Model selection, estimation and forecasting in VAR models with short- and long-run restrictions: A Monte-Carlo study

Osmani Teixeira de Carvalho Guillén
Banco Central do Brasil
Av. Presidente Vargas, 730 - Centro
Rio de Janeiro, RJ 20071-001
Brazil

João Victor Issler
Graduate School of Economics – EPGE
Getulio Vargas Foundation
Praia de Botafogo 190 s. 1111
Rio de Janeiro, RJ 22253-900
Brazil
jissler@fgv.br

George Athanasopoulos
Department of Econometrics and Business Statistics
Monash University, Clayton, Victoria 3800
Australia

Farshid Vahid
School of Economics
The Australian National University
Canberra, ACT 0200
Australia

October 2007

Abstract

We study the joint determination of the lag length, the dimension of the cointegrating space and the rank of the matrix of short-run parameters of a vector autoregressive (VAR) model using model selection criteria. We consider model selection criteria which have data-dependent penalties for lack of parsimony, as well as the traditional ones. In order to compute the fit of each model, we propose an iterative procedure to compute the maximum likelihood estimates of parameters of a VAR model with short-run and long-run restrictions. Our Monte Carlo simulations measure the improvements in forecasting accuracy that can arise from the joint determination of lag-length and rank relative to the commonly used procedure of selecting lag-length only and then testing for cointegration.

Keywords: Reduced rank models, model selection criteria, forecasting accuracy.

JEL Classification: C32, C53.

1 Introduction

There is little evidence that allowing for cointegration improves the forecasting performance of a vector autoregressive (VAR) model. Engle and Yoo (1987) showed that unconstrained VAR models produced better short-term forecasts than cointegrated VARs. Even for long-term forecasts, their simulation showed that the theoretical advantage of cointegrated VARs for long-term forecasting was realised only when the cointegrating vectors were known. Clements and Hendry (1995), Lin and Tsay (1996), Hoffman and Rasche(1996), Christoffersen and Diebold (1998), Diebold and Kilian (2001), and Silverstovs, Engsted and Haldrup (2004) have since confirmed the conclusions of Engle and Yoo (1987). On the other hand, the evidence of the advantage of considering rank restrictions for short-term forecasting in stationary VAR (and VARMA) models is increasing (see, for example, Ahn and Reinsel, 1988, Vahid and Issler, 2002, and Athanasopoulos, 2007). In this paper we study this apparent dichotomy between the advantages of imposing rank restrictions on different parameter matrices of a model.

The first question we ask is to what extent can previous conclusions about the ineffectiveness of estimated long-run restrictions for forecasting be attributed to the modelling methodology, in particular to lag-order uncertainty. As Johansen (1991) and Gonzalo (1994) point out, VAR-order selection may affect proper inference on cointegrating rank and estimates of cointegrating vectors. Gonzalo and Pitarakis (1999) show that in large systems the usual model selection procedures for the lag order may severely underestimate the lag-order and consequently lead to wrong conclusions about the cointegrating rank. While we will investigate the accuracy of choosing the correct lag and rank here, our main focus will be on whether these inaccurate decisions are responsible for inferior forecasts of estimated VARs with long-run restrictions (from now on, we refer to such models as estimated vector error correction models or estimated VECMs for short).

In the context of stationary multivariate time series models, Vahid and Issler (2002) show that we can improve inference on both the lag-order and the rank of the parameter matrix if we determine them simultaneously with an appropriate model selection criterion such as the Hannan-Quinn criterion (HQ), or the Schwarz criterion (SC). They show that allowing for the possibility of reduced rank in the parameter matrix corrects the tendency of SC to underestimate the lag-order. They also show that this is likely to lead to significantly better short to medium term forecasts (1 to 8 step-ahead). Given that in the aforementioned papers the forecasting performance of VECMs have been measured only when lag-order is selected by a model selection criterion and then the cointegrating rank is determined by hypothesis testing, we ask the question whether simultaneous determination of lag-order and the ranks of the long-run and/or other parameter matrices can lead to a better forecasting performance of the resulting VECMs.

There is a well-founded and intuitive explanation for the difference between forecasting effects of allowing for cointegration and allowing for other types of rank restrictions in a VAR. It is inherently harder to extract information about low frequency dynamics from a finite time series sample than about the high frequency behaviour. Trends are more difficult to model than cycles. In a series of papers, Phillips and Ploberger have articulated this intuition (Phillips, 1996, Phillips and Ploberger, 1996, Ploberger and Phillips, 2003). In particular, they extend the results of Rissanen (1987) to time series with trending behaviour and show that the lower bound for how close an empirical model can get to the data generating process is larger when models include variables with trends, as in VECMs. They also suggest using this lower bound as a penalty for model selection in such situations. We explore this in the context of VECMs which may also have reduced rank on the parameters that governs their short-run dynamics.

Another objective of this paper is to investigate if rank restrictions on parameter matrices other than the long-run matrix improves the forecasting performance of VECMs. In the context of stationary time series, Vahid and Issler (2002) show that the cost of ignoring common cycle restrictions (i.e., restrictions that tie the short-run dynamics of series to a small number cycles less than the number of variables) is more than just the efficiency loss due to ignoring some restrictions. This is because the usual practice in applied work of choosing lag length by information criteria can lead to severely under-parameterised models in this case. Theoretically, little can be said about the forecasting performance of such severely misspecified models. Monte Carlo evidence shows that they perform poorly. Here we want to check this in the context of cointegrated VARs.

Our simulations cover three issues of model building, estimation, and forecasting. We examine the performance of standard information criteria ($IC(p)$) in choosing lag length p for cointegrated VARs with short-run restrictions. The consequences of this performance for the estimation of long-term parameters is also investigated. We also compare the performance of $IC(p)$ with that of $IC(p,r)$, which choose p and the rank of the short-run dynamic parameters system r simultaneously, and finally with that of $IC(p,r,q)$, which choose p, r and the cointegrating rank q simultaneously. We provide a comparison of forecasting accuracy of fitted VARs when only cointegration restrictions are imposed, when cointegration and short-run restrictions are jointly imposed, and when neither are imposed. These comparisons take into account the possibility of model misspecification in choosing the lag length of the VAR, the number of cointegrating vectors, and the rank of other parameter matrices. In order to estimate the parameters of a model with both long-run and short-run restrictions, we propose a simple iterative procedure similar to the one proposed by Hecq (2005). From these simulations, we also evaluate if there can be any improvements in the performance of tests of cointegrating rank and estimates of cointegrating vectors from better specification of short-run dynamics.

It is very difficult to claim that any result found in a Monte Carlo study is general, especially in multivariate time series. There are examples in the VAR literature of Monte Carlo designs as a result of which all model selection criteria over-estimate the true lag in small samples and therefore lead to the conclusion that the Schwarz criterion is the most accurate. The important feature of these designs is that they have a strong propagation mechanism. There are other designs with weak propagation mechanisms that result in all selection criteria underestimating the true lag and lead to the conclusion that AIC's asymptotic bias in overestimating the true lag may actually be useful in finite samples (see Vahid and Issler, 2002, for references). We pay particular attention to the design of the Monte Carlo, and present our results for different parts of the parameter space separately.

The outline of the paper is as follows. In Section 2 we study finite VARs with long-run and short-run restrictions and motivate their empirical relevance. We provide an overview of model selection criteria in Section 3, and in particular we discuss model selection criteria with data dependent penalty functions. In this section, we also outline an iterative procedure for computing the maximum likelihood estimates of parameters of a VECM with short-run restrictions. Section 4 describes our Monte-Carlo design; see also the discussion in the Appendix on DGP selection. Section 5 presents the simulation results and Section 6 concludes.

2 VAR models with long-run and short-run common factors

We start from the triangular representation of a cointegrated system used extensively in the theoretical cointegration literature (some early examples are Phillips and Hansen, 1990, Phillips and Loretan, 1991 and Saikkonen, 1992). We assume that the K-dimensional time series

$$y_t = \begin{pmatrix} y_{1t} \\ y_{2t} \end{pmatrix}, \quad t = 1, ..., T$$

where y_{1t} is $q \times 1$ (implying that y_{2t} is $(K-q) \times 1$) is generated from:

$$\begin{aligned} y_{1t} &= \beta y_{2t} + u_{1t} \\ \Delta y_{2t} &= u_{2t} \end{aligned} \tag{1}$$

where β is a $q \times (K-q)$ matrix of parameters, and

$$u_t = \begin{pmatrix} u_{1t} \\ u_{2t} \end{pmatrix}$$

is a strictly stationary process with mean zero and positive definite covariance matrix. This is a DGP of a system of K cointegrated I(1) variables with q cointegrating vectors, also referred to as a system of K I(1) variables with $K-q$ common stochastic trends (also some researchers refer to this as a system of K variables with $K-q$ unit roots, which can be ambiguous if used out of context and therefore we do not use it here)[1]. The extra feature that we add to this fairly general DGP is that u_t is generated from a VAR of finite order p and rank r ($< K$).

In empirical applications, the finite VAR(p) assumption is routine. This is in contrast to the theoretical literature on testing for cointegration, in which u_t is assumed to be an infinite VAR and a finite VAR(p) is used as an approximation (e.g. Saikkonen, 1992). Here, our emphasis is on building multivariate forecasting models rather than hypothesis testing. The finite VAR assumption is also routine when the objective is studying the maximum likelihood estimator of the cointegrating vectors, as in Johansen (1988).

The reduced rank assumption is considered for the following reasons. Firstly, this assumption means that all serial dependence in the K-dimensional vector time series u_t can be characterised by only $r < K$ serially dependent indices. This is a feature of most macroeconomic models, in which the short-run dynamics of the variables around their steady states are generated by a small number of serially correlated demand or supply shifters. Secondly, this assumption implies that there are $K-r$ linear combinations of u_t that are white noise. Gourieroux and Peaucelle (1990) call such time series "codependent" and interpret the white noise combinations as equilibrium combinations among stationary variables. This is justified on the grounds that although each variable has some persistence, the deviations from white noise combinations have no persistence at all. For instance, if an optimal control problem implies that the policy instrument should react to the current values of the target variables, then it is likely that there will be such a linear relationship between the observed variables up to a measurement noise. Finally, many papers in multivariate time series literature provide evidence of the usefulness of reduced rank VARs for forecasting (see, for example, Velu, Reinsel and Wickern, 1986,

[1] While in theory every linear system of K cointegrated I(1) variables with q cointegrating vectors can be represented in this way, in practice the decision on how to partition K-variables into y_{1t} and y_{2t} is not trivial, because y_{1t} are variables which must definitely have a non-zero coefficient in the cointegrating relationships.

and Ahn and Reinsel, 1998). Recently, Vahid and Issler (2002) have shown that failing to allow for the possibility of reduced rank structure can lead to developing seriously mis-specified vector autoregressive models that produce bad forecasts. Of course, all reasons stated here are also compelling reasons for abandoning the VAR assumption in favour of finite order VARMA models with simple embedded structures. This is our ultimate goal, but in this paper we limit ourselves to reduced rank finite VARs.

The dynamic equation for u_t is therefore given by (all intercepts are suppressed to ease the notation)

$$u_t = B_1 u_{t-1} + B_2 u_{t-2} + \cdots + B_p u_{t-p} + \varepsilon_t \tag{2}$$

where $B_1, B_2, ..., B_p$ are $K \times K$ matrices with $rank \left[\begin{array}{cccc} B_1 & B_2 & ... & B_p \end{array} \right] = r$, and ε_t is an i.i.d. sequence with mean zero and positive definite variance-covariance matrix and finite fourth moments. Note that the rank condition implies that each B_i has rank at most r and the intersection of the null-spaces of all B_i is a subspace of dimension $K - r$. The following lemma derives the vector error correction representation of this data generating process.

Lemma 1 *The data generating process given by equations (1) and (2) has a reduced rank vector error correction representation of the following type*

$$\Delta y_t = \gamma \left(\begin{array}{cc} I_q & -\beta \end{array} \right) y_{t-1} + \Gamma_1 \Delta y_{t-1} + \Gamma_2 \Delta y_{t-2} + \cdots + \Gamma_p \Delta y_{t-p} + \eta_t. \tag{3}$$

in which $rank \left[\begin{array}{cccc} \Gamma_1 & \Gamma_2 & ... & \Gamma_p \end{array} \right] \leq r$.

Proof. Subtracting y_{t-1} from both sides of the first equation in (1) and adding an subtracting βy_{2t-1} from the right side of the same equation leads to

$$\begin{array}{rcl} \Delta y_{1t} & = & -\left(y_{1t-1} - \beta y_{2t-1}\right) + \beta \Delta y_{2t} + u_{1t} \\ \Delta y_{2t} & = & u_{2t} \end{array}$$

which after substituting u_{2t} for Δy_{2t} on the first line, can be written as

$$\Delta y_t = - \left(\begin{array}{cc} I_q & -\beta \\ 0 & 0 \end{array} \right) y_{t-1} + v_t \tag{4}$$

where

$$v_t = \left(\begin{array}{cc} I_q & \beta \\ 0 & I_{K-q} \end{array} \right) u_t.$$

Since v_t is a full rank linear transformation of vector u_t, it will also be a VAR of order p and rank less than or equal to r, i.e.,

$$v_t = F_1 v_{t-1} + F_2 v_{t-2} + \cdots + F_p v_{t-p} + \eta_t$$

where $F_i = \left(\begin{array}{cc} I_q & \beta \\ 0 & I_{K-q} \end{array} \right) B_i \left(\begin{array}{cc} I_q & \beta \\ 0 & I_{K-q} \end{array} \right)^{-1}$ for $i = 1, ..., p$, and $\eta_t = \left(\begin{array}{cc} I_q & \beta \\ 0 & I_{K-q} \end{array} \right) \varepsilon_t$. The matrix $\left[\begin{array}{cccc} F_1 & F_2 & ... & F_p \end{array} \right]$ has rank less than or equal to r because it is the product of $\left[\begin{array}{cccc} B_1 & B_2 & ... & B_p \end{array} \right]$ and full rank matrices. Consider the characteristic polynomial of the vector autoregression that characterises the dynamics of v_t :

$$G(L) = I_K - F_1 L - F_2 L^2 - \cdots - F_p L^p.$$

Using the identity $G(L) = G(1) + G^*(L)(I_K - L)$, we can write this polynomial as:

$$\begin{aligned} G(L) &= I_K - F_1 - F_2 - \cdots - F_p + \\ &\left(\sum_{i=1}^{p} F_i + \sum_{i=2}^{p} F_i L + \cdots + \sum_{i=p-1}^{p} F_i L^{p-2} + F_p L^{p-1} \right) (I_K - L). \end{aligned}$$

Pre-multiplying both sides of (4) by $G(L)$, using the $G(1)+G^*(L)(I_K - L)$ formulation of $G(L)$ only when we apply it to the y_{t-1} term, and noting that $G(L)v_t = \eta_t$, we obtain,

$$\begin{aligned} G(L)\Delta y_t &= -(I_K - F_1 - F_2 - \cdots - F_p) \begin{pmatrix} I_q & -\beta \\ 0 & 0 \end{pmatrix} y_{t-1} \\ &\quad - \left(\sum_{i=1}^{p} F_i + \sum_{i=2}^{p} F_i L + \cdots + \sum_{i=p-1}^{p} F_i L^{p-2} + F_p L^{p-1} \right) \begin{pmatrix} I_q & -\beta \\ 0 & 0 \end{pmatrix} \Delta y_{t-1} + \eta_t. \end{aligned}$$

Expanding the left side of the equation and taking all lagged term to the right, and also denoting the first q columns of $-(I_K - F_1 - F_2 - \cdots - F_p)$ by γ, we obtain

$$\Delta y_t = \gamma \begin{pmatrix} I_q & -\beta \end{pmatrix} y_{t-1} + \sum_{j=1}^{p} \left(F_j - \sum_{i=j}^{p} F_i \begin{pmatrix} I_q & -\beta \\ 0 & 0 \end{pmatrix} \right) \Delta y_{t-j} + \eta_t.$$

Defining

$$\Gamma_j = F_j - \sum_{i=j}^{p} F_i \begin{pmatrix} I_q & -\beta \\ 0 & 0 \end{pmatrix}$$

for $j = 1, ..., p$, we note that each Γ_j is the result of elementary column operations on the matrix $\mathbf{F} = \begin{bmatrix} F_1 & F_2 & ... & F_p \end{bmatrix}$, they cannot have rank larger than the rank of $\mathbf{F}$. Moreover, all vectors in the null-space of $\mathbf{F}$ would also lie in the null-space of $\begin{bmatrix} \Gamma_1 & \Gamma_2 & ... & \Gamma_p \end{bmatrix}$. Therefore, $rank \begin{bmatrix} \Gamma_1 & \Gamma_2 & ... & \Gamma_p \end{bmatrix} \leq rank(\mathbf{F}) \leq r$. ■

This lemma shows that the triangular DGP (1) under the assumption that the dynamics of its stationary component (i.e. u_t) can be characterised by a small number of common factors, is equivalent to a VECM in which the coefficient matrices of lagged differences have reduced rank and their left null-spaces overlap. Hecq, Palm and Urbain (2006) call such a structure a VECM with weak serial correlation common features (WSCCF). It is instructive here to compare this structure with a DGP that embodies a stricter form of comovement, namely one that implies that the dynamics of the deviations of y_t from their Beveridge-Nelson (BN) trends can be characterised by a small number of cyclical terms.

Starting from the Wold representation for Δy_t

$$\Delta y_t = \Theta(L)\eta_t$$

where $\Theta(L)$ is an infinite moving average matrix polynomial with $\Theta_0 = I_K$ and absolutely summable coefficients and η_t are innovations in Δy_t, and using the matrix identity used in the proof of the lemma above, we get

$$y_t = \Theta(1) \sum_{i=0}^{\infty} \eta_{t-i} + \Theta^*(L)\eta_t,$$

where $\Theta_j^* = -\sum_{i=j+1}^{\infty} \Theta_i$. The first term is the vector of BN trends. These are random walks, and are simply the limit of the long-run forecast $y_{t+h|t}$ as $h \to \infty$. Cointegration

implies that $\Theta(1)$ has reduced rank, and hence the K random walk trends can be written in terms of a smaller number of common BN trends. Specifically, q cointegrating vectors is equivalent to $K-q$ common BN trends. Deviation from the BN trends, i.e. $\Theta^*(L)\eta_t$, are usually called the BN "cycles". The question is that if the reduced rank structure assumed for u_t in the triangular system above implies that the BN cycles can be characterised as linear combinations of r common factors. The answer is negative. Vahid and Engle (1993) analyse the restrictions that common trends and common cycles impose on a VECM. They show that in addition to a rank restriction similar to the one derived above on the coefficients of lagged differences, the left null-space of the coefficient of the lag level must also overlap with that of all other coefficient matrices. That is, the DGP with common BN cycles is a special case of the above under some additional restrictions.

One may question why we do not restrict attention to models with common BN cycles given that the reasons provided above to support the triangular structure, in particular the fact that most macro models imply that deviations from the steady state depend on a small number of common factors, more compellingly support a model with common BN cycles. However, Hecq, Palm and Urbain (2006) show that the uncertainty in determining the rank of the cointegrating space can adversely affect inference on common cycles, and they conclude that testing for weak common serial correlation features is a more accurate means of uncovering short-run restrictions in vector error correction models. Therefore, as a systematic approach to allow for more parsimonious models than the unrestricted VECM, it seems imprudent to consider only the strong form of serial correlation common features.

Our objective is to come up with a model development methodology that allows for cointegration and weak serial correlation common features. For stationary time series, Vahid and Issler (2002) show that allowing for reduced rank models is beneficial for forecasting. For partially non-stationary time series, there is an added dimension of cointegration. While the benefit of modelling cointegration per se for forecasting has been assessed to be negligible (Engle and Yoo, 1987, Clements and Hendry, 1995, Lin and Tsay, 1996), we examine the joint benefits of cointegration and short-run rank restrictions for forecasting partially non-stationary time series.

3 Model selection criteria and estimation of reduced-rank VAR models

The modal strategy in applied work for modelling a vector of I(1) variables is to use a model selection criterion for choosing the lag length of the VAR, then test for cointegration conditional on the lag-order, and finally estimate the VECM. Hardly ever any further step is taken to simplify the model, and if any test of the adequacy of the model is undertaken, it is usually a system test. For example, to test of the adequacy of the dynamic specification, additional lags of all variables are added to all equations, and an overall test of significance for testing K^2 restrictions is used. For stationary time series, Vahid and Issler (2002) show that model selection criteria severely underestimate the lag order in weak systems, i.e. in systems where the propagation mechanism is weak. They also show that using model selection criteria to choose lag order and rank simultaneously can remedy this shortcoming significantly. In modelling cointegrated I(1) variables, the underestimation of the lag order may have worse consequences because it also affects the quality of cointegration tests and estimates of cointegrating vectors.

Johansen (2002) analyses the finite sample performance of tests for the rank of the cointegrating space and suggests correction factors for improving the finite sample performance of such tests. The correction factor depends on the coefficients of lagged differences in the VECM (i.e. $\Gamma_1, \Gamma_2, \ldots, \Gamma_p$ in (3)), which makes the lag length p and estimates of $\Gamma_1, \Gamma_2, \ldots, \Gamma_p$ critical for the practical implementation of this correction factor. It is conceivable that if allowing for reduced rank VARs improves lag order selection and through that improves the quality of the estimates of $\Gamma_1, \Gamma_2, \ldots, \Gamma_p$, then the quality of finite sample inference on the rank of the cointegrating space will also improve. Hence, one issue that we investigate here is if the quality of inference on the rank of the cointegration space will improve when we choose p and rank of $\Gamma_1, \Gamma_2, \ldots, \Gamma_p$ using the model selection criteria suggested in Lütkepohl(1993, p. 202) and studied in Vahid and Issler (2002) relative to when we only choose the lag length using a traditional lag selection criterion. The lag-rank selection criteria are the analogues of the Akaike information criterion (AIC), the Hannan and Quinn criterion (HQ) and the Schwarz criterion (SC), and are defined as

$$AIC(p,r) = T \sum_{i=K-r+1}^{K} \ln\left(1-\lambda_i\left(p\right)\right) + 2(r(K-r)+rKp) \tag{5}$$

$$HQ\left(p,r\right) = T \sum_{i=K-r+1}^{K} \ln\left(1-\lambda_i\left(p\right)\right) + 2(r(K-r)+rKp)\ln\ln T \tag{6}$$

$$SC\left(p,r\right) = T \sum_{i=K-r+1}^{K} \ln\left(1-\lambda_i\left(p\right)\right) + (r(K-r)+rKp)\ln T, \tag{7}$$

where K is the dimension of (number of series in) the system, r is the rank of $\left[\begin{array}{cccc}\Gamma_1 & \Gamma_2 & \ldots & \Gamma_p\end{array}\right]$, p is the number of lagged differences in the VECM, T is the number of observations, and $\lambda_i\left(p\right)$ are the sample squared partial canonical correlations (PCCs) between Δy_t and the set of regressors $(\Delta y_{t-1}, ..., \Delta y_{t-p})$ after the linear influence of y_{t-1} (and deterministic terms such as a constant term and seasonal dummies if needed) is taken away from them, sorted from the smallest to the largest. Traditional model selection criteria are special cases of the above when rank is assumed to be full, i.e. when r is equal to K. Here, the question of the rank of Π, the coefficient of y_{t-1} in the VECM, is set aside, and taking the linear influence of y_{t-1} away from the dependent variable and the lagged dependent variables concentrates the likelihood on $\left[\begin{array}{cccc}\Gamma_1 & \Gamma_2 & \ldots & \Gamma_p\end{array}\right]$. Using well-known results on the relationship between the log-likelihood function and squared canonical correlations between the dependent and explanatory variables in reduced rank regressions (Anderson, 1951), the value of the above model selection criteria for $r = 1, \ldots, K$ can be easily calculated by a single canonical correlation estimation for each p. Choosing p and r that minimise one of these criteria, we can write the VECM as,

$$\begin{aligned}\Delta\, y_t &= \Pi\, y_{t-1} + CD_1\Delta\, y_{t-1} + CD_2\Delta\, y_{t-2} + \cdots + CD_p\Delta\, y_{t-p} + \eta_t \\ &= \Pi\, y_{t-1} + C\left[D_1\Delta\, y_{t-1} + D_2\Delta\, y_{t-2} + \cdots + D_p\Delta\, y_{t-p}\right] + \eta_t\end{aligned}$$

where C is a full rank $K \times r$ matrix and D_i are $r \times K$ matrices for $i = 1, ..., p$. Since $CD_i = CHH^{-1}D_i$ for any invertible $r \times r$ matrix H, these matrices are identified only after imposing r^2 normalising restrictions. The PCC procedure produces maximum likelihood estimates of $[D_1, D_2, \ldots, D_p]$ imposing the normalisation that the sample variance covariance matrix of $\left[\hat{D}_1\Delta\, y_{t-1} + \hat{D}_2\Delta\, y_{t-2} + \cdots + \hat{D}_p\Delta\, y_{t-p}\right]$ is identity. These are the

canonical variates corresponding to the largest r squared PCCs. The squared PCCs between Δy_t and y_{t-1} after taking the linear influence of $\left[\hat{D}_1 \Delta y_{t-1} + \hat{D}_2 \Delta y_{t-2} + \cdots + \hat{D}_p \Delta y_{t-p}\right]$ away, are then used as the eigenvalues on which the Johansen cointegration test is based on.

Strictly speaking, the above procedure is a slight departure from the likelihood ratio framework. This is because $\left[\hat{D}_1, \hat{D}_2, \ldots, \hat{D}_p\right]$ are maximum likelihood estimates of $[D_1, D_2, \ldots, D_p]$ only when Π is unrestricted, although they are consistent estimates even when Π has reduced rank. When Π is restricted to have rank $q\,(< K)$, the maximum likelihood estimates of $[D_1, D_2, \ldots, D_p]$ will change, and the maximised value of the log likelihood function cannot be simply written as a function of the PCCs between Δy_t and y_{t-1} conditional on $\left[\hat{D}_1 \Delta y_{t-1} + \hat{D}_2 \Delta y_{t-2} + \cdots + \hat{D}_p \Delta y_{t-p}\right]$. However, it is easy to compute the maximum of the log-likelihood function under the assumption of number of lag differences p, cointegration rank q, and rank of lagged difference coefficients r by iterating the following steps:

Step 0. Estimate $\left[\hat{D}_1, \hat{D}_2, \ldots, \hat{D}_p\right]$ as above, i.e. from a PCC procedure between Δy_t and $(\Delta y_{t-1}, ..., \Delta y_{t-p})$ controlling for y_{t-1}.

Step 1. Compute the PCCs between Δy_t and y_{t-1} conditional on $\left[\hat{D}_1 \Delta y_{t-1} + \hat{D}_2 \Delta y_{t-2} + \cdots + \hat{D}_p \Delta y_{t-p}\right]$. Take the q canonical variates $\hat{\alpha}' y_{t-1}$ corresponding to the q largest squared PCCs as estimates of cointegrating relationships. Regress Δy_t on $\hat{\alpha}' y_{t-1}$ and $\left[\hat{D}_1 \Delta y_{t-1} + \hat{D}_2 \Delta y_{t-2} + \cdots + \hat{D}_p \Delta y_{t-p}\right]$, and compute $\ln\left|\hat{\Omega}\right|$, the logarithm of the determinant of the residual variance matrix.

Step 2. Compute the PCCs between Δy_t and $(\Delta y_{t-1}, ..., \Delta y_{t-p})$ conditional on $\hat{\alpha}' y_{t-1}$. Take the r canonical variates $\left[\hat{D}_1 \Delta y_{t-1} + \hat{D}_2 \Delta y_{t-2} + \cdots + \hat{D}_p \Delta y_{t-p}\right]$ corresponding to the largest r PCCs as estimates of $[D_1 \Delta y_{t-1} + D_2 \Delta y_{t-2} + \cdots + D_p \Delta y_{t-p}]$. Regress Δy_t on $\hat{\alpha}' y_{t-1}$ and $\left[\hat{D}_1 \Delta y_{t-1} + \hat{D}_2 \Delta y_{t-2} + \cdots + \hat{D}_p \Delta y_{t-p}\right]$, and compute $\ln\left|\hat{\Omega}\right|$, the logarithm of the determinant of the residual variance matrix. If this is different from the corresponding value computed in Step 1, go to Step 1. Otherwise, stop.

The value of the $\ln\left|\hat{\Omega}\right|$ becomes smaller at each stage until it achieves its minimum, which we denote by $\ln\left|\hat{\Omega}_{p,q,r}\right|$. The values of $\hat{\alpha}$ and $\left[\hat{D}_1, \hat{D}_2, \ldots, \hat{D}_p\right]$ in the final stage will be the maximum likelihood estimators of α and $[D_1, D_2, \ldots, D_p]$. The maximum likelihood estimates of other parameters are simply the coefficient estimates of the final regression. The set of variables that are partialled out at each stage should include constants and other deterministic terms if needed.

We also consider determining three parameters p, r and q using model selection criteria. Gonzalo and Pitarakis (1999) study the performance of the AIC, the HQ, and the SC in choosing the rank of the cointegrating space. They only study a VAR(1) in levels and assume that the lag order is known. They find out that as the dimension of the system increases, the performance of both AIC and SC in selecting the correct cointegrating rank deteriorates, even when sample size is as large as 400. The bad performance of the AIC is not a surprise because AIC is not consistent and is expected to

choose a larger model than the correct model even asymptotically. When the dimension of the system is increased, AIC has more chance to fit the noise and choose a larger model. However, HQ and SC are consistent criteria, and the severe underestimation of rank by the SC when the dimension increases is of practical importance. Vahid and Issler (2002) find a similar result for SC when choosing the lag order in stationary VARs. They fix the dimension of the system, but they vary the strength of the propagation mechanism of the data generating process (DGP) measured by the trace of $\Omega_{\Delta y}\Omega_{\varepsilon}^{-1}$ where $\Omega_{\Delta y}$ is the variance of the generated series and Ω_{ε} is the variance of the error process that drives the DGP. In effect, increasing the dimension of the system in Gonzalo and Pitarakis (1999) is another way of weakening the propagation mechanism of the system. Vahid and Issler (2002) also observe that SC severely underestimates the lag length, but allowing for it to choose rank as well as the lag length significantly remedies this shortcoming. These results are particularly significant for building VARs for forecasting, where the popular belief is that SC should be used for choosing the lag length because more parsimonious models are likely to produce better forecasts. Vahid and Issler (2002) show that this belief is likely to lead to very poor forecasts when the true DGP's propagation mechanism is weak, as it is with most systems of macroeconomic variables.

We consider two classes of model selection criteria. First, we consider direct extensions of the AIC, HQ and SC to the case where the rank of cointegrating space, which is the same as the rank of Π, is also a parameter to be selected by the criteria. Specifically, we consider

$$AIC(p,r,q) = T\ln\left|\hat{\Omega}_{p,q,r}\right| + 2\left(q(K-q)+Kq+r(K-r)+rKp\right) \quad (8)$$

$$HQ\left(p,r,q\right) = T\ln\left|\hat{\Omega}_{p,q,r}\right| + 2\left(q(K-q)+Kq+r(K-r)+rKp\right)\ln\ln T \quad (9)$$

$$SC\left(p,r,q\right) = T\ln\left|\hat{\Omega}_{p,q,r}\right| + \left(q(K-q)+Kq+r(K-r)+rKp\right)\ln T, \quad (10)$$

where $\ln\left|\hat{\Omega}_{p,q,r}\right|$ (the minimised value of the logarithm of the determinant of the variance of the residuals of the VECM of order p, with Π having rank q and $\left[\begin{array}{cccc}\Gamma_1 & \Gamma_2 & ... & \Gamma_p\end{array}\right]$ having rank r) is computed by the iterative algorithm described above. Obviously, when $q = 0$ or $q = K$, we are back in the straightforward reduced rank regression framework where one set of eigenvalue calculations for each p provides the value of the log-likelihood function for $r = 1, ..., K$. Similarly, when $r = K$, we are back in the usual VECM estimation, and no iterations are needed.

We also consider a model selection criterion with a data dependent penalty function. Such model selection criteria date back at least to Poskitt (1987), Rissanen (1987) and Wallace and Freeman (1987). Poskitt (1987) sets up the usual Bayesian decision problem where the action depends on a parametric model estimated from a sample of T observations. Using the fact that for large T the log-likelihood is close to quadratic and assuming that utility of the best choice given a model is proportional to Van Emden (1971) measure of covariance complexity of the model (i.e. a measure of how precisely the model parameters can be estimated from the data), he shows that the posterior expected utility is maximised by choosing the model M that minimises

$$\delta_M = -2\ln l_M\left(\hat{\theta}\right) + d\left(1+\ln T+\ln\left(tr\left(\left[FIM_M\left(\hat{\theta}\right)/T\right]^{-1}\right)\right)/d\right) \quad (11)$$

where $l_M\left(\hat{\theta}\right)$ is the value of the likelihood function of model M evaluated at its maximum likelihood estimator $\hat{\theta}$, d is the number of the parameters of model M, $FIM_M\left(\hat{\theta}\right)$

is the Fisher information matrix of model M (i.e., $\left[-\partial^2 \ln l_M(\theta)/\partial\theta\partial\theta'\right]$) evaluated at $\hat{\theta}$, and tr is the trace operator. The last term computes the average of the variances of the maximum likelihood parameter estimates and favours models with more precise parameter estimates. The other penalty terms $d(1+\ln T)$ is of the same order of the penalty of the Schwarz criterion, and has been suggested in modified versions of the AIC. Using the same logic as Akaike (1973), Bozdogan (1987) derived $d(1+\ln T)$ instead of Akaike's $2d$ penalty, and suggested that as a correction for the AIC to make AIC consistent.

The model selection criterion that we consider in this paper is closer to those inspired by the work of Wallace and Freeman (1987). The idea is based on coding and information theory. A model is a medium for transmitting data from a sender to a receiver. The sender and receiver agree that the data is transmitted using a parametric class of models indexed by θ, and their common knowledge is embodied in the prior density $h(\theta)$. The length of the most efficient code for transmitting data using a model with estimated parameters $\hat{\theta}$ is $-\ln l_M\left(\hat{\theta}\right)$, but the estimated parameters themselves need to be transmitted also. The precision with which $\hat{\theta}$ is transmitted can also be optimised to make the message length as small as possible. Assuming that $\ln l_M\left(\hat{\theta}\right)$ is approximately quadratic, the most efficient code length for transmitting the parameters with optimal precision and then transmitting the data using the model will be:

$$MML = -\ln l_M\left(\hat{\theta}\right) - \ln h\left(\hat{\theta}\right) + \frac{1}{2}\ln\left|FIM_M\left(\hat{\theta}\right)\right| + \frac{d}{2}(1+\ln\kappa_d) \tag{12}$$

where κ_d is a constant, decreasing function of d. With stationary data, the $FIM_M\left(\hat{\theta}\right)$, which was defined above, is a $d\times d$ positive definite matrix whose elements are growing at the same order as T, and hence its eigenvalues will be growing at the same rate. In that case, its determinant, which is the product of its eigenvalues, grows at the same rate as T^d, and its logarithm therefore grows at the same rate as $d\ln T$, which is the same as the penalty term in the Schwarz criterion (please note that (12) has to be multiplied by 2 to be comparable with other criteria presented above). Hence, MML has a data dependent penalty function and behaves like the Schwarz criterion when the data are stationary and T is large. Dowe and Wallace (1999) have suggested more precise approximations to the theoretical message length, and Wallace and several coauthors have suggested suggest using (12) as an objective function for estimation (see Wallace, 2002, for a summary of the progress in this line of research).

Rissanen (1987) also uses coding and information theory and derives a bound for how close any empirical model of the data can get to the true data generating process P_θ. For stationary data, this bound is exactly the penalty of the Schwarz criterion. Recently, Ploberger and Phillips (2003) have extended Rissanen's work and shown its particular usefulness for model selection in time series analysis. The have generalised Rissanen's result to show that the distance between any empirical model and the P_θ is larger or equal to $\frac{1}{2}\ln\left|E_\theta\left(FIM_M\right)\right|$ almost everywhere on the parameter space[2]. In fact they show that this is true even when P_θ is the "pseudo-true" DGP, i.e. the closest DGP in a parametric class to the true DGP. This leads to Phillips (1996) and Phillips and Ploberger (1996) posterior information criterion (PIC):

$$PIC = -2\ln l_M\left(\hat{\theta}\right) + \ln\left|FIM_M\left(\hat{\theta}\right)\right|$$

[2] Ploberger and Phillips (2003) use the outerproduct formulation of the information matrix, which has the same expected value as the negative of the second derivative under P_θ.

which is basically the same as the MML criterion. The contribution of these recent group of papers has been to show the particular importance of this in application to partially nonstationary time series. With stationary series, all eigenvalues of FIM_M grow at the same rate as T. However, if in model M, one of the parameters is the coefficient of a variable with a deterministic linear trend, then the corresponding eigenvalues of FIM_M grows at the rate T^3, implying that the penalty for that parameter must be 3 times as much as a parameter for a stationary variable. Similarly, an I(1) variable would warrant a penalty twice as large as a stationary variable. This theory confirms that it is harder to get closer to the DGP when variables are trending.

Chao and Phillips (1999) used PIC for simultaneous selection of the lag length and cointegration rank in VARs. They reformulated PIC into a form that was convenient for their proof of consistency. They showed that in a K-variable vector error correction (VEC) model with p lagged differences and q cointegrating vectors, PIC penalty grows at the rate $\left(K^2p + 2q\left(K-q\right) + Kq\right)\ln T$ in contrast to SC penalty which is $\left(K^2p + q\left(K-q\right) + Kq\right)\ln T$. The question is that can one use $\left(K^2p + 2q\left(K-q\right) + Kq\right)\ln T$ as a penalty term to choose lag-length and the cointegration rank? The answer is negative, because $\left(K^2p + 2q\left(K-q\right) + Kq\right)$ is not a monotonically increasing function of q (its derivative is $3K - 4q$), which means that in some situations it would be possible to increase the number of cointegrating vectors (and inevitably improve the fit) and also decrease the penalty. The reason for this apparent inconsistency may be that the above rate is derived under a specific null hypothesis. Whereas in stationary case moving across models does not change the hypothesis of stationarity of the data, here different assumptions about the number of unit roots in the model changes the order of magnitude of variables according to that model. However, the observed data doesn't change, and therefore the data-dependent penalty $\ln\left|FIM_M\left(\hat{\theta}\right)\right|$ will adjust to the true order of magnitude of the data. The details of the Fisher information matrix for the reduced rank VECM is given in the appendix.

4 Monte-Carlo design for VARs with short- and long-term restrictions

One of the critical issues in any Monte-Carlo study is that of diversity of Data Generating Processes (DGPs), which allows sampling a large subset of the parameter space that includes sufficiently distinct members. One of the challenges in our context is that we want the design to include VECMs with short-term restrictions and satisfy conditions for stationarity. To make the Monte-Carlo simulation manageable, we use as DGP a three-dimensional VAR. The simple real business cycle models and also the simplest closed economy monetary dynamic stochastice general equilibrium models are three-dimensional (see for example King et al., 1991 and Rudebusch and Svennson, 1999). We consider VARs in levels with lag lengths of 2 and 3, which translates to 1 and 2 lagged differences in the VECM. This choice allows us to study the consequences of both under- or over-parameterization of the estimated VAR.

For each choice of cointegration rank q and short-run rank r, we use 100 DGPs. From each DGP, we generated 1,000 samples of 100 observations, and 1000 samples of 200 observations (the actual generated samples were longer and the initial part of each generated sample is discarded to reduce the effect of initial conditions). In summary, our results are based on 1,000 samples of 100 different DGPs – a total of 100,000 different samples – for each of either $T = 100$ or $T = 200$ observations.

As discussed in Vahid and Issler, it is worth sorting results by a measure of the

strength of the propagation mechanism of the DGP, i.e. a signal-to-noise ratio or a system R^2 measure. Here, we selected two different set of parameters with the following characteristics: the first has the median of the system R^2 measure between 0.4 and 0.5, with 3% larger than 0.6 and none greater than 0.7. The second has the median of the system R^2 between 0.7 and 0.8, with 22% larger than 0.8, none greater than 0.9, and none smaller than 0.7.

The Monte-Carlo procedure can be summarized as follows. Using each of our 100 DGPs, we generated 1,000 samples (once with 100, and again with 200 observations). Then, we recorded the lag length chosen by traditional (full-rank) information criteria, labelled $IC(p)$: $AIC(p)$, $HQ(p)$ and $SC(p)$, and the corresponding lag length chosen by alternative information criteria, labelled $IC(p,r)$: $AIC(p,r)$, $HQ(p,r)$ and $SC(p,r)$, in (5)-(7), with $q = 3$.

For choices made using $IC(p)$ we used Johansen's(1989, 1991) trace test at 5% to choose q and then estimated a VECM with no SCCF restrictions. Their out-of-sample forecasting accuracy measures were recorded up to 16 periods ahead. For choices made using $IC(p,r)$, we used the algorithm described in detail in the last section to obtain a triplet (p,r,q) in each case, with a resulting VECM estimated using SCCF restrictions. Their respective out-of-sample forecasting accuracy measures were recorded up to 16 periods ahead. Out-of-sample forecasting accuracy measures were then compared for these two types of VAR estimates.

4.1 Measuring forecast accuracy

We measure the accuracy of forecasts with the traditional trace of the mean-squared forecast error matrix (*TMSFE*) and the determinant of the mean-squared forecast error matrix (|*MSFE*|) at different horizons. The discussion in Lin and Tsay(1996) justifies the use of these two types of loss functions in measuring forecast accuracy. We also compute Clements and Hendry's(1993) *generalized forecast error second moment* (*GFESM).* *GFESM* is the determinant of the expected value of the outer product of the vector of stacked forecast errors of all future times up to the horizon of interest. For example, if forecasts up to h quarters ahead are of interest, this measure will be:

$$GFESM = \left| E \begin{pmatrix} \tilde{\varepsilon}_{t+1} \\ \tilde{\varepsilon}_{t+2} \\ \vdots \\ \tilde{\varepsilon}_{t+h} \end{pmatrix} \begin{pmatrix} \tilde{\varepsilon}_{t+1} \\ \tilde{\varepsilon}_{t+2} \\ \vdots \\ \tilde{\varepsilon}_{t+h} \end{pmatrix}' \right|$$

where $\tilde{\varepsilon}_{t+h}$ is the n-dimensional forecast error at horizon h of our n-variable model. This measure is invariant to elementary operations that involve different variables (*TMSFE* is not invariant to such transformations), and also to elementary operations that involve the same variable at different horizons (neither *TMSFE* nor |*MSFE*| is invariant to such transformations). In our Monte-Carlo, the above expectation is evaluated for every model, by averaging over the simulations.

There is one complication associated with simulating 100 different DGPs. Simple averaging across different DGPs is not appropriate, because the forecast errors of different DGPs do not have identical variance-covariance matrices. Lütkepohl(1985) normalizes the forecast errors by their true variance-covariance matrix in each case before aggregating. Unfortunately, this would be a very time consuming procedure for a measure like *GFESM,* which involves stacked errors over many horizons. Instead, for each information criterion, we calculate the percentage gain in forecasting measures, comparing the full-rank models selected by $IC(p)$, with the reduced-rank models chosen by $IC(p,r)$.

The percentage gain is computed using natural logs of ratios of respective loss functions, since this implies symmetry of results for gains and losses. This procedure is done at every iteration, for every DGP, and the final results are then averaged.

5 Monte-Carlo simulation results

5.1 Selection of lag, rank, and the number of cointegrating vectors

See Tables 1 and 2. Discussion to be written.

5.2 Forecasts

To be written.

6 Conclusion

To be written.

References

[1] Ahn, S.K. and G.C. Reinsel (1988), "Nested reduced-rank autoregressive models for multiple time series", *Journal of the American Statistical Association*, 83, 849-856.

[2] Ahn, S.K. and G.C. Reinsel (1990), "Estimation for partially nonstationary multivariate autoregressive models," *Journal of the American Statistical Association*, 85, 813-823.

[3] Christoffersen, P.F. and Diebold, F.X. (1998), "Cointegration and long-horizon forecasting," *Journal of Business and Economic Statistics*, vol. 16, 4, pp. 450-458.

[4] Clements, M.P. and D.F. Hendry (1993), "On the limitations of comparing mean squared forecast errors," *Journal of Forecasting*, 12, 617-637 (with discussions).

[5] Clements, M.P. and D.F. Hendry (1995), "Forecasting in cointegrated systems", *Journal of Applied Econometrics*, 10, 127-146.

[6] Diebold, F.X. and Kilian L. (2001), "Measuring predictability: Theory and macroeconomic applications," *Journal of Applied Econometrics*, 16 (6): 657-669.

[7] Engle, R.F. and C.W.J. Granger (1987), "Cointegration and error correction: Representation, estimation and testing", *Econometrica*, 55, 251-276.

[8] Engle, R.F. and Kozicki, S.(1993), "Testing for Common Features," *Journal of Business and Economic Statistics*, vol. 11, pp. 369-395, with discussions.

[9] Engle, R.F. and S. Yoo (1987), "Forecasting and testing in cointegrated systems", *Journal of Econometrics*, 35, 143-159.

[10] Gonzalo, J (1994), "Five alternative methods of estimating long-run equilibrium relationships," *Journal of Econometrics*, 60 (1-2): 203-233.

[11] Granger, C.W.J.(1981), "Some properties of time series data and their use in econometric model specification", *Journal of Econometrics*, 16, pp. 121-130.

[12] Hecq, A., Palm, F.C., Urbain, J.-P. (2005), "Testing for common cyclical features in VAR models with cointegration," *Journal of Econometrics,* 132,117-142.

[13] Hoffman, D.L. and Rasche, R.H. (1996), "Assessing forecast performance in a cointegrated system," *Journal of Applied Econometrics*, 11 (5): 495-517.

[14] Issler, J.V. and Vahid, F. (2001), "Common Cycles and the Importance of Transitory Shocks to Macroeconomic Aggregates," *Journal of Monetary Economics,* 47, 449-475.

[15] Johansen, S.(1988), "Statistical Analysis of Cointegrating Vectors," *Journal of Economic Dynamics and Control*, vol. 12, pp. 231-254.

[16] Johansen, S.(1991), "Estimation and Hypothesis Testing of Cointegration Vectors in Gaussian Vector Autoregressive Models," *Econometrica*, vol. 59, pp. 1551-1580.

[17] Johansen, S.(2002), "A small sample correction of the test for cointegrating rank in the vector autoregressive model", *Econometrica*, 70, 1929-1961.

[18] Lin, J.L. and R.S. Tsay (1996), "Cointegration constraints and forecasting: An empirical examination", *Journal of Applied Econometrics,* 11, 519-538.

[19] Lütkepohl, H. (1985), "Comparison of criteria for estimating the order of a vector autoregressive process", *Journal of Time Series Analysis*, 6, 35-52.

[20] Phillips, P.C.B. (1996), "Economic Model Determination," *Econometrica,* 64, 763-812.

[21] Phillips, P.C.B. and Ploberger, W. (1996), "An asymptotic theory of Bayesian inference for time series," *Econometrica*, 64, 381-413.

[22] Ploberger, W. and Phillips, P.C.B. (2003), "Empirical limits for time series econometric models," *Econometrica*, 71, 627-673.

[23] Poskitt, D.S. (1987), "Precision, complexity and Bayesian model determination," *Journal of the Royal Statistical Society, Series B,* 49, 199-208.

[24] Rissanen, J.J. (1987), "Stochastic Complexity," (with discussion), *Journal of the Royal Statistical Society, Series B* 49, 223-239 and 252-265.

[25] Silverstovs, B., Engsted T. and Haldrup N. (2004), "Long-run forecasting in multicointegrated systems," *Journal of Forecasting*, 23 (5): 315-335.

[26] Vahid, F. and R.F. Engle (1993), "Common trends and common cycles", *Journal of Applied Econometrics*, 8, 341-360.

[27] Vahid, F. and Issler, J.V. (2002), "The importance of common cyclical features in VAR analysis: A Monte Carlo study," *Journal of Econometrics*, 109 (2): 341-363.

[28] Wallace, C.S. (2005), *Statistical and Inductive Inference by Minimum Message Length,* Berlin: Springer.

[29] Wallace, C.S. and Dowe, D. (1999), "Refinements of MDL and MML coding," *Computer Journal*, 42, 330-337.

[30] Wallace, C.S. and Freeman, P.R. (1987), "Estimation and inference by compact coding," *Journal of the Royal Statistical Society, Series B,* 49, 240-252.

A The Fisher information matrix of the reduced rank VECM

Assuming that the first observation in the sample is labelled observation $-p+1$ and the sample contains $T+p$ observations, we write the K-variable reduced rank VECM:

$$\Delta y_t = \gamma' \begin{pmatrix} I_q & \beta' \end{pmatrix} y_{t-1} + \begin{pmatrix} I_r \\ C' \end{pmatrix} [D_1 \Delta y_{t-1} + D_2 \Delta y_{t-2} + \cdots + D_p \Delta y_{t-p}] + \mu + e_t$$

in stacked form:

$$\Delta Y = Y_{-1} \begin{pmatrix} I_q \\ \beta \end{pmatrix} \gamma + WD \begin{pmatrix} I_r & C \end{pmatrix} + \iota_T \mu' + E$$

where

$$\begin{aligned}
\underset{T\times K}{\Delta Y} &= \begin{bmatrix} \Delta y_1' \\ \vdots \\ \Delta y_T' \end{bmatrix}, \quad \underset{T\times K}{Y_{-1}} = \begin{bmatrix} y_0' \\ \vdots \\ y_{T-1}' \end{bmatrix}, \quad \underset{T\times K}{E} = \begin{bmatrix} e_1' \\ \vdots \\ e_T' \end{bmatrix} \\
\underset{T\times Kp}{W} &= \begin{pmatrix} \Delta Y_{-1} & \cdots & \Delta Y_{-p} \end{pmatrix} = \begin{bmatrix} \Delta y_0' & \cdots & \Delta y_{-p+1}' \\ \vdots & \vdots & \vdots \\ \Delta y_{T-1}' & \cdots & \Delta y_{T-p}' \end{bmatrix} \\
\underset{Kp\times r}{D} &= \begin{pmatrix} D_1' \\ \vdots \\ D_p' \end{pmatrix}
\end{aligned}$$

and ι_T is a $T \times 1$ vector of ones. When e_t are $N(0, \Omega)$ and serially uncorrelated, the log-likelihood function conditional on the first p observations being known is:

$$\begin{aligned}
\ln l(\theta, \omega) &= -\frac{KT}{2} \ln(2\pi) - \frac{T}{2} \ln|\Omega| - \frac{1}{2} \sum_{t=1}^{T} e_t' \Omega^{-1} e_t \\
&= -\frac{KT}{2} \ln(2\pi) - \frac{T}{2} \ln|\Omega| - \frac{1}{2} tr\left(E\Omega^{-1}E'\right)
\end{aligned}$$

where

$$\theta = \begin{pmatrix} vec(\beta) \\ vec(\gamma) \\ vec(D) \\ vec(C) \\ \mu \end{pmatrix}$$

is a $(K-q)q + Kq + Kpr + r(K-r) + K$ matrix of mean parameters, and $\omega = vech(\Omega)$ is a $K(K+1)/2$ vector of unique elements of the variance matrix. The differential of the log-likelihood is (see Magnus and Neudecker, 1988):

$$\begin{aligned}
d \ln l(\theta, \omega) &= -\frac{T}{2} tr \Omega^{-1} d\Omega + \frac{1}{2} tr\left(\Omega^{-1} d\Omega \Omega^{-1} E'E\right) - \frac{1}{2} tr\left(\Omega^{-1} E' dE\right) - \frac{1}{2} tr\left(\Omega^{-1} dE' E\right) \\
&= \frac{1}{2} tr\left(\Omega^{-1}\left(E'E - T\Omega\right)\Omega^{-1} d\Omega\right) - tr\left(\Omega^{-1} E' dE\right)
\end{aligned}$$

and the second differential is:

$$\begin{aligned}
d^2 \ln l(\theta, \omega) &= tr\left(d\Omega^{-1}\left(E'E - T\Omega\right)\Omega^{-1} d\Omega\right) + \frac{1}{2} tr\left(\Omega^{-1}\left(2E'dE - Td\Omega\right)\Omega^{-1} d\Omega\right) \\
&\quad - tr\left(d\Omega^{-1} E' dE\right) - tr\left(\Omega^{-1} dE' dE\right).
\end{aligned}$$

Since we eventually want to evaluate the Fisher information matrix at the maximum likelihood estimator, and at the maximum likelihood estimator $\hat{E}'\hat{E} - T\hat{\Omega} = 0$, and also $\hat{\Omega}^{-1}\hat{E}'dE/d\theta = 0$ (these are apparent from the first differentials), we can delete these terms from the second differential, and use $tr(AB) = vec(A')'vec(B)$ to obtain:

$$\begin{aligned} d^2 \ln l(\theta,\omega) &= -\frac{T}{2} tr\left(\Omega^{-1}d\Omega\Omega^{-1}d\Omega\right) - tr\left(\Omega^{-1}dE'dE\right) \\ &= -\frac{T}{2}(d\omega)'\mathbf{D}_K'\left(\Omega^{-1}\otimes\Omega^{-1}\right)\mathbf{D}_K d\omega - (vec(dE))'\left(\Omega^{-1}\otimes I_T\right)vec(dE) \end{aligned}$$

where $\mathbf{D}_K$ is the "duplication matrix". From the model, we can see that

$$dE = -Y_{-1}\begin{pmatrix} 0 \\ d\beta \end{pmatrix}\gamma - Y_{-1}\begin{pmatrix} I_q \\ \beta \end{pmatrix}d\gamma - WdD\begin{pmatrix} I_r & C \end{pmatrix} - WD\begin{pmatrix} 0 & dC \end{pmatrix} - \iota_T d\mu'$$

and therefore

$$vec(dE) = -\begin{bmatrix} \gamma'\otimes Y_{-1}^{(2)} & I_K \otimes Y_{-1}\begin{pmatrix} I_q \\ \beta \end{pmatrix} & \begin{pmatrix} I_r \\ C' \end{pmatrix}\otimes W & \begin{pmatrix} 0 \\ I_{K-r} \end{pmatrix}\otimes WD & I_K\otimes\iota_T \end{bmatrix} d\theta.$$

Hence the elements of the Fisher information matrix are:

$$\begin{aligned} FIM_{11} &= \gamma\Omega^{-1}\gamma'\otimes Y_{-1}^{(2)\prime}Y_{-1}^{(2)}, \quad FIM_{12} = \gamma\Omega^{-1}\otimes Y_{-1}^{(2)\prime}Y_{-1}\begin{pmatrix} I_q \\ \beta \end{pmatrix}, \\ FIM_{13} &= \gamma\Omega^{-1}\begin{pmatrix} I_r \\ C' \end{pmatrix}\otimes Y_{-1}^{(2)\prime}W, \quad FIM_{14} = \gamma\Omega^{-1}\begin{pmatrix} 0 \\ I_{K-r} \end{pmatrix}\otimes Y_{-1}^{(2)\prime}WD \\ FIM_{15} &= \gamma\Omega^{-1}\otimes Y_{-1}^{(2)\prime}\iota_T \\ FIM_{22} &= \Omega^{-1}\otimes\begin{pmatrix} I_q & \beta' \end{pmatrix}Y_{-1}'Y_{-1}\begin{pmatrix} I_q \\ \beta \end{pmatrix}, \quad FIM_{23} = \Omega^{-1}\begin{pmatrix} I_r \\ C' \end{pmatrix}\otimes\begin{pmatrix} I_q & \beta' \end{pmatrix}Y_{-1}'W \\ FIM_{24} &= \Omega^{-1}\begin{pmatrix} 0 \\ I_{K-r} \end{pmatrix}\otimes\begin{pmatrix} I_q & \beta' \end{pmatrix}Y_{-1}'WD, \quad FIM_{25} = \Omega^{-1}\otimes\begin{pmatrix} I_q & \beta' \end{pmatrix}Y_{-1}'\iota_T \\ FIM_{33} &= \begin{pmatrix} I_r & C \end{pmatrix}\Omega^{-1}\begin{pmatrix} I_r \\ C' \end{pmatrix}\otimes W'W, \quad FIM_{34} = \begin{pmatrix} I_r & C \end{pmatrix}\Omega^{-1}\begin{pmatrix} 0 \\ I_{K-r} \end{pmatrix}\otimes W'WD \\ FIM_{35} &= \begin{pmatrix} I_r & C \end{pmatrix}\Omega^{-1}\otimes W'\iota_T \\ FIM_{44} &= \begin{pmatrix} 0 & I_{K-r} \end{pmatrix}\Omega^{-1}\begin{pmatrix} 0 \\ I_{K-r} \end{pmatrix}\otimes D'W'WD, \quad FIM_{45} = \begin{pmatrix} 0 & I_{K-r} \end{pmatrix}\Omega^{-1}\otimes D'W'\iota_T \\ FIM_{55} &= \Omega^{-1}\otimes\iota_T'\iota_T = \Omega^{-1}\times T \end{aligned}$$

B Tables

Table 1: Performance of $IC(p, r, q)$ in a weak 1,1,2 design
and its comparison with the usual application of the Johansen method

	$T = 100$				$T = 200$			
AIC	$q < q^*$	$q = q^*$	$q > q^*$	Total	$q < q^*$	$q = q^*$	$q > q^*$	Total
$r < r^*$	(1,0,0)	(3,0,0)	(5,0,0)	(8,0,0)	(0,0,0)	(2,0,0)	(3,0,0)	(5,0,0)
$r = r^*$	(0,1,5)	(0,9,10)	(0,16,15)	(0,26,30)	(0,1,4)	(0,11,8)	(0,23,12)	(0,35,24)
$r > r^*$	(0,1,10)	(0,3,8)	(0,5,9)	(0,9,26)	(0,1,10)	(0,3,8)	(0,5,9)	(0,10,26)
Total	(1,2,15)	(3,12,18)	(5,21,23)	(8,35,57)	(0,2,13)	(2,15,16)	(3,29,21)	(5,45,50)
AIC+J	(4,13,2)	(17,29,3)	(11,20,2)	(31,62.7)	(0,3,0)	(9,48,3)	(5,31,2)	(14,81,5)
HQ	$q < q^*$	$q = q^*$	$q > q^*$	Total	$q < q^*$	$q = q^*$	$q > q^*$	Total
$r < r^*$	(1,0,0)	(13,0,0)	(13,0,0)	(27,0,0)	(0,0,0)	(8,0,0)	(7,0,0)	(15,0,0)
$r = r^*$	(0,3,3)	(0,22,6)	(0,20,6)	(0,45,15)	(0,3,3)	(0,32,6)	(0,26,6)	(0,59,14)
$r > r^*$	(0,1,2)	(0,3,2)	(0,3,2)	(0,7,5)	(0,1,2)	(0,2,3)	(0,2,3)	(0,5,7)
Total	(1,5,5)	(13,25,8)	(13,23,8)	(27,53,20)	(0,2,4)	(8,34,8)	(7,28,8)	(15,64,21)
HQ+J	(7,8,0)	(31,18,0)	(22,13,0)	(61,39,0)	(0,2,0)	(22,37,0)	(15,24,0)	(37,63,0)
SC	$q < q^*$	$q = q^*$	$q > q^*$	Total	$q < q^*$	$q = q^*$	$q > q^*$	Total
$r < r^*$	(9,0,0)	(31,0,0)	(14,0,0)	(55,0,0)	(1,0,0)	(25,0,0)	(8,0,0)	(34,0,0)
$r = r^*$	(0,9,1)	(0,22,1)	(0,9,1)	(0,40,3)	(0,4,1)	(0,42,2)	(0,12,1)	(0,59,3)
$r > r^*$	(0,1,0)	(0,1,0)	(0,1,0)	(0,2,0)	(0,0,0)	(0,1,0)	(0,1,0)	(0,3,1)
Total	(9,10,1)	(31,22,1)	(14,10,1)	(55,42,3)	(1,5,1)	(25,44,2)	(8,13,1)	(35,62,4)
SC+J	(10,4,0)	(43,7,0)	(31,5,0)	(85,15,0)	(1,2,0)	(38,20,0)	(27,13,0)	(66,34,0)

Note: The total of the three entries (a,b,c) in each cell show the percentage of times that the selected model fell in the category that is identified by the column and row labels. The number (a) shows the percentage where $p < p^*$, (b) shows the percent where $p = p^*$, and (c) shows the percent where $p > p^*$. The row labled X+J, shows this information for the method that is commonly used in practice where lag-length is chosen by model selection criterion X, and then Johansen procedure is used for determining q.

Table 2: Performance of $IC\,(p,r,q)$ in a 1,1,2 design
and its comparison with the usual application of the Johansen method

	$T=100$				$T=200$			
AIC	$q<q^*$	$q=q^*$	$q>q^*$	Total	$q<q^*$	$q=q^*$	$q>q^*$	Total
$r<r^*$	(0,0,0)	(2,0,0)	(3,0,0)	(5,0,0)	(0,0,0)	(1,0,0)	(1,0,0)	(2,0,0)
$r=r^*$	(0,0,0)	(0,17,9)	(0,31,19)	(0,48,28)	(0,0,0)	(0,19,6)	(0,42,13)	(0,61,19)
$r>r^*$	(0,0,0)	(0,4,2)	(0,7,5)	(0,11,7)	(0,0,0)	(0,4,2)	(0,9,3)	(0,13,5)
Total	(0,0,0)	(2,21,12)	(3,39,23)	(5,60,35)	(0,0,0)	(1,23,7)	(1,51,16)	(2,74,24)
AIC+J	(1,9,1)	(9,41,4)	(6,26,3)	(16,76,8)	(0,3,0)	(4,52,3)	(2,34,2)	(6,89,5)
HQ	$q<q^*$	$q=q^*$	$q>q^*$	Total	$q<q^*$	$q=q^*$	$q>q^*$	Total
$r<r^*$	(0,0,0)	(9,0,0)	(9,0,0)	(19,0,0)	(0,0,0)	(4,0,0)	(4,0,0)	(8,0,0)
$r=r^*$	(0,2,0)	(0,36,3)	(0,33,3)	(0,71,7)	(0,1,0)	(0,47,2)	(0,38,1)	(0,86,3)
$r>r^*$	(0,0,0)	(0,2,0)	(0,2,0)	(0,4,0)	(0,0,0)	(0,1,0)	(0,1,0)	(0,3,0)
Total	(0,3,0)	(9,38,3)	(9,34,3)	(19,75,7)	(0,1,0)	(4,48,2)	(4,39,1)	(8,89,3)
HQ+J	(2,7,0)	(20,34,0)	(15,22,0)	(37,63,0)	(0,2,0)	(11,48,0)	(8,31,0)	(19,81,0)
SC	$q<q^*$	$q=q^*$	$q>q^*$	Total	$q<q^*$	$q=q^*$	$q>q^*$	Total
$r<r^*$	(3,0,0)	(23,0,0)	(10,0,0)	(36,0,0)	(0,0,0)	(15,0,0)	(4,0,0)	(36,0,0)
$r=r^*$	(0,8,0)	(0,40,0)	(0,15,0)	(0,63,1)	(0,4,0)	(0,61,0)	(0,16,0)	(0,80,0)
$r>r^*$	(0,0,0)	(0,0,0)	(0,0,0)	(0,0,0)	(0,0,0)	(0,0,0)	(0,0,0)	(0,0,0)
Total	(3,8,0)	(23,41,0)	(10,15,0)	(36,64,1)	(0,4,0)	(15,61,0)	(4,16,0)	(19,81,0)
SC+J	(3,5,0)	(31,23,0)	(24,14,0)	(58,42,0)	(0,2,0)	(22,36,0)	(16,24,0)	(38,62,0)
MML	$q<q^*$	$q=q^*$	$q>q^*$	Total	$q<q^*$	$q=q^*$	$q>q^*$	Total
$r<r^*$	(6,0,0)	(23,0,0)	(1,0,0)	(30,0,0)	(1,0,0)	(14,0,0)	(1,0,0)	(15,0,0)
$r=r^*$	(0,13,0)	(0,51,0)	(0,3,0)	(0,67,0)	(0,6,0)	(0,75,0)	(0,2,0)	(0,83,1)
$r>r^*$	(0,0,0)	(0,1,0)	(1,0,0)	(1,1,0)	(0,0,0)	(0,1,0)	(0,0,0)	(0,1,0)
Total	(6,14,0)	(23,52,0)	(3,3,0)	(32,68,0)	(1,6,0)	(14,76,1)	(1,3,0)	(15,84,1)

Note: The total of the three entries (a,b,c) in each cell show the percentage of times that the selected model fell in the category that is identified by the column and row labels. The number (a) shows the percentage where $p<p^*$, (b) shows the percent where $p=p^*$, and (c) shows the percent where $p>p^*$. The row labled X+J, shows this information for the method that is commonly used in practice where lag-length is chosen by model selection criterion X, and then Johansen procedure is used for determining q.