Specification via model selection in vector error correction models

Jesùs Gonzalo[a,*], Jean-Yves Pitarakis[b]

[a] *Department of Econometrics and Statistics, Universidad Carlos III de Madrid, 28903 Getafe, Madrid, Spain*
[b] *Department of Economics, University of Reading, Reading, UK*

Received 24 October 1997; accepted 15 May 1998

Abstract

This paper proposes a model selection approach for the specification of the cointegrating rank in the VECM representation of VAR models. Asymptotic properties of estimates are derived and their features compared with the traditional likelihood ratio based approach.

Keywords: VAR; Model selection; Misspecification

JEL classification: C32; C52

1. Introduction

The specification of an accurate short and long run dynamic structure is a crucial preliminary step in multivariate ARMA type time series models. Although the determination of the lag structure and cointegrating rank (k and r hereafter) is usually not the final objective of a modelling exercise it might have a great impact on subsequent inferences about the persistence of shocks, impulse responses, variance decomposition, forecasting etc. The most commonly used strategy by practitioners involves first obtaining an estimate of the lag length k using some information theoretic criterion or sequential testing strategy and then determining the cointegrating rank within a p-dimensional VAR(k) specification using the likelihood ratio based approach proposed by Johansen (1991). Although the limiting distribution of the LR statistic depends solely on the number of common trends $p-r$ and not on the lag length k, in a misspecified VAR in which errors are not iid (due to an overly parsimonious lag choice for instance), inferences that rely on the iid based tabulated distributions will be wrong even asymptotically. Another undesirable feature of the sequential testing approach is the fact that by construction it will not lead to a consistent estimate of r due to the constraint imposed by the size of the test. This aspect might become particularly burdensome in large dimensional systems where the build up of Type I errors can be considerable (see Gonzalo and Pitarakis (1995)). The primary objective of this paper is to evaluate the asymptotic and finite sample properties of a model selection based approach for the estimation of r and compare its behaviour with that of the traditional LR based

[*] Corresponding author. Fax: (34-1) 624-9849; e-mail: jgonzalo@elrond.uc3m.es

strategy. Given the extensive coverage of the lag length selection issue in the recent literature (see Lütkepohl, 1985, 1991; Gonzalo and Pitarakis, 1997, for an overview) our analysis will only focus on the implications of a misspecified lag length on the competing approaches. The plan of the paper is as follows. Section 2 will lay out the methodology and the asymptotic properties of the model selection based estimates. Section 3 will evaluate their properties from a practitioner's point of view and Section 4 concludes.

2. The model selection approach

2.1. Asymptotic properties

We assume that the data are driven by a p dimensional VAR(k) model in which the long run impact matrix has rank r where $r=0,\ldots,p$ and $k=1,\ldots,K$ with K assumed to be a known upperbound for k. We further let k_0 and r_0 denote the true lag length and cointegrating rank, respectively. More specifically we consider the following vector error correction representation for $X_t=(x_{1t},\ldots,x_{pt})'$

$$\Delta X_t = \Pi X_{t-1} + \sum_{j=1}^{k_0-1} \Gamma_j \Delta X_{t-j} + \epsilon_t \tag{1}$$

with ϵ_t assumed to be a Gaussian $iid(0,\Omega)$ process and $\Omega > 0$. Assuming that a particular estimate $\hat{k}$ of the unknown lag length has been obtained, we now view the estimation of r as a model selection problem where one chooses a model among a finite portfolio of $p+1$ competing models as follows

$$\hat{r} = \mathrm{Min}_{0\le \mathrm{r} \le \mathrm{p}}[IC(r|\hat{k})] \tag{2}$$

where

$$IC(r|\hat{k}) = \log|\hat{\Omega}(r|\hat{k})| + \frac{c_T}{T} m_r. \tag{3}$$

$\hat{\Omega}(r|\hat{k})$ in Eq. (3) denotes the residual covariance matrix estimated from Eq. (1) under the restriction that $rank(\Pi)=r$ and with $\hat{k}$ fitted lags. m_r represents the number of freely estimated parameters, with $m_r=2pr-r^2$ when $\hat{k}=1$ and more generally $m_r=p^2(\hat{k}-1)+2pr-r^2$ and c_T is a deterministic penalty term. When $c_T=2$, Eq. (3) reduces to the well known AIC criterion (Akaike, 1969, 1974), when $c_T=\log T$ we have the BIC criterion (Schwarz, 1978) and $c_T=2 \log\log T$ refers to the HQ criterion (Hannan and Quinn, 1979). The following two propositions establish the main results of the paper.

Proposition 2.1: *Letting r_0 denote the true rank of Π in Eq. (1) and $\hat{r}$ the estimated rank obtained from Eq. (2) with $\hat{k} \ge k_0$, then $\hat{r} \xrightarrow{p} r_0$ iff: (i) $\frac{c_T}{T} \to 0$; and (ii) $c_T \to \infty$.*

Proof: See Appendix A.

Proposition 2.2: *Letting r_0 denote the true rank of Π in Eq. (1) and $\hat{r}$ the estimated rank obtained from Eq. (2) with $\hat{k} < k_0$, then $\hat{r} \xrightarrow{p} r_0$ iff: (i) $\frac{c_T}{T} \to 0$; and (ii) $c_T \to \infty$.*

Proof: See Appendix A.

It is well known that the LR based sequential testing strategy will lead to wrong inferences if applied to a VAR(k) in which $k<k_0$ since the iid errors assumption gets violated. An important feature of the model selection approach as pointed out in Proposition 2.2 is that a correctly specified or overparameterized model is not a prerequisite for obtaining consistent estimates of the cointegrating rank, provided that the penalty term used in the criterion satisfies conditions (i) and (ii). It is true that when proceeding the conventional way it might always be possible to whiten the errors by adding extra lagged dependent variables on the right-hand side of Eq. (1) but given the enormous distortions one may face when degrees of freedom are scarce common sense often suggests a certain principle of parsimony, especially when the system dimension is large. The consistency result in Proposition 2.2 allows greater flexibility and reduced risk in the trade off between the selected order and the available degrees of freedom.

2.2. Computational aspects

The minimization problem in Eq. (2) involves the estimation of $\hat{\Omega}(r|\hat{k})$ across all possible values of r, rendering the approach less practical than a straightforward computation of the LR statistic. It is however possible to transform the expression of IC(r) so that it incorporates only the eigenvalues (canonical correlations between ΔX_t and X_{t-1}) used in the computation of the LR statistic, thus rendering the approach straightforward to implement. Using the same notation as in Johansen (1991), we let $S_{ij}=1/T\ \Sigma\hat{u}_{it}\hat{u}'_{jt}$ for $i,j=0{,}1$ where $\hat{u}_{0t}$ and $\hat{u}_{1t}$ are the residuals from the $\Delta X_t=A_1\Delta X_{t-1}+\ldots+A_{k-1}\Delta X_{t-k+1}+u_{0t}$ and $X_{t-1}=B_1\Delta X_{t-1}+\ldots+B_{k-1}\Delta X_{t-k+1}+u_{1t}$ regressions, respectively, when the model under investigation is a VAR(k). Next we let $\hat{\lambda}_i$ with $i=1\ldots,p$ denote the eigenvalues of $S_{11}^{-1}S_{10}S_{00}^{-1}S_{01}$ with $\hat{\lambda}_1\geq\ldots\geq\hat{\lambda}_p$ and where the LR statistic is given by $LR=-T\ \Sigma_{i=r+1}^{p}\log(1-\hat{\lambda}_i)$. Since the eigenvalues of $S_{11}^{-1}S_{10}S_{00}^{-1}S_{01}$ are the same as the ones of $S_{00}^{-1}S_{01}S_{11}^{-1}S_{10}$ and since $\hat{\Omega}=S_{00}-S_{01}S_{11}^{-1}S_{10}$ it follows that $|S_{00}|^{-1}|\hat{\Omega}|=|I_p-S_{00}^{-1}S_{01}S_{11}^{-1}S_{10}|$, leading to the following relationship

$$\log|\hat{\Omega}(r)| = \log|S_{00}| + \sum_{i=1}^{r}\log(1-\hat{\lambda}_i). \tag{4}$$

Thus we can rewrite IC(r) as

$$IC(r) = \log|S_{00}| + \sum_{i=1}^{r}\log(1-\hat{\lambda}_i) + \frac{c_T}{T}m_r. \tag{5}$$

To simplify the implementation of the approach even further we could focus on the minimization of $\overline{IC(r)}=IC(r)-IC(p)$ for $r=0,\ldots,p-1$ where $\overline{IC(r)}$ is given by

$$\overline{IC(r)} = -T\sum_{i=r+1}^{p}\log(1-\hat{\lambda}_i) - c_T(p-r)^2 \tag{6}$$

and $\overline{IC(p)}=0$. Thus, instead of having to run a set of reduced rank regressions, the only input required to implement the above approach is the set of eigenvalues entering the LR statistic and which are readily available from most packages.

3. Performance study

In this section we evaluate the finite sample performance of the model selection approach and compare its behaviour with that of the LR based strategy. We consider the implementation of the competing approaches on both properly specified and misspecified models[1] in order to also highlight the relative robustness of each technique when the fitted model is underparameterized. Within the model selection approach in addition to the well known AIC, BIC, HQ type penalties we also introduce an additional criterion (LCIC thereafter), the penalty of which is given by $(\log T + 2\log\log T)/2$ (i.e. a linear combination of the BIC and HQ). This alternative criterion clearly satisfies both requirements of Proposition 2.1 or 2.3 and will therefore also lead to consistent estimates of the cointegrating rank. The literature on model selection criteria is rich in ad hoc suggestions of alternative penalty terms that could potentially overcome the empirically established overly parsimonious estimates obtained by the BIC or the drawbacks of using weak or constant penalty terms. In Zhang (1992), for instance, the author argued that penalty terms should lie between 1.5 and 5 under most circumstances. Our motivation for introducing the LCIC criterion follows this same line of thought with the main concern of having a penalty term that could not only overcome excessive parsimony or overranking in finite samples but also continue to satisfy the consistency requirements.

We initially focused on a simple trivariate family of models given by $x_{1t} = \rho x_{1t-1} + \epsilon_{1t}$, $\Delta x_{it} = \epsilon_{it}$ $(i = 2,3)$, with $\epsilon_t \equiv \text{NID}(0, I_3)$ and $\rho \in [0.6, 1]$ with increments of magnitude 0.05, thus allowing our experiments to encompass both power and ‘size’ aspects. The sample size ranged from $T = 150$ to $T = 650$ with increments of 100 across $N = 2000$ replications. For this experiment, all inference strategies have been evaluated on a correctly specified model (i.e. by fitting $k = 1$ lags). There are three main points that can be drawn from the correct decision frequencies (empirical probabilities of selecting the true rank) displayed in Table 1. The AIC based correct decision frequencies display practically no variability across all sample sizes, pointing to the correct rank approximately 50–60% of the times across all values of ρ and T. Contrary to its popularity and overall good performance when used for lag length determination purposes, in this framework it is clearly unreliable with a strong tendency to overrank and no tendency to improve as T increases. Turning to the BIC criterion, our results suggest that it performed as well as the LR based approach for values of ρ up to 0.80 but subsequently failed to move away from $r_0 = 0$, clearly unable to detect the presence of a weak cointegrating relationship even for samples as large as $T = 650$. Despite the fact that the criterion leads to consistent estimates the consistency property is clearly not noticeable in finite and even moderately large sample sizes when the alternative is close to the null. It is only when we experimented with values of T greater than 1000 that we started observing a progression towards r_0. The HQ criterion on the other hand showed a relatively good performance, consistently outperforming the LR based approach even across small sample sizes. Finally the LCIC criterion also showed a behaviour very similar to that of the LR based strategy tracking its performance very closely. In summary, when fitting the correct model, model selection criteria such as the HQ and LCIC performed very similarly to the LR but none of the competing strategies stood out as a clear overall outperformer.

Next we focused on experiments where the model estimated with only one lag $(k = 1)$ is misspecified do to either VAR(1) or VMA(1) errors in the DGP. The latter case imply a true infinite

[1] For our purpose properly specified and misspecified models refer to models where $k = k_0$ and $k < k_0$, respectively.

Table 1
Correct decision frequencies (%)

DGP: $x_{1t}=\rho x_{1t-1}+\epsilon_{1t}$, $\Delta x_{it}=\epsilon_{it}$ $(i=2,3)$

ρ		$T=150$	$T=250$	$T=350$	$T=450$	$T=550$	$T=650$
0.60	AIC	64	63	64	62	65	64
0.60	BIC	97	99	100	100	100	100
0.60	HQ	90	92	94	94	94	95
0.60	LR	94	95	94	94	95	95
0.60	LCIC	96	98	98	99	99	100
0.65	AIC	63	65	63	65	64	64
0.65	BIC	92	99	100	100	100	100
0.65	HQ	92	92	93	93	95	95
0.65	LR	94	95	94	94	95	94
0.65	LCIC	96	98	98	98	99	99
0.70	AIC	64	64	63	65	63	63
0.70	BIC	76	99	100	100	100	100
0.70	HQ	90	92	94	94	94	94
0.70	LR	89	94	95	95	94	94
0.70	LCIC	92	98	99	98	99	99
0.75	AIC	64	64	66	64	63	65
0.75	BIC	49	95	100	100	100	100
0.75	HQ	89	92	95	93	95	95
0.75	LR	77	94	96	94	95	96
0.75	LCIC	80	97	99	99	99	99
0.80	AIC	64	64	64	64	64	64
0.80	BIC	23	73	98	100	100	100
0.80	HQ	78	92	94	95	96	96
0.80	LR	54	92	95	95	96	95
0.80	LCIC	52	93	98	99	99	99
0.85	AIC	62	64	64	63	63	62
0.85	BIC	7	31	68	94	100	100
0.85	HQ	55	87	93	94	96	97
0.85	LR	31	75	94	95	96	96
0.85	LCIC	25	64	93	98	99	99
0.90	AIC	55	63	66	66	66	64
0.90	BIC	2	5	14	36	61	83
0.90	HQ	27	54	81	92	95	95
0.90	LR	14	36	67	88	94	95
0.90	LCIC	10	21	44	76	92	98
0.95	AIC	41	50	58	64	63	64
0.95	BIC	1	1	1	1	2	4
0.95	HQ	12	17	23	38	51	69
0.95	LR	6	10	16	27	42	59
0.95	LCIC	4	5	5	9	14	25
1.00	AIC	47	47	50	51	50	49
1.00	BIC	100	100	100	100	100	100
1.00	HQ	90	93	94	96	95	96
1.00	LR	96	96	96	95	96	96
1.00	LCIC	98	99	99	100	100	100

order VAR in differences and the former a level VAR in which $k_0 = 2$. The motivation here is to evaluate the performance of the various strategies when a misspecified model is fitted to the data. Although the literature provides well established techniques for whitening potentially dependent errors, misspecification is a real risk when dealing with limited sample sizes. Table 2 presents correct decision frequencies based on the following DGP: $\Delta x_{it} = u_{it}$, $u_{it} = \rho_i u_{it-1} + \epsilon_{it}$ with $\rho_1 = 0.5$, $\rho_2 = 0.3$, $\rho_3 = 0.2$ and where $\epsilon_t \equiv \text{NID}(0, I_3)$. Thus, the true model is in fact a purely nonstationary VAR(2) with $r_0 = 0$. We considered the various rank selection strategies based on underfitted ($k = 1 < k_0 = 2$) and correctly specified ($k = k_0$) models. Results are summarized in Table 2. At this stage it is worth emphasizing the fact that within the underfitted case inferences based on the LR testing strategy will be wrong since its 'true' asymptotic distribution will depend on the parameters driving the error process and will not be the one tabulated in the literature under the iid errors assumption. This is somehow reflected by the correct decision frequencies which although are reasonable in magnitude, do not reflect the true ability of LR to detect the correct rank. Indeed it is interesting to observe the clustering around 80% even for samples as large as $T = 650$. The 'consistent' model selection criteria (BIC, HQ, LCIC) on the other hand show a clear and rapid tendency to converge towards the correct rank even when the estimated model is underparameterized. Although one might argue that since the BIC has a tendency to cluster at $r = 0$ the figures might not reflect its true ability to select the true rank, this criticism is not valid for the HQ and LCIC criteria which displayed an excellent performance in smaller sample sizes and also converged rapidly towards $r_0 = 0$ as T was allowed to increase. When we reconsidered the same experiments by fitting the correct lag length (i.e. setting $k = 2$ in the estimated models) the model selection criteria and the LR statistic showed a very similar behaviour. The LR by construction selected the correct magnitude approximately 95% of the times and the model selection criteria converged rapidly by selecting the true rank close to 100% of the times. In summary this set of experiments suggest that when models are misspecified model selection criteria such as the HQ and LCIC may be more reliable than the standard testing approach.

Finally, we evaluated the various strategies within a DGP with $r_0 = 0$ driven by moving average errors. More specifically we considered the following model: $\Delta x_{it} = \epsilon_{it} - \theta_i \epsilon_{it-1}$ $(i = 1,2,3)$ with

Table 2
Correct decision frequencies (%)

DGP: $\Delta x_{it} = u_{it}$, $u_{it} = \rho_i u_{it-1} + \epsilon_{it}$ $(i = 1,2,3)$, $\rho_1 = 0.5$, $\rho_2 = 0.3$, $\rho_3 = 0.1$

$k < k_0$	$T = 150$	$T = 250$	$T = 350$	$T = 450$	$T = 550$	$T = 650$
AIC	30	33	34	31	34	33
BIC	92	94	95	97	97	98
HQ	66	72	75	77	78	80
LR	79	79	80	79	79	79
LCIC	83	87	89	91	91	93
$k = k_0$	$T = 150$	$T = 250$	$T = 350$	$T = 450$	$T = 550$	$T = 650$
AIC	46	46	48	47	47	48
BIC	99	100	100	100	100	100
HQ	87	92	94	94	96	96
LR	94	96	95	96	97	96
LCIC	97	99	99	100	100	100

Table 3
Correct decision frequencies (%)

DGP: $\Delta x_{it}=\epsilon_{it}-\theta_i\epsilon_{it-1}$ $(i=1,2,3)$, $\theta_1=0.5$, $\theta_2=0.2$, $\theta_3=1$						
$k=1<k_0=\infty$	$T=150$	$T=250$	$T=350$	$T=450$	$T=550$	$T=650$
AIC	16	14	12	12	14	12
BIC	81	85	87	87	88	87
HQ	50	49	55	54	55	56
LR	63	59	60	58	58	57
LCIC	70	71	75	75	76	77

$\theta_1=0.5$, $\theta_2=0.2$, $\theta_3=0.1$ and $\epsilon_t \equiv \text{NID}(0,I_3)$. The correct decision frequencies based on models fitted with only one lag are presented in Table 3. The LR based frequencies are again clustered at around 60% across all sample sizes, with no improvement tendency as T increases. The testing strategy is again outperformed by the HQ and especially the LCIC based approaches in both moderate and larger sample sizes.

Overall our experiments suggest that when the estimated model is misspecified due to residual autocorrelation in the error process, the model selection criteria are more reliable than the conventional LR based approach, despite the fact that a substantial sample size might be required for the convergence to the truth to be visible. When the rank estimate is based on models fitted with the correct lag structure, both the model selection and LR based strategies display very comparable behaviour but the former did not seem to offer substantial improvements. This latter conclusion on correctly specified models also supports earlier simulation based evidence documented in Reimers (1993).

4. Concluding remarks

In this paper our objective was to evaluate both the theoretical and applied properties of a model selection based approach for the estimation of the cointegrating rank in multivariate time series models. We established that model selection based estimates have desirable asymptotic properties and are more robust to underparameterization than the ones obtained via the LR testing approach. In finite samples, although the performance of the IC based approach tracks very closely the LR based one when both procedures are applied to a correctly specified model, when the estimated models are underparameterized we found that the model selection procedure may provide significant improvements.

Acknowledgements

Financial support from the Spanish Secretary of Education (PB 950298 DGICYT) and the European Union Human Capital and Mobility Program is gratefully acknowledged.

Appendix A

Proof of Proposition 2.1: Letting $l>r_0$, from Eq. (5) we have $P[IC(l)<IC(r_0)]=P[-T\Sigma_{i=r_0+1}^{l} \log(1-\hat{\lambda}_i)>c_T(2pl-l^2-2pr_0+r_0^2)]$. Since $-T\Sigma_{i=r_0+1}^{l} \log(1-\hat{\lambda}_i)$ is $O_p(1)$, and the right-hand side diverges towards infinity from condition (ii), we have that $\lim_{T\to\infty} P[IC(l)<IC(r_0)]=0$ implying that overrranking does not occur asymptotically. For $l<r_0$ we have $P[IC(l)<IC(r_0)]=P[\Sigma_{i=l+1}^{r_0} \log(1-\hat{\lambda}_i)<\frac{c_T}{T}(2pr_0-r_0^2+l^2-2pl)]$ and since $p\lim(-\Sigma_{i=l+1}^{r_0} \log(1-\hat{\lambda}_i))>0$, from condition (i) the right-hand side will converge to zero, leading to $\lim_{T\to\infty}P[IC(l)<IC(r_0)]=0$, thus implying that underranking does not occur asymptotically. Taken together the above two results imply that $\hat{r}\xrightarrow{p}r_0$. In order to show that the requirements (i) and (ii) on the penalty term are necessary, let us suppose that c_T is bounded by some constant δ. Condition (i) still holds and $\lim_{T\to\infty}P[IC(l)<IC(r_0)]=0$ $\forall l<r_0$. For $l>r_0$ we have $P[IC(l)<IC(r_0)]=P[-T\Sigma_{i=r_0+1}^{l} \log(1-\hat{\lambda}_i)>c_T(2pl-l^2-2pr_0+r_0^2)]$ which will be non-zero since the right-hand side does not converge towards infinity when c_T is bounded. There is, therefore, a positive probability of overranking (i.e. selecting $l>r_0$). In order to see that (i) is necessary suppose that it fails, with $\frac{c_T}{T}\to c>0$. Clearly (ii) is satisfied and for $l>r_0$ we have $\lim_{T\to\infty}P[IC(l)<IC(r_0)]=0$. When $l<r_0$, $\lim_{T\to\infty}P[IC(l)<IC(r_0)]=P[-\Sigma_{i=l+1}^{r_0} \log(1-\hat{\lambda}_i)<c(2pr_0-r_0^2+l^2-2pl)]$ and since $c>0$ the result follows.

Proof of Proposition 2.2: When $\hat{k}<k_0$, the quantity $-T\Sigma_{i=r_0+1}^{l} \log(1-\hat{\lambda}_i)$ will not converge to the same asymptotic distribution as in the correctly specified or overfitted model but will still remain $O_p(1)$, thus ensuring that $\lim_{T\to\infty}P[IC(l)<IC(r_0)]=0$ when $l>r_0$ provided that $c_T\to\infty$. Similarly, $p\lim(-\Sigma_{i=l+1}^{r_0} \log(1-\hat{\lambda}_i))$ will remain positive despite the fact that it might not converge towards the same limit as when $\hat{k}\geq k_0$. This will, therefore, ensure that $\lim_{T\to\infty}P[IC(l)<IC(r_0)]=0$ when $l<r_0$ provided that $\frac{c_T}{T}\to 0$.

References

Akaike, H., 1969. Fitting autoregressive models for prediction. Annals of the Institute of Statistical Mathematics 21, 243–247.

Akaike, H., 1974. A new look at the statistical model identification, IEEE Transactions on Automatic Control AC-19, 667–673.

Gonzalo, J., Pitarakis, J.Y., 1997. Lag length estimation in large dimensional systems, Economics Department Discussion Paper, The University of Reading.

Gonzalo, J., Pitarakis, J.Y., 1995. Comovements in large systems, CORE Discussion Paper No. 9465.

Hannan, E., Quinn, B., 1979. The determination of the order of an autoregression. Journal of The Royal Statistical Society, Ser. B 41, 190–195.

Johansen, S., 1991. Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models. Econometrica 59, 1551–1580.

Lütkepohl, H., 1985. Comparison of criteria for estimating the order of a vector autoregressive process. Journal of Time Series Analysis 14, 47–69.

Lütkepohl, H., 1991. Introduction to Multiple Time Series Analysis, Springer-Verlag, Berlin.

Reimers, H.E., 1993. Comparisons of tests for multivariate cointegration. Statistical Papers 33, 335–359.

Schwarz, G., 1978. Estimating the dimension of a model. Annals of Statistics 6, 461–464.

Zhang, P., 1992. On the distributional properties of model selection criteria. Journal of the American Statistical Association 87, 732–737.