. Thus even in moderately large systems the risk of overparameterization is negligible and therefore the AIC criterion may also lead to consistent like estimates since $\lim_{T\to\infty} P[AIC(p_0+1) < AIC(p_0)] = O(e^{-K^2})$. For the BIC and HQ criteria, the probability of overfitting converges to zero as $T \to \infty$ since for both criteria $c_T \to \infty$, implying that $\lim_{T\to\infty} P[IC(p_0+h) < IC(p_0)] = 0$. It is worth pointing out however the influence that the system dimension K will have on this latter probability. Specifically for the probability of fitting one spurious variate under the BIC we have

$$\begin{aligned} P[BIC(p_0+1) < BIC(p_0)] &= P[T \log \frac{|\hat{\Omega}(p_0)|}{|\hat{\Omega}(p_0+1)|} > K^2 \log T] \\ &\approx P[\chi^2(K^2) > K^2 \log T] \\ &\approx O\left(\frac{(\log T)^{\frac{K^2}{2}-1}}{T^{\frac{K^2}{2}}}\right) \end{aligned}$$

with the last approximation following from the asymptotic expansion of the incomplete gamma function. Proceeding similarly for the HQ type penalty, we obtain $P[HQ(p_0+1) < HQ(p_0)] \approx O((\log\log T)^{\frac{K^2}{2}-1}(\log T)^{-\frac{K^2}{2}})$. Thus although the BIC's convergence rate may appear as very desirable it also casts serious doubts on its ability to move away from the lowest possible lag length when the system dimension is large.

2.2 Overfitting in Finite Samples

So far the validity of our arguments has been conditional upon the availability of a sufficiently large sample size so as to ensure that the distribution of $T(\log|\hat{\Omega}(p_0)| - \log|\hat{\Omega}(p_0+h)|)$ is accurately approximated by a $\chi^2(K^2\ h)$ random variable. Typically in finite samples, the degrees of freedom limitations will introduce severe upward biases in the estimated covariance matrices resulting in a rightward shift of the empirical distribution relative to that of the theoretical χ^2. Thus despite the evidence from the above large sample based results, even in large dimensional systems the AIC criterion might still end up pointing to very high lag orders if the inflated $\chi^2(K^2\ h)$ dominates the deterministic term $2K^2\ h$. This effect could be particularly strong if a large system dimension is combined with a large value of the upper bound p_{max}. When this happens it would be innacurate to attribute the causes of the resulting overparameterization to AIC's "overfitting nature" since it arises solely from the degrees of freedom restrictions. The chances of this occurring for the BIC are negligible however since $K^2\ h \log T$ will be extremely large (at least twice as large as $2K^2\ h$) even for a relatively small T.

To gain further insight into this latter point we simulated data from a ten dimensional $VAR(p_0 = 1)$ using samples of size T=90, 150, 250 and 1000 and with a $VAR(p = 2)$ as the fitted model. The empirically obtained 95% critical values of the LR statistic for testing $\Phi_2 = 0$ were 184.08, 158.22, 143.15 and 129.88 respectively, compared with the theoretical $\chi^2_{95\%}(100)$ counterpart of 124.35. Since for the AIC criterion we have $K^2 \ h \ c_T = 100 \times 1 \times 2 = 200$ it is clear that under moderately small samples overfitting might occur frequently. For the BIC on the other hand even under T=90 we have $K^2 \ h \ c_T \simeq 500$ suggesting that overfitting is unlikely to occur no matter how inflated the finite sample distribution of LR is. Note that the above empirical percentiles were highly robust to the stationarity properties and parameter values of the DGP, having experimented across various stationary, purely nonstationary and cointegrated specifications. In finite samples and large dimensional systems AIC's overfitting feature will arise only if T is small relative to the system dimension K and the chosen upper bound p_{max}.

It is possible to be more explicit about this claim by using existing results on finite sample corrections. Indeed the important discrepancies between the finite sample and asymptotic distributions are a well documented issue in the multivariate analysis literature. Since Bartlett (1947), numerous authors introduced correction factors to various expressions of the likelihood ratio statistic in order to make the moments of the finite sample distributions match those of the asymptotic distribution, up to a certain order of magnitude. At this stage and for the clarity of the exposition it is useful to reformulate the IC based lag length selection problem by focusing on a slightly modified objective function we denote by $\overline{IC}(p) = IC(p) - IC(p_{max})$ with $IC(p)$ defined as in (2). Note that the selection of an optimal p by minimizing $\overline{IC}(p)$ is a program identical to the one in (3). The modified criterion can be written as

$$\overline{IC}(p) \quad = \quad T \log \frac{|\hat{\Omega}(p)|}{|\hat{\Omega}(p_{max})|} - K^2 c_T (p_{max} - p) \qquad (4)$$

where we can recognize the expression of the LR statistic, asymptotically distributed as $\chi^2(K^2(p_{max} - p))$, in its first right hand side component (note that in this modified framework we have $\overline{IC}(p_{max}) = 0$ by construction).

By appealing to existing results on finite sample corrections it is now possible to gain further insight on the effects that a limited sample size might have on the choice of p. In the context of VAR models, Sims (1980) for instance proposed a finite sample correction to the LR statistic based on replacing the normalizing factor T by $T-\delta$ with δ denoting the number of parameters estimated in each equation of the model. Within the above framework therefore, the small sample adjusted LR statistic is given by $LR^c = (T - Kp_{max}) \log(|\hat{\Omega}(p)|/|\hat{\Omega}(p_{max})|)$. Interestingly, this correction is also equivalent to the theoretically derived adjustment obtained by Fujikoshi (1977) in the context of static canonical correlation analysis. To our knowledge an explicit and theoretically derived small sample adjustment for the LR statistic does not exist in the VAR literature, however numerous simulation studies (Reinsel and Ahn (1992), Cheung and Lai (1993), Gonzalo and Pitarakis (1995, 1998, 1999) among others) have shown that the above correction improves significantly upon the raw LR statistic in both stationary or cointegrated VAR's and is commonly used in the time series literature. It is also important to point out that this simple small sample adjustment has often been criticized on the grounds that it does not always provide a good approximation of the tail areas, allowing solely a good match of the first moment of LR^c with that of a $\chi^2(K^2(p_{max}-p))$ random variable. This is potentially a serious problem when the adjusted statistic is used for hypothesis testing, here however our focus being on expected values rather than tail areas it should serve our purpose quite accurately. Indeed, our motivation here is to obtain a quantitative indication of the average ability of the IC approach not to overfit. Consider for instance the quantity $E[\overline{IC}(p_{max}) - \overline{IC}(p_0)]$ and let us focus on the loose requirement that on average the model selection procedure selects p_0 over p_{max}. Using the expression of $\overline{IC}(p)$ in (4) the requirement that $E[\overline{IC}(p_0)] < 0$ can be written as

$$K^2 c_T(p_{max} - p_0) \quad > \quad E\left[T \log \frac{|\hat{\Omega}(p_0)|}{|\hat{\Omega}(p_{max})|}\right].$$

Next, assuming that the distribution of LR^c is accurately approximated by the asymptotic $\chi^2(K^2(p_{max} - p_0))$ even for moderately small magnitudes of T and rewriting the above expression as

$$K^2 c_T(p_{max} - p_0) \quad > \quad \frac{T}{T - Kp_{max}} E\left[(T - Kp_{max}) \log \frac{|\hat{\Omega}(p_0)|}{|\hat{\Omega}(p_{max})|}\right]$$

and making use of the fact that $E[\chi^2(K^2(p_{max} - p_0)] = K^2(p_{max} - p_0)$ leads to the requirement that

$$T \quad > \quad \frac{Kc_T\, p_{max}}{c_T - 1} \tag{5}$$

for $E[\overline{IC}(p_0)] < 0$ to hold. As an illustration, consider the case of the AIC criterion with $K = 10$, $p_{max} = 7$ and $p_0 = 1$. In order to ensure that "on average" p_0 is chosen over p_{max} we would need $T > 140$. Regardless of the DGP's parameter structure if the above condition is not satisfied the AIC will often point to lag lengths greater than p_0. This simple scenario has occurred quite frequently in applied work but its reasons have been attributed solely to AIC's natural tendency to overfit. In Ho & Sorensen (1996) for instance, the authors estimated a seven dimensional VAR with $p_{max} = 4$ and found that the AIC was systematically selecting $\hat{p} = 4$. Our previous results provide a clear explanation for this finding and highlight the dangers of using the AIC under these conditions. In the case of the HQ criterion the requirement drops to $T > 104$ and for the BIC we would need $T > 90$. Thus although when K is large the AIC does not overfit asymptotically, in small samples (small compared to p_{max} and K) it might repeatedly select the preset upper bound if the latter is not chosen carefully. In summary, with moderate or large sample sizes none of the model selection criteria will overfit. This is also true for the AIC in large dimensional systems with T sufficiently large relative to p_{max} and K. If T is small relative to p_{max} and K (ie. $T < Kc_T \ p_{max}/(c_T - 1)$ for instance) then the AIC criterion and to a lesser extent HQ might frequently point to lag lengths close to the upper bound. In those instances our analysis suggests that it might be beneficial to adjust the LR component of $\overline{IC}(p)$ in a way similar to the Bartlett type small sample adjustment applied to the LR statistic.

2.3 Underfitting

Regarding the probability of underfitting, it is well known that for all criteria it vanishes asymptotically regardless of the location of the roots of $|\Phi(z)| = 0$. Indeed, if we consider the probability of selecting p_0 over $p_0 - 1$ for instance, we have

$$(6) \qquad P[IC(p_0) < IC(p_0 - 1)] \ = \ P[\log|\hat{\Omega}(p_0 - 1)| - \log|\hat{\Omega}(p_0)| > \frac{c_T K^2}{T}],$$

and since $|\hat{\Omega}(p_0)| < |\hat{\Omega}(i)| \ \forall i = 1, \ldots, p_0 - 1$, the probability in (6) will converge to one provided that c_T/T tends to zero. Most finite sample simulation studies however found that criteria such as the BIC might often lead to an overly parsimonious model. The AIC on the other hand has been rarely found to underfit. Here we argue that if the sample size is moderate (greater than $Kc_T \ p_{max}/(c_T - 1)$ for instance) and the system dimension large, all criteria including the AIC might lead to lag lengths artificially clustered at very low levels. The problem will arise from the K^2 term adjacent to c_T/T in (6) which even for T moderately large may leave the factor $K^2 c_T/T$ too high. This suggests that in finite samples an increased system dimension may adversely affect the probability of underfitting by increasing the possibility of selecting underspecified models. This phenomenon can be illustrated by focusing on a set of simple generic models which will also allow us to isolate the impact of the stationarity properties of the system. We initially consider the following K dimensional VAR driven by VAR(1) errors

$$\Delta X_t = u_t \qquad u_t = Ru_{t-1} + \epsilon_t$$

with $R = diag(\rho_1, \ldots, \rho_K)$, $|\rho_i| < 1$ for $i = 1, \ldots, K$, and ϵ_t denoting a gaussian vector white noise process with $E(\epsilon_t \epsilon_t') = \Omega_\epsilon > 0$. The above model can be rewritten as $\Delta X_t = \Pi X_{t-1} + R\Delta X_{t-1} + \epsilon_t$ with $\Pi = 0$. Equivalently it can also be viewed as a $VAR(p_0 = 2)$ in levels. Suppose that instead of the true model we fit $\Delta X_t = \Psi X_{t-1} + \nu_t$ thus omitting the relevant lagged dependent variable. It is well known that in this underfitted model we will continue to have $\hat{\Psi}_{ols} \xrightarrow{p} 0$ due to the I(1)'ness of the components of X_t. Letting $\hat{\Omega}_1$ denote the residual covariance matrix obtained from the underfitted model, it is then straightforward to show that $\hat{\Omega}_1 = diag(\frac{1}{1-\rho_1^2}, \ldots, \frac{1}{1-\rho_K^2})\Omega_\epsilon + o_p(1)$. On the other hand, if we were fitting the correct model we would have $\hat{\Omega}_2 = \Omega_\epsilon + o_p(1)$ with $\hat{\Omega}_2$ denoting the residual covariance from the correctly specified model. Recalling the general expression of the model selection criteria given in (2) and putting $\Omega_\epsilon = I_K$ we have

$$IC(p = 2) - IC(p = 1) \quad = \quad K^2 \frac{c_T}{T} + \sum_{i=1}^{K} \log(1 - \rho_i^2) + o_p(1) \tag{7}$$

and since $p_0 = 2$, we need

$$K^2 \frac{c_T}{T} + \sum_{i=1}^{K} \log(1 - \rho_i^2) + o_p(1) \quad < \quad 0, \tag{8}$$

so as not to underfit. Although (8) will always hold asymptotically, if K is large it is very likely that even a criterion such as the AIC might underfit since the "negativity" of $\sum_{i=1}^{K} \log(1 - \rho_i^2)$ will be masked by a large value of $K^2 \frac{c_T}{T}$. This also highlights the reason why a criterion such as the BIC for which $K^2 c_T / T$ is likely to be very large relative to the second negative component in (8) might persistently point to lag lengths below p_0. The above results also illustrate the importance of the magnitudes of the chosen values for the parameters driving the error process. It is clear that one needs to be cautious when properties of model selection criteria are established under specific DGP's. In fact virtually any property can be obtained by a proper manipulation of the parameters of the DGP.

The above example can also be used to assess the influence of the stationarity properties of the data on the probability of underfitting. Indeed instead of focusing on a system of I(1) variables, we can consider the following $VAR(p_0 = 2)$ specification

$$
\begin{aligned}
X_t &= AX_{t-1} + u_t \\
u_t &= Ru_{t-1} + \epsilon_t
\end{aligned}
$$

where for simplicity we let $A = diag(\alpha_1, \ldots, \alpha_K)$, and R and Ω_ϵ defined as above. In a purely stationary system $|\alpha_i| < 1 \; \forall i = 1, \ldots, K$. In this context it is straightforward to show that $\log|\hat{\Omega}(p = 1)| \xrightarrow{p} -\sum_{i=1}^{K} \log(1 - \rho_i^2 \alpha_i^2)$ and $\log|\hat{\Omega}(p = 2)| \xrightarrow{p} 0$. We can therefore write

$$
IC(p = 2) - IC(p = 1) = K^2 \frac{c_T}{T} + \sum_{i=1}^{K} \log(1 - \rho_i^2 \alpha_i^2) + o_p(1) \tag{9}
$$

which can be compared with (8) in the purely nonstationary case. It is clear that when $\alpha_i = 1 \; \forall i$, the possibility that $IC(p = 2) < IC(p = 1)$ will be much greater than when $|\alpha_i| < 1$, suggesting that in finite samples the presence of unit roots will help push the inequality in the desired direction. Finally for the cointegrated case we consider the specification given by $X_{it} = \alpha_i X_{it-1} + u_{it}$ with $|\alpha_i| < 1$ for $i = 1, \ldots, r$ and $\Delta X_{it} = u_{it}$ for $i = r + 1, \ldots, K$ and the $u_{it}'s$ specified as above. Thus the true model is now a $VAR(p_0 = 2)$ with cointegrating rank r. Proceeding as above we obtain

$$
\begin{aligned}
IC(p = 2) - IC(p = 1) &= K^2 \frac{c_T}{T} + \sum_{i=1}^{r} \log(1 - \alpha_i \rho_i^2) + \\
&\quad \sum_{i=r+1}^{K} \log(1 - \rho_i^2) + o_p(1)
\end{aligned} \tag{10}
$$

thus illustrating the fact that in relation to the effects of K on the probability of underfitting in finite samples, the cointegrated case will correspond to an intermediate scenario between the purely $I(1)$ and purely $I(0)$ cases.

2.4 Local Poperties

In this section we investigate the behaviour of the various model selection criteria when the entries of the coefficient matrix Φ_{p_0} in Model (1) are allowed to shrink towards zero as the sample size increases. In other words while the true model has lag length p_0, the latter will be approaching $p_0 - 1$ as $T \to \infty$. Intuitively, the smaller the entries of Φ_{p_0} the more difficult it will be to distinguish between p_0 and $p_0 - 1$ thus raising the risk of underfitting. Although all model selection criteria whose penalty terms satisfy $c_T/T \to 0$ do not underfit asymptotically, for moderate samples and large system dimensions in particular the probability of underestimating p_0 might be very high. This way of proceeding will also allow us to formally isolate the factors that influence the ability of the model selection criteria to correctly detect the true lag length and is very similar in spirit to the local power analysis conducted in the context of standard hypotheses tests. For simplicity, we focus on a $VAR(p_0 = 2)$ model with I(0) variables and restrict ourselves to a binary decision problem by imposing $p_{max} = 2$ and operating under $p_0 \geq 1$. We have

$$X_t = \Phi_1 X_{t-1} + \Phi_2 X_{t-2} + \epsilon_t \tag{11}$$

where $\Phi_2 = \Lambda/\sqrt{T}$ with Λ a $K \times K$ constant coefficient matrix. We also define $\Phi' = (\Phi_1, \Phi_2)$, $Z_t' = (X_{t-1}', X_{t-2}')$ and $\phi_{2K^2\times 1} = vec(\Phi)$. Letting $\lambda_{K^2\times 1} = vec(\Lambda)$ we introduce a $K^2 \times 2K^2$ restriction matrix G conformable with ϕ such that $G\phi = 0_{K^2\times 1}$ corresponds to a VAR(1) model (i.e. $\Phi_2 = 0$) and $G\phi = \lambda/\sqrt{T}$ corresponds to a VAR(2) local to a VAR(1). The following proposition summarizes the ability of the IC based model selection procedure to detect the true lag length $p_0 = 2$.

Proposition 2.2 *Under DGP (11) the probability of correct decision $P[IC(2) < IC(1)]$ is such that* $\lim_{T\to\infty} P[IC(2) < IC(1)] = P[\chi^2(K^2, \psi^2) > K^2\ c_T]$, *where $\chi^2(K^2, \psi^2)$ is a noncentral chi-square random variable with K^2 degrees of freedom and noncentrality parameter $\psi^2 = \lambda'[G(\Omega \otimes E(Z_t Z_t')^{-1})G']^{-1}\lambda$.*

The above result can be used to assess analytically the "power" properties of the model selection approach. The components of the noncentrality parameter ψ^2 include the elements that will affect the probability of correct decision of the model selection criteria. From Kendall and Stuart (1961) the noncentral χ^2 distribution can be approximated by a centered one as follows

$$\chi^2(K^2, \psi^2) \approx h\chi^2(m) \tag{12}$$

where $h = (K^2 + 2\psi^2)/(K^2 + \psi^2)$ and $m = (K^2 + \psi^2)^2/(K^2 + 2\psi^2)$. By focusing on simple DGP's that could allow the calculation of the noncentrality parameter to be done analytically, we can establish the main factors affecting the probability of correct decision as well as the degree of their importance. Consider the VAR(2) in (11) with $\Phi_1 = diag(\phi_1, \ldots, \phi_K)$, $\Phi_2 = diag(\frac{\lambda}{\sqrt{T}}, \ldots, \frac{\lambda}{\sqrt{T}})$ and $\Omega_\epsilon = I_K$ for instance. It is straightforward to establish that the noncentrality parameter is given by

$$\psi^2 = \frac{K\lambda^2}{1-\frac{\lambda^2}{T}}, \tag{13}$$

illustrating the fact that the correct decision frequencies depend on the system dimension, the sample size and the magnitude of the λ parameter. Note that the magnitude of the parameters appearing in Φ_1 does not affect the correct decision frequencies. More general models allowing for nonzero cross correlations and general covariances can also be handled using a symbolic algebra package such as Mathematica or Maple. In order to evaluate the accuracy of the analytical power we used the above DGP to compute both Monte-Carlo and analytical probabilities of correct decisions. Although the empirical and analytical powers did not coincide, they were rarely more than 10% apart, suggesting that the analytical asymptotic power is sufficiently accurate even in finite samples. Table 1 presents the analytical powers corresponding to (11) across a wide range of system dimensions and parameters for all three types of model selection criteria. For the chosen parameterization the magnitudes clearly illustrate the rapid deterioration of the BIC based results as the system dimension is allowed to grow beyond $K = 3$.

Table 1 about here

Overall the results suggest that the AIC criterion is the best performer, especially in large dimensional systems where criteria such as the BIC are totally unable to move away from the lowest possible lag length even under very large sample sizes. Recalling that we operate under $p_{max} = 2$, it is also important to observe that as K grows the incorrect decisions resulting from the AIC criterion are clustered at $p = 1$, the underfitted specification. This also supports our discussion centered around (9) where we argued that even for a small penalty magnitude such as $c_T = 2$, the size of the system dimension might prevent the inequality in (9) to take the desired sign in finite samples.

2.5 Further Empirical Evidence

Our previous results aimed to establish and explain the diversity of outcomes that could arise when evaluating the performance of alternative lag length selection methods. Given the large number of individual factors and their joint interactions influencing the overall properties of each criteria, our analysis allowed us to isolate features that would have required an impracticably large number of

DGP parameterizations for them to be uncovered via direct simulations. Here our aim is to use our previous analysis as a framework for designing a selective set of DGPs so as to provide further insight on the sensitivity of each model selection criterion to factors such as the sample size, system dimension, preset upper bound p_{max} and more importantly to highlight the fact that even slightly altered DGP parameterizations may lead to contradictory features for the same criterion.

We initially considered a ten dimensional system of I(1) variables. The true lag length was set to $p_0 = 1$ and we experimented across various values of the upper bound p_{max} and sample size T ($p_{max} = \{3, 5, 7\}$ and $T = \{90, 100, 150, 200, 250\}$). The correct decision frequencies corresponding to each criterion are displayed in Table 2a. Although we implemented the model selection approach for $0 \leq p \leq p_{max}$ we only present the frequencies corresponding to choices of $p \geq 1$ since none of the criteria pointed to $p = 0$ throughout all replications. Due to the I(1)'ness of our DGPs this latter point should not be interpreted as a strong ability of the model selection criteria not to underfit however. Indeed for the model selection criteria to point to $p = 0$ we need $P[IC(0) < IC(1)]$ which can also be rewritten as $P[\log|\hat{\Omega}(0)| - \log|\hat{\Omega}(1)| < K^2 c_T/T]$. Since $\hat{\Omega} = \sum_t X_t X_t'/T$ and given that X_t is an I(1) vector process it follows that $\hat{\Omega}(0) = O_p(T)$ which makes the probability $P[IC(0) < IC(1)]$ converge to zero extremely fast. From the results in Table 2a the consistency of the AIC based lag length estimate is striking. For values of $T \geq 150$, the AIC selected $p = 1$, 100% of the times, behaving exactly as the BIC and HQ. The frequencies corresponding to T=100 clearly highlight the importance of the selected upper bound p_{max}. Indeed, although for $p_{max} = 3$ the correct decision frequency corresponding to the AIC is approximately 99% under both T=90 and T=100, as we increase p_{max} to 5 and 7 we can observe that the AIC is systematically pointing to the upper bound, confirming our previous discussion and result in (5).

Table 2a about here

When we conducted the same experiments across smaller system dimensions while maintaining the restriction that p_{max} should be such that the lower bound in (5) is kept identical to the $K = 10$ scenario (e.g. $\{K, p_{max}\} = \{5, 14\}$ under $c_T = 2$ requires $T > 140$ as does $\{K, p_{max}\} = \{10, 7\}$) we found the correct decision frequency patterns to be both quantitatively and qualitatively similar to the ones presented in Table 2a. The magnitude and patterns of the above frequencies also remained unchanged when we introduced I(0) components into the system, confirming the irrelevance of the stationarity properties of the data for the probability of overfitting. The fact that the BIC points to $p_0 = 1$ $\forall p_{max}$ and $\forall T$ also casts some doubt on its genuine ability to move away from the lowest possible lag length. Indeed this might be due to the strength of its penalty which combined with the dimensionality factor makes it spuriously select $p = 1$.

In order to explore alternative scenarios under which underfitting is likely to occur in finite samples we next focus on a class of $VAR(p_0 = 2)$ models, concentrating on the individual and joint influence of factors isolated in our analysis in (8)-(10) and (13). We consider two types of $VAR(p_0 = 2)$ specifications, having *large* and *small* parameter magnitudes (ie. $\rho_i^{\prime}s$ in (7)) respectively. The chosen parameterization for the first DGP leads to $\sum_{i=1}^{K} \log(1-\rho_i^2) = -2.74$ while the second one leads to $\sum_{i=1}^{K} \log(1-\rho_i^2) = -0.42$. Our result in (8) suggests that for moderately small samples the BIC will point to p=1 most of the time even in the "strong parameter value" case. This is indeed confirmed by the empirical results presented in Table 2b which suggest that the BIC requires samples much greater than T=200 to achieve acceptable correct decision frequencies. The AIC based estimates on the other hand are converging to $p_0 = 2$ quite rapidly with the AIC selecting the true lag length close to 100% of the times for $T \geq 150$ and any magnitude of p_{max}.

Tables 2b-2c about here

It is also important to observe that the AIC does not overfit unless T is extremely small relative to p_{max} and K. When we reconsidered the same experiment with smaller magnitudes of the $\rho_i^{\prime}s$ the correct decision frequencies (see Table 2c) were reduced by half for the AIC which continued however to remain by far the best performing criterion since the BIC and HQ were totally unable to select any lag length other than the lower bound $p = 1$ close to 100% of the times. Another feature also worth emphasizing is that even for the AIC criterion all wrong decisions are clustered at $p < p_0$, the underfitted model, confirming our analysis in (8) and our results in Table 1. Thus our overall findings strongly suggest that in large dimensional systems and for moderately large sample sizes underfitting is the main problem practitioners should concentrate on even when using the AIC criterion. Regarding the relative performance of the criteria considered in this study, the AIC is clearly the best performer in large dimensional systems. In a related study Koreisha & Pukkila (1993) also investigated the influence of the system dimension on the behavior of standard information theoretic criteria via an extensive set of Monte-Carlo experiments based on purely stationary VAR(1) and VAR(2) models. In addition to providing further theoretical support and an analysis of the causes of some of their findings, our results suggest that the sample sizes considered in their study forced an overemphasis on the overfitting aspect. Indeed our findings suggest that underfitting might be a more serious and common problem in large dimensional systems.

So far our framework has assumed the order of the VAR to be finite and bounded by p_{max}. It is also important to evaluate the properties of the lag length estimation techniques when the error process of the VAR contains (invertible) moving average components, with the latter implying the the true DGP has a $VAR(\infty)$ representation. Within this framework it is still possible to approximate the $VAR(\infty)$ by a truncated $VAR(p)$ version and obtain consistent estimates of the parameter matrices provided that the truncation lag p is allowed to grow at an appropriate rate with the sample size (see Berk (1974), Lewis & Reinsel (1985), Ng & Perron (1995)). The key issue that arises in this context is the quality of the different methods for the selection of an appropriate truncation lag. In the context of a univariate autoregression Ng and Perron (1995) showed that under the presence of moving average errors with large MA parameter magnitudes, model selection criteria such as the BIC or AIC are unable to select large values of p unless an impracticably large sample size becomes available. Here we initially explore the same issue in a K dimensional VAR context by considering a simple stationary VAR(1) model driven by VMA(1) errors and written as $X_t = \epsilon_t - \Theta\epsilon_{t-1}$. Assuming $\Theta = diag(\theta_1, \ldots, \theta_K)$ for simplicity and putting $\Omega_\epsilon = I_K$ then we can write

$$IC(p+1) - IC(p) \approx \frac{c_T K^2}{T} + \sum_{i=1}^{K} \log\left[\frac{(1-\theta_i^{2(p+3)})(1-\theta_i^{2(p+1)})}{(1-\theta_i^{2(p+2)})^2}\right] \tag{14}$$

provided that T is sufficiently large so that $|\hat{\Omega}(p)| \approx \prod_{i=1}^{K}(1-\theta_i^{2(p+2)})(1-\theta_i^{2(p+1)})^{-1}$. Ideally (11) should continue to remain negative for sufficiently large values of p but the presence of the K^2 factor clearly highlights the fact that even an AIC type penalty may not allow the IC approach to select large lag lengths even when the $|\theta_i|'s$ are large, unless an extremely large sample size is available. As a numerical illustration, letting $K = 5$, $T = 1000$ and $\theta_i = -0.4\ \forall i = 1, \ldots, K$ we have $IC(2) - IC(1) < 0$ while $IC(3) - IC(2) > 0$.

In this $VAR(\infty)$ context it is obviously difficult to analyze the properties of alternative lag length selection techniques without having a benchmark to evaluate the costs of an inappropriate truncation (e.g. validity of the subsequent distribution theory of cointegration tests, accuracy of forecasts, validity of the resulting impulse response functions, granger-causality tests). Although the true lag length is infinite, the parameters of the AR representation are declining geometrically, thus if the parameters of the MA process are not too large in absolute value, a small truncation lag co uld possibly lead to approximately white noise residuals. Although it is beyond the scope of this paper to extend the univariate results presented in Hall (1994) and Ng and Perron (1995) to this VAR framework, here we adopt the view that an LM test for residual autocorrelation could be used to evaluate the quality of the selected truncation lag. For this purpose we simulated a ten-dimensional VAR(1) model driven by VMA(1) errors given by $\Delta X_t = \epsilon_t - \Theta\epsilon_{t-1}$, setting $\Theta = diag(0.8, 0.7, 0.6, 0.4, 0.2, 0.65, 0, 0, 0, 0)$ and letting $p_{max} = \{3, 5, 7\}$. Across all sample sizes, we found that both the BIC and HQ were unable to move away from $p = 1$. The AIC on the other hand pointed to $p = 2$ most of the time (approximately 83% of the times for $T \geq 250$ and any magnitude of p_{max}, with the remaining frquencies concentrated at $p = 1$). When we performed an LM test of residual autocorrelation across all the replications using the lag length chosen by the AIC and T=250 ,the test could not reject the white noise hypothesis approximately 85% of the times at a 5% level (compared with 3% for the BIC based estimated lag length) thus suggesting that $p = 2$ might be a reasonable truncation lag for our chosen DGP.

3 General to Specific LR based Testing

Instead of using an information theoretic approach for choosing an appropriate lag length, it is also possible to use a sequential testing strategy that focuses on the significance of the coefficient matrices in the VAR. A scheme commonly used in applied work involves testing $H_0^i : \Phi_{p_{max}-i+1} = 0$ versus $H_1^i : \Phi_{p_{max}-i+1} \neq 0 | \Phi_{p_{max}} = \ldots = \Phi_{p_{max}-i+2} = 0$ for $i = 1, \ldots,\ p_{max}$ using the $\chi^2(K^2)$ distributed $LR = T(\log|\hat{\Omega}(p_{max} - i)| - \log|\hat{\Omega}(p_{max} - i + 1)|)$ test statistic. The procedure stops when a null hypothesis is rejected for the first time, leading to $\hat{p} = p_{max} - i + 1$ (see Lütkepohl (1992), Ch.2). Although alternative testing schemes such as a specific to general approach have also been proposed in the literature, numerous studies that focused on univariate time series models documented the overall superiority of the GS approach (Hall (1994), Ng and Perron (1995)) relative to alternative testing schemes and accordingly in what follows we concentrate solely on the testing strategy outlined above. This general to specific approach has been criticized on the grounds that it does not lead to a consistent estimator of p_0 since the probability of overfitting does not vanish asymptotically. Also, the buildup of Type I errors could become considerable when the test involves long sequences, as it is the case when the chosen maximum lag length p_{max} is large.

The literature on model selection criteria has often argued that selecting the lag length via an information theoretic criterion is similar to performing a likelihood ratio based test with the critical values determined "internally" by the chosen penalty term rather than by the χ^2 distribution's specific cutoff points. This statement is not entirely correct however. In what follows we define $\hat{p}_{IC}^{(j)} = \arg\min_{j \leq p \leq p_{max}} IC(p)$ for $j = \{0, 1\}$ and let $\hat{p}_{GS}$ denote the corresponding lag order obtained via the GS testing approach. We also let c_α denote the cut-off point from the $\chi^2(K^2)$ distribution used in the GS testing approach (i.e. c_α is such that $P[\chi^2(K^2) > c_\alpha] = \alpha$). The following proposition summarizes our main result

Proposition 3.1 *Under assumptions (A1)-(A2) and if $c_T = c_\alpha/K^2$ $\forall T$ we have $\hat{p}_{IC}^{(0)} = \hat{p}_{GS}$ whenever $\hat{p}_{GS} \in [0, p_{max} - 1]$ if the polynomial in (A2) has all its roots outside the unit circle and $\hat{p}_{IC}^{(1)} = \hat{p}_{GS}$ whenever $\hat{p}_{GS} \in [1, p_{max} - 1]$ if the polynomial in (A2) has at least one root on the unit circle.*

The above proposition can allow us to make interesting parallels between the IC and GS testing approaches. This is particularly useful in this context since the overall significance level of the GS testing approach is difficult to determine. It is however important to emphasize the fact that $c_T = c_\alpha/K^2$ will not be able to force the IC approach to choose the same lag length as the GS testing approach when the latter leads to $p = p_{max}$. It is only when *two* nested models with $p_{max} = 2$ are being compared that one can obtain a unique penalty (c_α/K^2) which guarantees the same choice of p across the two methods $\forall p$. An important implication of the above proposition is that as the system dimension increases, the use of the GS testing approach will lead to greater and greater lag lengths since as K increases it becomes less and less costly to overfit. It is also clear that the lag length selected by the GS approach will always be greater than the one obtained using the usual model selection criteria, since the probability of overfitting is a decreasing function of the penalty term.

We next evaluated the empirical performance of the GS approach by considering the same DGPs as in our earlier experiments. Within this finite order autoregressive and large dimensional framework our results unanimously confirmed the excessive tendency of the GS approach to point to lag orders close to p_{max} even under the most favourable parameter configurations and sample sizes. Under $T = 250$ for instance and considering the same DGP as in Table 2a, the GS testing strategy pointed to the true order $p_0 = 1$ close to 60%, 16% and 1% of the times under $p_{max} = 3$, $p_{max} = 5$ and 7 respectively, with most of the wrong frequencies clustered around $p = p_{max}$. More importantly across all previous experiments presented in Tables 2a-2c there was no single scenario under which the AIC criterion underperformed the GS approach. In order to also evaluate the behaviour of the GS approach under moving average errors and compare its decision frequency patterns with the model selection criteria we reconsidered the previously introduced VARMA(1,1) specification given by $\Delta X_t = \epsilon_t - \Theta\epsilon_{t-1}$. Recall that under this scenario the largest lag length selected by the model selection criteria was $p = 2$. Within this testing framework the GS approach on the other hand led to lag lengths concentrated around p_{max} most of the times, pointing to $p_{max} = 5$ approximately 55% of the times under T=250 and to $p_{max} = 7$ approximately 85% of the times. Interestingly these latter frequencies are also similar to the ones obtained under a finite VAR(2) DGP, thus raising doubts about the ability of the GS approach to select lag lengths other than or close to p_{max}.

4 Conclusions & Implications for Applied Research

In applied work, the frequent interest in dynamic interrelationships among economic variables across different countries, sectors or regions makes large dimensional VAR's a common framework of analysis. Although the specification of their dynamic structure is not of direct interest, its accuracy is crucial for subsequent inferences. In this paper we have shown that the commonly used model selection criteria for choosing an optimal lag length can be extremely sensitive to factors such as the system dimension and the preset upper bound. Contrary to the common belief that the AIC criterion has a tendency to overfit we found that in large dimensional systems the opposite is more likely to happen under moderate sample sizes. Furthermore, AIC's well known non zero asymptotic probability of overfitting is negligible in medium sized systems and zero in larger ones. We also derived a lower bound for the sample size under which the AIC will repeatedly point to the preset upper bound thus explaining various anomalies in the literature. From a practical point of view our results strongly point in favor of an AIC based approach for selecting lag lengths in large dimensional systems.

APPENDIX

Proof of Proposition 2.1 From (2) the requirement that $IC(p_0+h) < IC(p_0)$ can be formulated as $\log|\hat{\Omega}(p_0)| - \log|\hat{\Omega}(p_0+h)| > K^2h\frac{c_T}{T}$ implying that $\lim_{T\to\infty} P[IC(p_0+h) < IC(p_0)] = \lim_{T\to\infty} P(T(\log|\hat{\Omega}(p_0)| - \log|\hat{\Omega}(p_0+h)|) > K^2hc_T]$ where $T(\log|\hat{\Omega}(p_0)| - \log|\hat{\Omega}(p_0+h)|)$ is the likelihood ratio statistic for testing the null hypothesis $H_0 : G\phi = 0$ in (1) with $\phi_{K^2\times(p_0+h)} = vec(\Phi)$, $\Phi' = (\Phi_1, \ldots, \Phi_{p_0}, \Phi_{p_0+1}, \ldots, \Phi_{p_0+h})$ and G a known $(K^2h \times K^2(p_0+h))$ restriction matrix of rank K^2h. Under assumption (A1) and assuming also that all the roots of the polynomial in (A2) lie outside the unit circle we have (see Lütkepohl (1991, Ch. 3)) $\sqrt{T}(\hat{\phi} - \phi) \overset{d}{\to} N(0, \Omega \otimes Q^{-1})$ where $\hat{\phi}$ denote the MLE of ϕ, $z_t' = (x_{t-1}', \ldots, x_{t-(p_0+h)}')$ and $Q = E(z_t z_t')$. Thus under $H_0 : G\phi = 0$ we can write $\sqrt{T}G\hat{\phi} \overset{d}{\to} N(0, G(\Omega \otimes Q^{-1})G')$ also implying that the null limiting distribution of the Wald statistic, asymptotically equivalent to the LR is given by $T\hat{\phi}G'[G(\Omega \otimes Q^{-1})G']^{-1}G\hat{\phi} \overset{d}{\to} \chi^2(K^2h)$. Since the law of large numbers ensures that $plim\ \sum_t z_t z_t'/T = Q$ and $plim\ \hat{\Omega} = \Omega$ as $T \to \infty$, the quantity given by $\hat{\phi}G'[G(\hat{\Omega} \otimes (\sum z_t z_t')^{-1})G']^{-1}G\hat{\phi}$ will also be distributed as $\chi^2(K^2h)$ as required. Next, when X_t has I(1) components the original VAR can be reparameterized in such a way that the restrictions implied by the above null hypothesis can be reformulated as restrictions imposed on the parameter matrices corresponding to stationary regressors only. Indeed assuming $p_0 \geq 1$ and letting $(I_K - \Phi_1 L - \ldots - \Phi_{p_0+h}L^{p_0+h}) = (I_K - \Pi L) - (\Gamma_1 L + \ldots + \Gamma_{p_0+h-1}L^{p_0+h-1})(1-L)$ with $\Pi \equiv \Phi_1 + \ldots + \Phi_{p_0+h}$ and $\Gamma_s \equiv -(\Phi_{s+1} + \ldots + \Phi_{p_0+h})$ for $s = 1, 2, \ldots, p_0+h-1$ the $VAR(p_0+h)$ can now be reparameterized as $X_t = \Pi X_{t-1} + \Gamma_1 \Delta X_{t-1} + \ldots + \Gamma_{p_0+h-1}\Delta X_{t-(p_0+h-1)} + \epsilon_t$. Then the null hypothesis in the original model is equivalent to $H_0' : \Gamma_{p_0} = \ldots = \Gamma_{p_0+h-1} = 0$ in the reparameterized version. Since the restrictions implied by H_0' involve coefficients on stationary regressors only, the likelihood ratio statistic will have the same asymptotic distribution as in the I(0) case.

Proof of Proposition 2.2 Using the same notation as in the proof of Proposition 2.1, and noting that the restriction $G\phi = \frac{1}{\sqrt{T}}\lambda$ implies $\sqrt{T}G\phi = \lambda$ we have $\sqrt{T}(G\hat{\phi} - G\phi) \to N(\lambda, G(\Omega \otimes Q^{-1})G')$ instead of the central multivariate normal limiting distribution that we had under $G\phi = 0$. Since the quadratic form of a non central normal random vector with identity covariance is non central χ^2, the result follows.

Proof of Proposition 2.3 Letting $\omega_i = T\log|\hat{\Omega}(i)|$ where $\hat{\Omega}_i$ denotes the residual covariance matrix from a fitted VAR(i) specification and using (2)-(3) with $1 \leq p \leq p_{max}$ we have that the model selection approach will point to lag length p $\forall p \in [1, p_{max}]$ when

$$\begin{aligned} \omega_i - \omega_p &> (p-i)K^2c_T \ \ \forall\ i = 1, \ldots, p-1 \\ \omega_p - \omega_{i+1} &< (i+1-p)K^2c_T \ \ \forall\ i = p, \ldots, p_{max}-1. \end{aligned}$$

Similarly, for p obtained via the likelihood ratio based GS approach with $1 \leq p \leq p_{max}$, the estimated p is such that

$$\begin{aligned} \omega_{p-1} - \omega_p &> c_\alpha \\ \omega_i - \omega_{i+1} &< c_\alpha \ \ \forall i = p, \ldots, p_{max}-1. \end{aligned}$$

The result then follows by observing that when $c_T = c_\alpha/K^2$, the conditions that lead to the choice of $p = 1, 2, \ldots, (p_{max}-1)$ under the GS approach are identical to the ones that make the IC approach point to the same value of p. It is only when the GS strategy leads to $p = p_{max}$ that the two approaches might lead to distinct lag choices, since the above conditions do not overlap. When the roots of the polynomial in (A2) are known to lie strictly outside the unit circle the result follows by proceeding in an identical manner as above, with $0 \leq p \leq p_{max}$.

TABLE 1 Analytical Correct Decision Frequencies ($p_0 = 2$)

	$\lambda = 3, T = 150$			$\lambda = 3, T = 400$		
K	AIC	BIC	HQ	AIC	BIC	HQ
2	98%	59%	89%	98%	38%	81%
3	98%	24%	76%	97%	8%	62%
4	97%	4%	55%	96%	0%	34%
5	94%	0%	29%	93%	0%	12%
6	90%	0%	10%	88%	0%	2%
7	83%	0%	2%	79%	0%	0%
8	72%	0%	0%	67%	0%	0%
9	57%	0%	0%	52%	0%	0%
10	41%	0%	0%	36%	0%	0%

TABLE 2a $\Delta X_{it} = \epsilon_{it}$ for $i = 1, \ldots, K$ $(K = 10,\ p_0 = 1))$

$p_{max} = 3$	T=90			T=100			T=150			T=200			T=250		
p	AIC	BIC	HQ	AIC	BIC	HQ	AIC	BIC	HQ	AIC	BIC	HQ	AIC	BIC	HQ
1	98.8	100	100	99.4	100	100	100	100	100	100	100	100	100	100	100
2	0.5	0	0	0.3	0	0	0	0	0	0	0	0	0	0	0
3	0.7	0	0	0.3	0	0	0	0	0	0	0	0	0	0	0

$p_{max} = 5$	T=90			T=100			T=150			T=200			T=250		
p	AIC	BIC	HQ	AIC	BIC	HQ	AIC	BIC	HQ	AIC	BIC	HQ	AIC	BIC	HQ
1	40.1	100	100	91.8	100	100	100	100	100	100	100	100	100	100	100
2	0.1	0	0	0.2	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	59.7	0	0	8	0	0	0	0	0	0	0	0	0	0	0

$p_{max} = 7$	T=90			T=100			T=150			T=200			T=250		
p	AIC	BIC	HQ	AIC	BIC	HQ	AIC	BIC	HQ	AIC	BIC	HQ	AIC	BIC	HQ
1	0	99.9	0.4	0	100	99.2	100	100	100	100	100	100	100	100	100
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	100	0.1	99.6	100	0	0.8	0	0	0	0	0	0	0	0	0

TABLE 2b $\Delta x_{it} = u_{it},\ u_{it} = \rho_i u_{it-1} + \epsilon_{it}\ \mathit{for}\ i = 1, \ldots, K$

$\rho_1 = 0.3, \rho_2 = 0.7, \rho_3 = 0.5, \rho_4 = 0.6, \rho_5 = 0.8, \rho_6 = 0.2, \rho_7 = 0.5, \rho_8 = 0.4, \rho_9 = 0.0, \rho_{10} = 0.0$

$K = 10,\ p_0 = 2,\ \sum_{i=1}^{10} \log(1 - \rho_i^2) = -2.74$

$p_{max} = 3$	T=90			T=100			T=150			T=200			T=250		
p	AIC	BIC	HQ	AIC	BIC	HQ	AIC	BIC	HQ	AIC	BIC	HQ	AIC	BIC	HQ
1	0.3	100	69.8	0.3	100	54.5	0	95.5	1.5	0	32.4	0	0	0.5	0
2	75.8	0	30.2	89.6	0	45.5	100	4.5	98.5	100	67.6	100	100	99.5	100
3	23.9	0	0	10.1	0	0	0	0	0	0	0	0	0	0	0
$p_{max} = 5$	T=90			T=100			T=150			T=200			T=250		
p	AIC	BIC	HQ	AIC	BIC	HQ	AIC	BIC	HQ	AIC	BIC	HQ	AIC	BIC	HQ
1	0	100	69.8	0.1	100	54.5	0	95.5	1.5	0	32.4	0	0	0.5	0
2	0.8	0	30.2	25.6	0	45.5	99.8	4.5	98.5	100	67.6	100	100	99.5	100
3	0.1	0	0	1.2	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0.8	0	0	0.1	0	0	0	0	0	0	0	0
5	99.1	0	0	72.3	0	0	0.1	0	0	0	0	0	0	0	0
$p_{max} = 7$	T=90			T=100			T=150			T=200			T=250		
p	AIC	BIC	HQ	AIC	BIC	HQ	AIC	BIC	HQ	AIC	BIC	HQ	AIC	BIC	HQ
1	0	93.5	0	0	100	31.1	0	95.5	1.5	0	32.4	0	0	0.5	0
2	0	0	0	0	0	19.1	98.4	4.5	98.5	100	67.6	100	100	99.5	100
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0.1	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0.1	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0.1	0	0	0	0	0	0	0	0
7	100	6.5	100	100	0	49.8	1.3	0	0	0	0	0	0	0	0

TABLE 2c $\Delta x_{it} = u_{it},\, u_{it} = \rho_i u_{it-1} + \epsilon_{it}\;\; for\;\; i = 1, \ldots, 10$
$\rho_1 = 0.2, \rho_2 = 0.3, \rho_3 = 0.1, \rho_4 = 0.4, \rho_5 = 0.15, \rho_6 = 0.25, \rho_7 = 0.1, \rho_8 = 0, \rho_9 = 0, \rho_{10} = 0$
$K = 10,\, p_0 = 2,\, \sum_{i=1}^{10} \log(1 - \rho_i^2) = -0.42$

$p_{max} = 3$	T=90			T=100			T=150			T=200			T=250		
p	AIC	BIC	HQ	AIC	BIC	HQ	AIC	BIC	HQ	AIC	BIC	HQ	AIC	BIC	HQ
1	88.8	100	100	92	100	100	91.6	100	100	77.1	100	100	50.6	100	0
2	7.1	0	0	7.1	0	0	8.4	0	0	22.9	0	0	49.4	0	0
3	4.1	0	0	0.9	0	0	0	0	0	0	0	0	0	0	0
$p_{max} = 5$	T=90			T=100			T=150			T=200			T=250		
p	AIC	BIC	HQ	AIC	BIC	HQ	AIC	BIC	HQ	AIC	BIC	HQ	AIC	BIC	HQ
1	19.1	100	100	70.6	100	100	91.6	100	100	77.1	100	100	50.6	100	100
2	0.3	0	0	3.1	0	0	8.4	0	0	22.9	0	0	49.4	0	0
3	0.1	0	0	0.1	0	0	0	0	0	0	0	0	0	0	0
4	0.1	0	0	0.2	0	0	0	0	0	0	0	0	0	0	0
5	80.4	0	0	26	0	0	0	0	0	0	0	0	0	0	0
$p_{max} = 7$	T=90			T=100			T=150			T=200			T=250		
p	AIC	BIC	HQ	AIC	BIC	HQ	AIC	BIC	HQ	AIC	BIC	HQ	AIC	BIC	HQ
1	0	99.4	0	0	100	97.4	91.5	100	100	77.1	100	100	50.6	100	100
2	0	0	0	0	0	0	8.4	0	0	22.9	0	0	49.4	0	0
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	100	0.6	100	100	0	2.6	0.1	0	0	0	0	0	0	0	0

REFERENCES

Akaike, H. (1974). "A New Look at the Statistical Model Identification, " *IEEE Transactions on Automatic Control*, AC-19, 667-673.

Bartlett, M.S. (1954). "A Note on the Multiplying Factors for various χ^2 approximations," *Journal of the Royal Statistical Society Ser. B*, 16, 296-298.

Berk, K. (1974). "Consistent Autoregressive Spectral Estimates," *Annals of Statistics*, 2, 489-502.

Cheung, Y.W. and Lai, K.S. (1993). "Finite Sample Sizes of Johansen's Likelihood Ratio Test for Cointegration," *Oxford Bulletin of Economics and Statistics*, 55, 313-328.

Engle, R. and C. W. J. Granger (1987). "Co-Integration and error correction: representation, estimation and testing," *Econometrica,* Vol. 55, pp. 251-276.

Fujikoshi, Y. (1977). "Asymptotic Expansions for the Distributions of some Multivariate Tests," *Multivariate Analysis-IV, 55-71, P.R. Krishnaiah, ed.*

Gonzalo, J. and Pitarakis, J.Y. (1995). "Comovements in Large Systems," *CORE Discussion Paper, No. 9465.*

Gonzalo, J. and Pitarakis, J.Y. (1998). "Specification via model selection in vector error correction models," *Economics Letters,* Vol. 60, pp. 321-328.

Gonzalo, J. and Pitarakis, J.Y. (1999). "Dimensionality Effects in Large Dimensional Systems," in R. Engle and H. White eds. *Cointegration, Causality and Forecasting: Volume in Honour of C.W.J Granger,* Oxford University Press.

Hannan, E. and Quinn, B. (1979). "The Determination of the Order of an Autoregression," *Journal of the Royal Statistical Society, Ser. B*, 41, 190-195.

Ho, M.S. and Sorensen, B. (1996). "Finding Cointegration Rank in High Dimensional Systems using the Johansen Test. An Illustration using Data based Monte-Carlo Simulations," *Review of Economics and Statistics*, 4, 726-732.

Kendall, M. and Stuart, A. (1961). *The Advanced Theory of Statistics. Vol.2,* (New-York: Charles Griffin)

Koreisha, S.G. and Pukkila, T. (1993). "Determining the order of a Vector Autoregression when the Number of Components is Large," *Journal of Time Series Analysis*, 14, 47-69.

Lewis, R. and Reinsel, G.C. (1985). "Prediction of Multivariate Time Se ri es by Autoregressive Model Fitting," *Journal of Multivariate Analysis*, 16, 393-411.

Lütkepohl, H. (1985). "Comparison of Criteria for Estimating the Order of a Vector Autoregressive Process," *Journal of Time Series Analysis*, 6, 35-52.

Ng, S. and Perron, P. (1995). "Unit Root Tests in ARMA Models with Data Dependent Methods for the Selection of the Truncation Lag," *Journal of the American Statistical Association*, 90, 268-281.

Paulsen, J. (1984). "Order Determination of Multivariate Autoregressive Time Series with Unit Roots," *Journal of Time Series Analysis*, 5, 115-127.

Paulsen, J. and D. Tjostheim (1985). "On the Estimation of Residual Variance and Order in Autoregressive Time Series," *Journal of the Royal Statistical Society, Series B*, 47, 216-228.

Pötscher, B.M. (1990). "Estimation of Autoregressive Moving- Average Order given an Infinite Number of Models and Approximation of Spectral Densities," *Journal of Time Series Analysis*, 11, 165-179.

Reinsel, G.C. and Ahn, S.K. (1992). "Vector Autoregressive Models with Unit Roots and Reduced Rank Structure: Estimation, Likelihood Ratio Test, and Forecasting," *Journal of Time Series Analysis*, 13, 353-375.

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

Tsay, R.S. (1984). "Order Selection in Nonstationary Autoregressive Models," *Annals of Statistics*, 12, 1425-1433.