Econometric Theory, **23**, 2007, 638–685. Printed in the United States of America.
DOI: 10.1017/S0266466607070284

A MODIFIED INFORMATION CRITERION FOR COINTEGRATION TESTS BASED ON A VAR APPROXIMATION

ZHONGJUN QU
University of Illinois at Urbana-Champaign

PIERRE PERRON
Boston University

We consider the cointegration tests of Johansen (1988, *Journal of Economic Dynamics and Control* 12, 231–254; 1991, *Econometrica* 59, 1551–1580) when a vector autoregressive (VAR) process of order k is used to approximate a more general linear process with a possibly infinite VAR representation. Traditional methods to select the lag order, such as Akaike's information criterion (AIC) or the Bayesian information criterion, often lead to too parsimonious a model with the implication that the cointegration tests suffer from substantial size distortions in finite samples. We extend the analysis of Ng and Perron (2001, *Econometrica* 69, 1519–1554) to derive a modified Akaike's information criterion (MAIC) in this multivariate setting. The idea is to use the information specified by the null hypothesis as it relates to restrictions on the parameters of the model to keep an extra term in the penalty function of the AIC. This MAIC takes a very simple form for which this extra term is simply the likelihood ratio test for testing the null hypothesis of r against more than r cointegrating vectors. We provide theoretical analyses of its validity and of the fact that cointegration tests constructed from a VAR whose lag order is selected using the MAIC have the same limit distribution as when the order is finite and known. We also provide theoretical and simulation analyses to show how the MAIC leads to VAR approximations that yield tests with drastically improved size properties with little loss of power.

1. INTRODUCTION

In an effort to model long-run economic relationships, a large literature on cointegration has developed, starting with the seminal work of Granger (1981) and Engle and Granger (1987). One branch of this literature pertains to finding the number of cointegrating relations, that is, the cointegrating rank of the under-

We are grateful to two referees for especially useful and constructive comments. Address correspondence to Zhongjun Qu, Department of Economics, University of Illinois at Urbana-Champaign, 1206 S. Sixth Street, Champaign, Illinois 61820, USA; e-mail: zqu@uiuc.edu.

 © 2007 Cambridge University Press 0266-4666/07 $15.00

https://doi.org/10.1017/S0266466607070284 Published online by Cambridge University Press

lying system. For that purpose, various statistics have been proposed, with the likelihood ratio (LR) tests of Johansen (1988, 1991) being widely used. The latter are based on estimating a vector autoregression of the form (omitting deterministic components)

$$\Delta y_t = \Pi y_{t-1} + \sum_{j=1}^{k} \Gamma_j \Delta y_{t-j} + e_t \qquad (t = k+2, \ldots, T).$$

Under the assumption of Gaussian errors and known lag order, the LR tests for the hypothesis that Π has rank, say, r are constructed using the technique of reduced rank regression involving the solution to an eigenvalue problem (see Anderson, 1951). The asymptotic distributions of the tests, again assuming known lag order, are derived by Johansen (1988, 1996) and Ahn and Reinsel (1990). These are nonstandard distributions that can be expressed in terms of Brownian motions (and also deterministic components, if they are included in the regression).

In practice, the lag order k is unknown and needs to be chosen prior to constructing the tests. Many studies have shown by simulations that the tests may suffer from serious size distortions if the lag order is not chosen appropriately; see, for example, Cheung and Lai (1993), Yap and Reinsel (1995), and Lütkepohl and Saikkonen (1999). To this end, it has been proposed to use information criteria such as Akaike's information criterion (AIC) (Akaike, 1973) and the Bayesian information criterion (BIC) proposed by Schwarz (1978). Lütkepohl and Saikkonen (1999) provide some theoretical and simulation evidence in support of such a proposal. They show that the AIC and the BIC are asymptotically valid criteria in the sense that the resulting cointegration tests will have the same limiting distributions as if the lag order is finite and known.

Recently, there also has been increased interest in extending the analysis of cointegrated systems beyond the finite-order vector autoregressive (VAR) case. Yap and Reinsel (1995) consider testing for cointegration in a vector autoregressive moving average model (VARMA) under the assumption that the orders for the AR and MA components are known a priori. Lütkepohl and Saikkonen (1997) study the analysis of impulse response functions in infinite-order VAR processes. Kuersteiner (2005) extends the analysis of Ng and Perron (1995) to prove the validity of a general-to-specific testing strategy to select the vector autoregression order when the process is a strictly stationary infinite-order VAR model and the goal is to get coefficient estimates that are consistent and asymptotically normal.

More closely related to our work is the paper by Saikkonen (1992), who considers estimating and testing for cointegration using a finite VAR approximation when the true process is an infinite-order vector autoregression. Saikkonen and Luukkonen (1997) justify the use of Johansen's LR tests for cointegration for such processes and give sufficient conditions on the lag order, under which the associated tests have the same limiting distributions as in the

https://doi.org/10.1017/S0266466607070284 Published online by Cambridge University Press

finite-order case. In this context, choosing a lag order amounts to finding a good enough approximation to the true data generating process (DGP) to yield a cointegration test that has an exact size close to nominal size while retaining acceptable power. Simulation results by Lütkepohl and Saikkonen (1999), and others presented in this paper, show that the use of the AIC or BIC leads to tests with unacceptable size distortions. The problem is particularly serious when the noise component has a moving average (MA) representation with a root close to one, in which case the lag lengths chosen by the AIC and the BIC are too small, resulting in poor approximations to the true process and consequently in tests with substantial size distortions. As shown subsequently, this can also occur in a finite VAR system when the coefficients of the last few lags are small but have nevertheless an important effect on the overall correlation structure of the series. Because it is important to have a reliable procedure that yields tests with acceptable size and power, it is important to find a useful alternative to traditional information criteria for selecting the lag order.

In this paper, we extend the work of Ng and Perron (2001), who proposed a modified Akaike's information criterion (MAIC) in the context of choosing the autoregressive lag order to construct unit root tests. The main idea underlying their procedure is to use the information specified by the null hypothesis, as it relates to restrictions on the parameters of the model, to keep an extra term in the penalty function of the AIC. We show that this strategy leads to the following simple MAIC in the context of testing for cointegration using an LR test based on a VAR approximation:

$$MAIC = \ln|\hat{\Sigma}| + \frac{2(\tau_k(r) + kn^2)}{N},$$

where $\hat{\Sigma}$ is an estimate of the covariance matrix of e_t, $N = T - K_T - 1$, with K_T the maximal lag order allowed, and the extra term $\tau_k(r)$, relative to the AIC, is simply the LR test of r against more than r cointegrating vectors. It permits an overall penalty function that better represents the cost associated with an increase in the lag order. Simulations show that when the true order is infinite, or when the true order is finite but there is a weak correlation for the last lags, this extra term is highly sensitive to changes in the lag order and implies a selected lag higher than that obtained with a standard AIC. This leads to drastic improvements in the exact size of the cointegration tests with little loss in power. The MAIC, via the extra term $\tau_k(r)$, also adapts to the cointegration rank, the dimension of the system, and the nature of the deterministic components in the data and the regression.

We provide theoretical justifications for its validity, showing that the resulting cointegration tests have the same limiting distributions as when the lag order is finite and known. We also give simulation evidence showing that, compared with the AIC and the BIC, the resulting procedure yields drastic size improvements while preserving the power of the test.

https://doi.org/10.1017/S0266466607070284 Published online by Cambridge University Press

The idea of using sample-dependent quantities in the penalty factor of an information criterion is not new. For instance, Phillips and Ploberger (1994) used a Bayesian approach to suggest the Posterior Information Criterion (PIC), whose penalty depends, in particular, on the second moment matrix of the regressors [see also the Fisher Information Criterion (FIC) of Wei, 1992; for a discussion of PIC in cointegration context see Phillips, 1996; Chao and Phillips, 1999]. But the motivation and performance are very different. The PIC and FIC allow different penalty parameters associated with stationary, integrated, or trending regressors. When performing unit root tests, these information criteria still yield tests that have high size distortions in the presence of a negative MA component (see Phillips and Ploberger, 1994, Table 2). The motivation behind the statistic that we incorporate in the penalty term is very different and directly motivated by the testing problem under consideration. We work under the null hypothesis of some cointegrating rank, and the aim is to have the penalty term adapt to the nature of the serial correlation present. In this sense, our MAIC is tailored to work for the problem of testing for cointegration and is unlikely to be suitable for other problems such as the selection of a model for forecasting purposes, in which case the PIC and FIC are likely to perform better. Also note that our approach involves a mix of information criteria and traditional LR testing procedures and hence is not immune to the problem of postmodel selection bias; see Leeb and Pötscher (2005).

The structure of the paper is as follows. Section 2 presents the models and cointegration test considered. Section 3 derives the exact form of the MAIC, stated earlier, in the context of a finite-order vector autoregression and illustrates its performance with one specific example. Section 4 shows its validity in the general case of a linear process having an infinite VAR representation and shows that cointegration tests constructed from a VAR approximation with a lag order selected by the MAIC have the same limit distribution as when the order is finite and known. The analysis also includes extensions to cover the case where various types of deterministic components are included. Section 5 studies the properties of the MAIC in a local asymptotic framework whereby the MA root is local to one, which explains why it provides an improvement over the traditional AIC when the true process is of infinite order. Section 6 presents simulations to show how, in such a case, the use of the MAIC yields tests with improved finite-sample properties, compared to what can be achieved using the AIC or BIC. Section 7 offers brief concluding remarks and an Appendix some technical derivations.

2. PRELIMINARIES

Let $y_t = (y_{1,t}, \ldots, y_{n,t})'$ be an n-vector of time series variables available for $t = 1, \ldots, T$. The cointegration test of Johansen (1988, 1991) is based on a VAR(k) approximation of the form

$$\Delta y_t = \mu + \Pi y_{t-1} + \sum_{j=1}^{k} \Gamma_j \Delta y_{t-j} + e_t \qquad (t = k+2, \ldots, T).$$

https://doi.org/10.1017/S0266466607070284 Published online by Cambridge University Press

Under the hypothesis of r cointegrating vectors, the matrix Π has reduced rank, such that $\Pi = \alpha\beta'$ where α is an $n \times r$ matrix of adjustment coefficients and β is an $n \times r$ matrix whose columns contain the cointegrating vectors. In the literature, three cases are usually considered with respect to the nature of the deterministic components. These are

$$\text{Case A: } \Delta y_t = \alpha\beta' y_{t-1} + \sum_{j=1}^{k} \Gamma_j \Delta y_{t-j} + e_t, \tag{1}$$

$$\text{Case B: } \Delta y_t = \alpha(u + \beta' y_{t-1}) + \sum_{j=1}^{k} \Gamma_j \Delta y_{t-j} + e_t, \tag{2}$$

$$\text{Case C: } \Delta y_t = u + \alpha\beta' y_{t-1} + \sum_{j=1}^{k} \Gamma_j \Delta y_{t-j} + e_t. \tag{3}$$

For Case A, no deterministic component is present, for Case B the constant is restricted, and for Case C, the constant is unrestricted and allows trend in the I(1) series.[1]

The construction of the test is based on an eigenvalue problem. Let R_{0t} be the residuals from a regression of Δy_t on $(\Delta y_{t-1}, \ldots, \Delta y_{t-k})$ and R_{1t} be the residuals from a regression of y_{t-1} on $(\Delta y_{t-1}, \ldots, \Delta y_{t-k})$. For Case B, the dependent variable is $(1, y_{t-1})$ in the construction of R_{1t}, and for Case C, a constant is added to the list of regressors. Define $S_{ij} = T^{-1} \sum_{t=1}^{T} R_{it} R_{jt}'$ and let $(\hat{\lambda}_{r+1}, \ldots, \hat{\lambda}_n)$ be the $n - r$ smallest eigenvalues of the system $|\lambda S_{11} - S_{10} S_{00}^{-1} S_{01}| = 0$. To test the null hypothesis of r cointegrating vectors versus the alternative that there are more than r cointegrating vectors, Johansen's trace statistic[2] (with a finite-sample correction as suggested by Reinsel and Ahn, 1992) is

$$LR = -(N - n(k+1) - \gamma_\delta) \sum_{j=r+1}^{n} \log(1 - \hat{\lambda}_j), \tag{4}$$

where $N = T - k - 1$ denotes the sample size used for estimation, k is the lag order, and γ_δ is the number of parameters associated with the deterministic components. Johansen (1988, 1991) derived the limit distribution of the LR test assuming a known finite order k. Saikkonen and Luukkonen (1997) extended the analysis to show that the same limit distribution applies when the data are generated by a general linear process of weak dependence, in which case the true autoregressive order is infinite. This will be the case if the order k of the VAR approximation is chosen in such a way that it increases to infinity as the sample size increases but at a controlled rate (less than $T^{1/3}$); see also Lütkepohl and Saikkonen (1999). They recommend the use of an information criterion to select k, such as the BIC or AIC.

In economic applications, a nonfinite true autoregressive order is often the result of an MA component in the VARMA representation of the system. Such cases are indeed relevant in practice. For example, suppose that some compo-

https://doi.org/10.1017/S0266466607070284 Published online by Cambridge University Press

nent of the system $y_{i,t}$ is the sum of a random walk and a white noise measurement error; then $\Delta y_{i,t}$ is an MA(1) process with a negative MA coefficient. In practice, the log of aggregate output is often modeled as an unobserved component autoregressive moving average (ARMA) model, consisting of a stochastic trend that accounts for the long-term growth and a stationary component that represents the cycle. This implies an equivalent ARMA representation for the first difference of $y_{i,t}$. Of related interest, Yap and Reinsel (1995) consider a system of three U.S. monthly interest rate series and argue that it can be modeled as a VARMA(1,1) process with the largest (in absolute value) MA root being -0.40. In principle, it is possible to test for cointegration using a VARMA process directly (Yap and Reinsel, 1995). However, in the presence of a negative MA component, estimates such as those based on maximum likelihood can have finite-sample properties that make them difficult to assess (e.g., the issue of multiple local optima and the so-called pile-up problem). Hence, it is often simpler and more reliable to adopt a vector autoregression, which is easy to estimate and, as we shall see, yields tests with good properties provided the lag length is selected appropriately.

To illustrate the potential pitfalls of using an information criterion to select the lag length in the context of a system containing an MA component, we present the size of the LR test from a simple simulation exercise (see also Lütkepohl and Saikkonen, 1999). Consider a bivariate process with no cointegration, that is,

$$\Delta y_t = \varepsilon_t + \varphi \varepsilon_{t-1} \qquad (t = 1, \ldots, T), \tag{5}$$

where $\Delta y_t = (\Delta y_{1,t}, \Delta y_{2,t})'$, $\varepsilon_t \sim i.i.d.\ N(0, I_2)$, and $y_0 = \varepsilon_0 = 0$. The sample size is $T = 200$, and the number of replications is 10,000. The LR test is constructed from a VAR(k) approximation with k selected using either the BIC or the AIC restricting the search to values of k between 0 and $K_T = int(12(T/100)^{1/4})$, which is often employed in practice. The experiments are repeated for various values of φ between -0.8 and 0.8. To evaluate the size of the LR test we use the appropriate asymptotic critical values corresponding to Cases A–C, for tests with 5% nominal size. The results are presented in Table 1. First, they clearly show that serious size distortions exist when a large negative MA root is present. In the case of a large positive MA root, size distortions also exist, though they are not as serious. Second, using the BIC implies tests with substantially worse properties than if the AIC is used, given that the former imposes a stronger penalty that results in a smaller order being selected. Third, the tests show more severe size distortions when more deterministic components are included in the estimated system. This is particularly relevant in practice because most economic variables contain a nonzero mean or a trend.

With this background information, the rest of the paper is devoted to the derivation of an alternative information criterion following the work of Ng and

https://doi.org/10.1017/S0266466607070284 Published online by Cambridge University Press

TABLE 1. Finite-sample size of the LR test for a 5% nominal level

	Case A		Case B		Case C	
φ	AIC	BIC	AIC	BIC	AIC	BIC
−0.8	0.336	0.759	0.463	0.890	0.515	0.916
−0.6	0.128	0.299	0.168	0.408	0.201	0.466
−0.4	0.083	0.124	0.111	0.152	0.122	0.181
0.0	0.053	0.048	0.066	0.060	0.061	0.056
0.4	0.067	0.071	0.087	0.085	0.083	0.095
0.6	0.065	0.068	0.086	0.089	0.084	0.089
0.8	0.085	0.079	0.107	0.097	0.100	0.091

Note: Bivariate system generated by $\Delta y_t = \varepsilon_t + \varphi\varepsilon_{t-1}$, with $\varepsilon_t \sim i.i.d.\ N(0, I_2), T = 200$.

Perron (2001), which allows tests with finite-sample size close to nominal size, and this with little loss of power.

3. THE MODIFIED AKAIKE'S INFORMATION CRITERION

To derive the form of the modified information criterion (MIC) we shall start by assuming that the data are generated by a finite-order vector autoregression with independent and identically distributed (i.i.d.) normal innovations, and we will subsequently analyze its large-sample properties in a more general context. We also abstract from deterministic components in both the DGP and the estimated system. As discussed in Ng and Perron (2005), when selecting the lag order, it is important to estimate each candidate model with the same numbers of effective observations. Given the fact that variations in the order k imply different numbers of initial conditions, this means that all candidate models are estimated with $N = T - K_T - 1$ observations, where K_T is the maximal value that defines the range over which k can be a candidate, the lower bound being zero. So the model used for the n-vector y_t is

$$\Delta y_t = \Pi y_{t-1} + \sum_{j=1}^{k} \Gamma_j \Delta y_{t-j} + e_t \qquad \textbf{(6)}$$

for $t = K_T + 2, \ldots, T$, where $e_t \sim i.i.d.\ N(0, \Sigma)$. Under the null hypothesis that the cointegrating rank of the system is r, we have $\Pi = \alpha\beta'$ as defined earlier. To facilitate the derivations, it is useful to adopt a reparameterization of the model. Define $w_t = (\Delta y'_{t-1}, \ldots, \Delta y'_{t-k})'$ and $\Gamma = (\Gamma_1, \ldots, \Gamma_k)$; then (6) is equivalent to

$$\Delta y_t = \Pi y_{t-1} + \Gamma w_t + e_t,$$

https://doi.org/10.1017/S0266466607070284 Published online by Cambridge University Press

where we note that y_{t-1} is I(1), an integrated process of order one, and w_t is I(0) in the sense that its partial sums satisfy a functional central limit theorem (FCLT). In what follows we shall simply refer to these components as the nonstationary and stationary components. Following Anderson (1951), we project y_{t-1} onto the range space of w_t and define y^*_{t-1} and Γ^* by

$$y^*_{t-1} = y_{t-1} - \sum_{s=K_T+2}^{T} y_{s-1} w_s' \left(\sum_{s=K_T+2}^{T} w_s w_s' \right)^{-1} w_t,$$

$$\Gamma^* = \Gamma + \Pi \sum_{t=K_T+2}^{T} y_{t-1} w_t' \left(\sum_{t=K_T+2}^{T} w_t w_t' \right)^{-1},$$

and we can write (6) as

$$\Delta y_t = \Pi y^*_{t-1} + \Gamma^* w_t + e_t,$$

which is useful because the regressors y^*_{t-1} and w_t are orthogonal. Finally, define $\Psi = (\Pi, \Gamma^*)$, $\theta = (\Psi, \Sigma)$, and $x_{t-1} = (y^{*\prime}_{t-1}, w_t')'$; (6) can then be equivalently represented as

$$\Delta y_t = \Psi x_{t-1} + e_t.$$

The aim of an information criterion in this context is to select the order k of the VAR approximation. We shall for now assume that the true order of the preceding model is k_0 and a model of order $k \geq k_0$ is estimated. This is a strong assumption, and it is done to simplify the derivation. The same technique is commonly used in other studies, for example, Linhart and Zucchini (1986, p. 245) and Hurvich and Tsai (1989). We use this assumption for Sections 3.1 and 3.2 when deriving the MIC. In Section 4, we relax it and allow for the case where the approximating family does not include the true model.

Hence, the log-likelihood function conditional on the first $K_T + 1$ observations is

$$l(y;\theta) = \log L(y;\theta) = \frac{-Nn}{2} \log(2\pi) - \frac{N}{2} \log|\Sigma| - \frac{1}{2} \Sigma_{t=K_T+2}^{T} (\Delta y_t - \Psi x_{t-1})' \Sigma^{-1} (\Delta y_t - \Psi x_{t-1})$$

with $\theta = (\Psi, \Sigma)$ the parameters of the model.

3.1. General Observations on the Construction of Akaike's Information Criterion

In this section, we briefly discuss the principles underlying the derivation of the AIC. For a detailed discussion, the readers are referred to Akaike (1973,

https://doi.org/10.1017/S0266466607070284 Published online by Cambridge University Press

1974), Cavanaugh (1997), and Burnham and Anderson (1998, pp. 43–50). See also the paper by Shibata (1980), who shows that the AIC asymptotically minimizes the mean squared error for the estimation of the parameters and also the integrated mean squared error for the estimation of the spectral density when the underlying model is a stationary Gaussian autoregression with infinite order. Pötscher (1989) studies its property when applied to a nonstationary autoregressive model, while assuming the true process is of finite order.

The treatment follows Cavanaugh (1997), with slightly different notations. Suppose we have the observed data $y = (y_1, \ldots, y_T)$ whose log-likelihood function is $l(y;\theta)$. Let θ_0 denote the true values of the parameters.[3] The goal is to select a model, here represented by a specific value of the parameter θ, that is a good approximation to the true process. The choice is based on minimizing the Kullback–Leibler distance $d_T(\theta,\theta_0) = E_0\{l(y;\theta_0) - l(y;\theta)\}$, where E_0 is the expectation of the objective function taken with respect to the true distribution of y. Because the first term is invariant to the choice of the model, minimizing this distance is equivalent to maximizing $E_0\{l(y;\theta)\}$. For any given model, θ is unknown, and a natural choice is to replace it by its maximum likelihood estimate (MLE) $\hat{\theta}(y)$. Also, because we do not wish the result to apply conditional on the particular sample available, we shall consider a process y_F, with the same distribution as y but generated by a different history for the innovations, and shall average over such possible processes. Hence, the aim is to maximize

$$E_0\{E_0\{l(y_F;\theta)\}|_{\theta=\hat{\theta}(y)}\}.$$

Akaike's information criterion is then a quantity, say, *AIC*, whose expectation is twice the function to be minimized (aside from an error of order $o(1)$), that is,

$$E_0 AIC = 2E_0\{E_0\{l(y_F;\theta)\}|_{\theta=\hat{\theta}(y)}\} + o(1).$$

To proceed further, consider the following decomposition of the objective function:

$$\begin{aligned} E_0\{E_0\{l(y_F;\theta)\}|_{\theta=\hat{\theta}(y)}\} = {} & \{E_0 \underbrace{l(y;\theta_0) - E_0 l(y;\hat{\theta}(y))}_{\text{(I)}}\} \\ & + \{E_0\{E_0\{\underbrace{l(y_F;\theta)\}|_{\theta=\hat{\theta}(y)}\} - E_0 l(y_F;\theta_0)}_{\text{(II)}}\} \\ & + \underbrace{E_0 l(y;\hat{\theta}(y))}_{\text{(III)}}. \end{aligned} \tag{7}$$

https://doi.org/10.1017/S0266466607070284 Published online by Cambridge University Press

For term (I), because $k \geq k_0$, we can apply a second-order Taylor expansion around $\hat{\theta}(y)$, with an error term of order $o_p(1)$, that is,

$$
\begin{aligned}
&E_0 l(y;\theta_0) - E_0 l(y;\hat{\theta}(y)) \\
&\quad = -\frac{1}{2} E_0 \left\{ \operatorname{tr}\left(\hat{\Sigma}^{-1}(\hat{\Psi} - \Psi_0) \left(\sum_{t=K_T+2}^{T} x_{t-1} x_{t-1}' \right) (\hat{\Psi} - \Psi_0)' \right) \right\} + o_p(1) \\
&\quad = -\frac{1}{2} E_0 \left\{ \operatorname{tr}\left(\hat{\Sigma}^{-1}(\hat{\Pi} - \Pi_0) \left(\sum_{t=K_T+2}^{T} y_{t-1}^{*} y_{t-1}^{*\prime} \right) (\hat{\Pi} - \Pi_0)' \right) \right\} \\
&\qquad - \frac{1}{2} E_0 \left\{ \operatorname{tr}\left(\hat{\Sigma}^{-1}(\hat{\Gamma}^{*} - \Gamma_0^{*}) \left(\sum_{t=K_T+2}^{T} w_t w_t' \right) (\hat{\Gamma}^{*} - \Gamma_0^{*})' \right) \right\} + o_p(1),
\end{aligned}
$$

where the first-order term equals zero because the score is zero when evaluated at the MLE. The second equality follows because $y_{t-1}^{*\prime}$ and w_t' are orthogonal. We again apply a second-order Taylor expansion to term (II), this time around the true value of the coefficients, so that

$$
\begin{aligned}
&E_0\{E_0\{l(y_F;\theta)\}|_{\theta=\hat{\theta}(y)}\} - E_0 l(y_F;\theta_0) \\
&\quad = -\frac{1}{2} E_0 \left\{ \operatorname{tr}\left(\Sigma_0^{-1}(\hat{\Psi} - \Psi_0) \left(E_0 \sum_{t=K_T+2}^{T} x_{F,t-1} x_{F,t-1}' \right) (\hat{\Psi} - \Psi_0)' \right) \right\} + o_p(1) \\
&\quad = -\frac{1}{2} E_0 \left\{ \operatorname{tr}\left(\Sigma_0^{-1}(\hat{\Pi} - \Pi_0) \left(E_0 \sum_{t=K_T+2}^{T} y_{t-1}^{*} y_{t-1}^{*'} \right) (\hat{\Pi} - \Pi_0)' \right) \right\} \\
&\qquad - \frac{1}{2} E_0 \left\{ \operatorname{tr}\left(\Sigma_0^{-1}(\hat{\Gamma}^{*} - \Gamma_0^{*}) \left(E_0 \sum_{t=K_T+2}^{T} w_t w_t' \right) (\hat{\Gamma}^{*} - \Gamma_0^{*})' \right) \right\} + o_p(1),
\end{aligned}
$$

where again the first-order term is zero because the expected score is evaluated at the true values of the coefficients and the second equality follows because $y_{t-1}^{*\prime}$ and w_t' are orthogonal. The traditional derivation of the AIC proceeds as follows. First, we have the following result:

$$
\begin{aligned}
&\lim_{T\to\infty} E_0 \left\{ \operatorname{tr}\left(\hat{\Sigma}^{-1}(\hat{\Gamma}^{*} - \Gamma_0^{*}) \left(\sum_{t=K_T+2}^{T} w_t w_t' \right) (\hat{\Gamma}^{*} - \Gamma_0^{*})' \right) \right\} \\
&\quad = \lim_{T\to\infty} E_0 \left\{ \operatorname{tr}\left(\Sigma_0^{-1}(\hat{\Gamma}^{*} - \Gamma_0^{*}) \left(E_0 \sum_{t=K_T+2}^{T} w_t w_t' \right) (\hat{\Gamma}^{*} - \Gamma_0^{*})' \right) \right\} = kn^2
\end{aligned}
$$

because each quantity converges to the same chi-square random variable with kn^2 degrees of freedom given that w_t is a stationary process. Second, because $k \geq k_0$,

https://doi.org/10.1017/S0266466607070284 Published online by Cambridge University Press

$$
\begin{aligned}
A &\equiv A_1 + A_2 \\
&\equiv \operatorname{tr}\left(\hat{\Sigma}^{-1}(\hat{\Pi} - \Pi_0)\left(\sum_{t=K_T+2}^{T} y_{t-1}^{*} y_{t-1}^{*\prime}\right)(\hat{\Pi} - \Pi_0)'\right) \\
&\quad + \operatorname{tr}\left(\Sigma_0^{-1}(\hat{\Pi} - \Pi_0)\left(E_0 \sum_{t=K_T+2}^{T} y_{t-1}^{*} y_{t-1}^{*\prime}\right)(\hat{\Pi} - \Pi_0)'\right) \\
&\Rightarrow F_1 + F_2, \qquad \textbf{(8)}
\end{aligned}
$$

say, where $F_1 + F_2$ are different random variables (because, with y_{t-1}^{*} a nonstationary process, the limit of $N^{-2}\sum_{t=K_T+2}^{T} y_{t-1}^{*} y_{t-1}^{*\prime}$ is a random variable, unlike the case with stationary regressors), which are functions of a vector Wiener process and the cointegrating rank. The important thing is that they do not depend on k. Hence, $E_0(F_1 + F_2) = C$, and an asymptotically unbiased estimate of $2E_0\{E_0\{l(y_F;\theta)\}|_{\theta=\hat{\theta}(y)}\}$ (aside from an error of order $o(1)$) is given by

$$2E_0\, l(y;\hat{\theta}(y)) - C - 2kn^2,$$

and a quantity AIC such that $E_0(AIC)$ has the same value (again aside from an error of order $o(1)$) is

$$AIC = 2l(y;\hat{\theta}(y)) - C - 2kn^2, \qquad \textbf{(9)}$$

which reduces to $-N\ln|\hat{\Sigma}| - C - 2kn^2$, and maximizing this quantity is equivalent to minimizing the traditional $AIC = \ln|\hat{\Sigma}| + 2kn^2/N$.

The crucial point to make here is that, even though the term A defined by (8) does not depend on k in large samples, it is highly dependent on this lag order in finite samples, and approximating it by a constant may give a distorted picture of the adequate penalty to be imposed on additional lags. Following the main idea proposed by Ng and Perron (2001), we note that the null hypothesis specifies some structure on the parameters of the matrix Π, in particular its rank, as specified by the number of cointegrating vectors. The idea is then to use this information to replace the constant C by a random variable that, in the limit, has the same expected value as A provided $k \geq k_0$ (up to a constant to be made precise later) and, hence, satisfies the principles underlying the construction of the AIC. The resulting quantity is dependent on k in finite samples in such a way as to better account for the quality of the competing models as approximations to the true process.

3.2. An Alternative Approximation for the Term *A*

We now provide an alternative approximation for the term A that uses the information on the parameters imposed by the null hypothesis of a cointegrating rank r. We start by separating the process y_t^{*} into two parts: a stationary component and a nonstationary noncointegrated component. More specifically,

https://doi.org/10.1017/S0266466607070284 Published online by Cambridge University Press

because the cointegration rank of y_t, and hence that of y_t^*, is r, there exists an invertible matrix P, with $Q = P^{-1}$, such that (see Reinsel, 1993, p. 166) $Q(I_n + \Pi)P = diag[I_{n-r}, \Lambda_r] = J$, where J is the Jordan canonical form. Hence, we have

$$\Pi = P(J - I_n)Q = P_2(\Lambda_r - I_r)Q_2',$$

where $P = [P_1, P_2]$, $Q' = [Q_1, Q_2]$, with P_1 and Q_1 $n \times (n - r)$ matrices, so that $z_{1t}^* = Q_1' y_t^*$ is nonstationary and noncointegrated and $z_{2t}^* = Q_2' y_t^*$ is stationary (again in the sense that it is an I(0) process whose partial sum satisfies a FCLT). Also, $\Pi P_1 = 0$, and, equivalently, if $\Pi = \alpha\beta'$, then $\beta' P_1 = 0$ because α is of full rank. Define the scaling matrix

$$D = \begin{bmatrix} I_{n-r}/N & 0 \\ 0 & I_r/N^{1/2} \end{bmatrix}$$

with the first block corresponding to z_{1t}^* and the second to z_{2t}^*. Then, using these results,

$$A_1 = \mathrm{tr}\left[\hat{\Sigma}^{-1}(\hat{\Pi} - \Pi_0)PD^{-1}\left(D\sum_{t=K_T+2}^{T} Q y_{t-1}^* y_{t-1}^{*\prime} Q' D\right)D^{-1}P'(\hat{\Pi} - \Pi_0)'\right]$$

$$= \mathrm{tr}\left[\hat{\Sigma}^{-1}(\hat{\Pi} - \Pi_0)P_1\left(\sum_{t=K_T+2}^{T} z_{1,t-1}^* z_{1,t-1}^{*\prime}\right)P_1'(\hat{\Pi} - \Pi_0)'\right] + \zeta_{1,rn} + o_p(1),$$

where $\zeta_{1,rn}$ asymptotically has a chi-square distribution with rn degrees of freedom. The last equality follows because the matrix $\sum_{t=K_T+2}^{T} D z_{t-1}^* z_{t-1}^{*\prime} D$ is asymptotically block diagonal. The second block corresponds to the r stationary elements and, hence, has the usual chi-square limit distribution. Consider now the term A_2. Using similar arguments and the fact that the limit of $E_0\{N^{-1}\sum_{t=K_T+2}^{T} z_{2,t-1}^* z_{2,t-1}^{*\prime}\}$ is the same as that of $N^{-1}\sum_{t=K_T+2}^{T} z_{2,t-1}^* z_{2,t-1}^{*\prime}$ (because the series are stationary), we have

$$A_2 = \mathrm{tr}\left[\Sigma_0^{-1}(\hat{\Pi} - \Pi_0)P_1\left(E_0\sum_{t=K_T+2}^{T} z_{1,t-1}^* z_{1,t-1}^{*\prime}\right)P_1'(\hat{\Pi} - \Pi_0)'\right] + \zeta_{1,rn} + o_p(1).$$

Now let

$$\varphi_{rn} = \mathrm{tr}\left[\Sigma_0^{-1}(\hat{\Pi} - \Pi_0)P_1\left(E_0\left(\sum_{t=K_T+2}^{T} z_{1,t-1}^* z_{1,t-1}^{*\prime}\right) - \sum_{t=K_T+2}^{T} z_{1,t-1}^* z_{1,t-1}^{*\prime}\right)\right.$$

$$\left.\times P_1'(\hat{\Pi} - \Pi_0)'\right].$$

The limit of φ_{rn} is a function of Wiener processes and the underlying parameters of the model. The important thing is that it does not depend on k provided $k \geq k_0$ (Reinsel, 1993, p. 167). The term A_2 can then be written as

https://doi.org/10.1017/S0266466607070284 Published online by Cambridge University Press

$$A_2 = \mathrm{tr}\left[\Sigma_0^{-1}(\hat{\Pi} - \Pi_0)P_1\left(\sum_{t=K_T+2}^{T} z^*_{1,t-1}z^{*\prime}_{1,t-1}\right)P_1'(\hat{\Pi} - \Pi_0)'\right] + \zeta_{1,rn} + \varphi_{rn} + o_p(1).$$

Collecting terms, we have

$$A = \mathrm{tr}\left[2\hat{\Sigma}^{-1}(\hat{\Pi} - \Pi_0)P_1\left(\sum_{t=K_T+2}^{T} z^*_{1,t-1}z^{*\prime}_{1,t-1}\right)P_1'(\hat{\Pi} - \Pi_0)'\right] + 2\zeta_{1,rn} + \varphi_{rn} + o_p(1).$$

We now impose the null hypothesis that $\Pi_0 = \alpha_0\beta_0'$, and because these are unknown, we replace them with the MLE obtained assuming the reduced rank structure, denoted $\tilde{\Pi} = \tilde{\alpha}\tilde{\beta}'$. This induces a difference to A whose limit distribution again does not depend on k provided $k \geq k_0$. Hence, a quantity with the same expected value in large samples as A (up to a constant that does not depend on k provided $k \geq k_0$) is

$$\mathrm{tr}\left[2\hat{\Sigma}^{-1}(\hat{\Pi} - \tilde{\Pi})P_1\left(\sum_{t=K_T+2}^{T} z^*_{1,t-1}z^{*\prime}_{1,t-1}\right)P_1'(\hat{\Pi} - \tilde{\Pi})\right].$$

The preceding quantity, which is discussed in detail by Reinsel (1993), is asymptotically equivalent to twice the LR test for testing the null hypothesis of r cointegrating vectors as analyzed by Johansen (1991), which we denote by[4]

$$\tau_k(r) = -N\sum_{j=r+1}^{n}\log(1 - \hat{\lambda}_j).$$

The corresponding MAIC, substituting the constant C in (9) by $2\tau_k(r)$, is then

$$MAIC = \ln|\hat{\Sigma}| + \frac{2(\tau_k(r) + kn^2)}{N}.$$

The extra term in the penalty function of the MAIC is closely related to that derived by Ng and Perron (2001) in the univariate context, and the reasons why it yields improvements are similar. In Ng and Perron (2001), this extra term was related to a unit root test based on an autoregression and involved the sum of the autoregressive coefficients. In their case, the bias in the sum of the autoregressive coefficients is dependent on the order of the autoregression; especially when the MA polynomial has a root close to one. Their simulations show that this bias decreases as the order increases, thereby reducing the overall penalty. Because this bias flattens out as the autoregressive order increases, the overall shape of the penalty function is nonlinear. The same explanation applies here except that, in this multivariate context, the unit root test is replaced by the cointegration test and Π is the equivalent of the sum of the autoregressive

https://doi.org/10.1017/S0266466607070284 Published online by Cambridge University Press

coefficients, being a function of the sum of the coefficient matrices of the VAR polynomial.

More generally, we can consider a class of information criteria of the form

$$MIC = \ln|\hat{\Sigma}| + \frac{C_T(\tau_k(r) + kn^2)}{N}. \tag{10}$$

If $C_T = 2$, we have the MAIC, if $C_T = \ln \ln N$, we may call it the modified Hannan–Quinn criterion (Hannan and Quinn, 1979), and if $C_T = \ln N$, we have a modified BIC. As for the traditional class of information criteria, we can specify a class of MIC with C_T either constant or such that $C_T \to \infty$ and $C_T/T \to 0$ as $T \to \infty$. We will analyze the MIC at this level of generality in the next sections.

The MIC takes a simple form and is very easy to apply. Consider the problem of testing r against more than r cointegrating vectors. First, we need to choose an upper bound K_T, and the following vector autoregression needs to be estimated for $0 \leq k \leq K_T$:

$$\Delta y_t = \Pi y_{t-1} + \sum_{j=1}^{k} \Gamma_j \Delta y_{t-j} + e_t \qquad (t = K_T + 2, \ldots, T), \tag{11}$$

where deterministic components may be included depending on the underlying process. For each case, we need to construct the following two quantities:

$$\tau_k(r) = -N \sum_{j=r+1}^{n} \log(1 - \hat{\lambda}_j); \qquad \hat{\Sigma} = N^{-1} \sum_{t=K_T+2}^{T} \hat{e}_t \hat{e}_t', \tag{12}$$

where $N = T - K_T - 1$ and $\hat{e}_t$ are estimated residuals from (11). Note that a possible choice of the upper bound is $K_T = int(12(T/100)^{1/4})$, which is used by Ng and Perron (2001). Also, when constructing (12) no finite-sample adjustment is applied; the scaling used follows from the likelihood principle we have adopted.[5] Second, we need to substitute the preceding two quantities into (10), and k_{MIC} is chosen as the minimizer. Then, the hypothesis can be tested using k_{MIC} along with (4). In practice, often iterative tests are performed using a series of hypotheses, that is, testing for the cointegrating rank starting with $r = 0$ and moving upward. In such a context, we suggest reselecting the order separately for each null hypothesis. Simulations showed that this yields better performance than selecting the lag order once for all steps of the iterative procedure.

It is important to emphasize that the extra term in the MIC provides only a finite-sample adjustment. Asymptotically the MIC and the standard information criteria described are equivalent. Hence, to provide a theoretical explanation of why the extra term allows significant improvements in the exact size of the cointegration tests would require asymptotic expansions, which in this context are very complex and outside the scope of our analysis. Instead, we shall offer explanations for some special cases of interest using theoretical argu-

https://doi.org/10.1017/S0266466607070284 Published online by Cambridge University Press

ments and simulations. In a first step, we consider how the MIC can offer improvements even in finite-order vector autoregressions (see Section 3.3), a fact overlooked by Ng and Perron (2001). In a second step, we consider the behavior of the MIC when the process is an MA of order one. In this case, we can provide theoretical explanations using a local asymptotic framework. This is done in Section 5 and complemented with simulation results in Section 6.

3.3. Size Improvement in Finite-Order VAR Models

We here consider how the MAIC can offer an improvement over traditional information criteria when the process is a finite-order vector autoregression. We work with the following simple example of a VAR(3) given by

$$A(L)\Delta y_t = \varepsilon_t,$$

$$A(L) = I + 0.2L + 0.1L^2 + 0.1L^3,$$

where y_t is an n-vector with $n = 2$, 3, and 4, $T = 200$, $\varepsilon_t \sim i.i.d.\ N(0, I_n)$ for $t > 0$, and $\varepsilon_t = 0$ for $t \leq 0$. An important feature of this process is that there is fairly weak correlation at lags 2 and 3 but that jointly they account for a substantial portion of the overall correlation because the first coefficient is also fairly small. To show the relative size properties of the MAIC, BIC, and AIC, we generated data according to the models given by equations (2) and (3) (the results for the model in equation (1) are similar and are not reported). All results are obtained using 10,000 replications. The exact sizes for the test with a 5% nominal size are reported in Table 2, part (a). It is clear that the use of the AIC and the BIC results in size distortions that increase with the dimension of the system and the number of deterministic components. On the other hand, the MAIC offers significant improvements, yielding exact sizes close to the 5% nominal level.

To better understand the role of the extra term $\tau_k(r)$ in allowing for tests with better exact sizes, we report in Figure 1 the median (over all replications) of $\tau_k(0) - \tau_0(0)$ as a function of k. There are several features worthy of note. First, the penalty term is decreasing until $k = 3$ (the true value), at which point it reaches a minimum, and there is also a further (less important) increase as k increases beyond 3. This nonlinear effect acts in a way to penalize any value of k different than 3. Second, the extent of the decrease is more important for the model in equation (3) than for the model in equation (2), which means that $\tau_k(r)$ also adapts to the nature of the deterministic components in the estimated system. Third, the decrease is also more important the larger the number of variables n in the system. These last two features explain why the MAIC is able to bring the exact size close to nominal size in all cases even though the size distortions implied by the use of the AIC or BIC are higher the larger the

https://doi.org/10.1017/S0266466607070284 Published online by Cambridge University Press

TABLE 2. Performance of AIC, BIC, and MAIC in the VAR(3) model

		Case B			Case C			
		MAIC	AIC	BIC	MAIC	AIC	BIC	
(a) Exact size of the LR test for a 5% nominal level								
$n = 2$		0.043	0.115	0.227	0.046	0.134	0.278	
$n = 3$		0.056	0.168	0.296	0.070	0.206	0.372	
$n = 4$		0.086	0.251	0.374	0.103	0.305	0.458	
		$k = 0$	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k > 5$
(b) Frequency distribution of the lag order selected (Case B)								
$n = 2$	MAIC	0.048	0.406	0.188	0.271	0.050	0.018	0.018
	AIC	0.186	0.448	0.139	0.166	0.035	0.014	0.012
	BIC	0.886	0.113	0.001	0.000	0.000	0.000	0.000
$n = 3$	MAIC	0.076	0.570	0.170	0.158	0.021	0.003	0.002
	AIC	0.307	0.519	0.086	0.078	0.008	0.002	0.001
	BIC	0.993	0.007	0.000	0.000	0.000	0.000	0.000
$n = 4$	MAIC	0.127	0.671	0.123	0.075	0.004	0.001	0.000
	AIC	0.504	0.435	0.039	0.019	0.002	0.000	0.000
	BIC	1.000	0.000	0.000	0.000	0.000	0.000	0.000
(c) Frequency distribution of the lag order selected (Case C)								
$n = 2$	MAIC	0.037	0.363	0.202	0.302	0.057	0.023	0.017
	AIC	0.178	0.471	0.143	0.153	0.032	0.012	0.011
	BIC	0.888	0.111	0.005	0.000	0.000	0.000	0.000
$n = 3$	MAIC	0.056	0.531	0.179	0.207	0.021	0.005	0.002
	AIC	0.307	0.527	0.082	0.074	0.007	0.002	0.002
	BIC	0.995	0.005	0.000	0.000	0.000	0.000	0.000
$n = 4$	MAIC	0.092	0.652	0.152	0.098	0.005	0.000	0.000
	AIC	0.491	0.440	0.043	0.024	0.002	0.000	0.000
	BIC	1.000	0.000	0.000	0.000	0.000	0.000	0.000

Note: $A(L)\Delta y_t = \varepsilon_t; A(L) = I + 0.2L + 0.1L^2 + 0.1L^3$ with $\varepsilon_t \sim i.i.d.\ N(0, I_n), T = 200$.

number of variables in the system or the number of deterministic components in the estimated regressions.

To see how the presence of the extra penalty term $\tau_k(r)$ affects the lag order selected, Table 2 presents the frequency distributions of the lag order selected by the MAIC, AIC, and BIC. The results show that the BIC selects 0 lag almost 100% of the time. The AIC selects either 0 or 1 lag most of the time. On the other hand, the MAIC selects 0 lags with small probability and higher order lags more often than the BIC or AIC. Note that the fact that the MAIC does not

https://doi.org/10.1017/S0266466607070284 Published online by Cambridge University Press

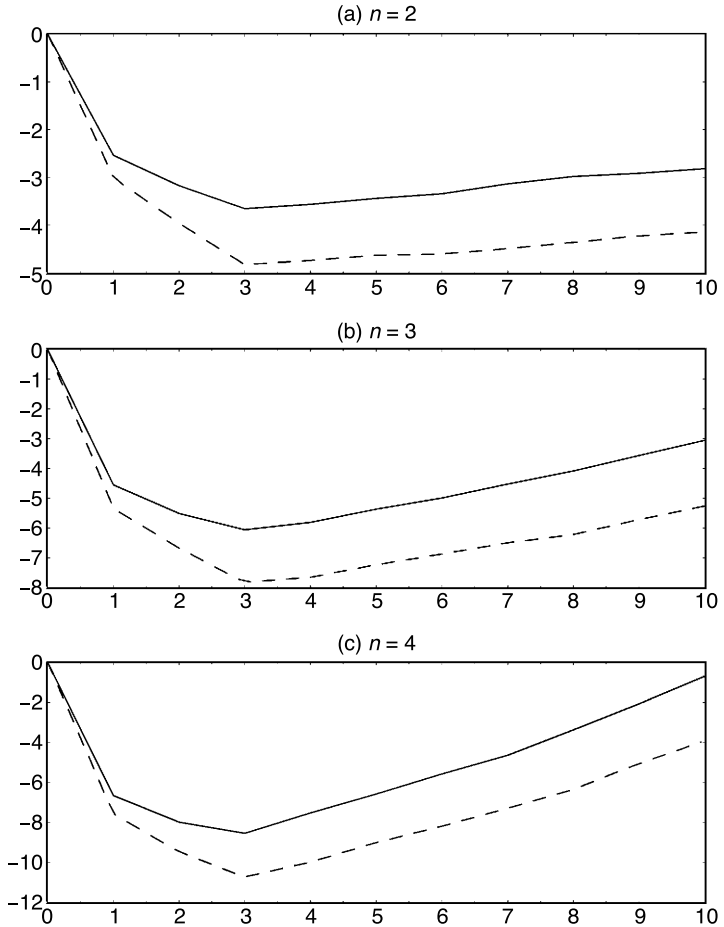


FIGURE 1. $\tau_k(0) - \tau_0(0)$ as a function of k: VAR(3), $T = 200$ (solid line: Case B; dashed line: Case C).

select 3 lags with the highest probability is not a defect and is because the MAIC adapts to the actual correlation in a particular sample. Indeed, unreported simulations showed that the exact size of the cointegration test using the MAIC is as good as when using a fixed number of lags set at the true value 3.

4. LAG ORDER SELECTION IN A GENERAL VAR MODEL WITH REDUCED RANK STRUCTURE

We now study the property of the MIC in a general framework. Following Lütkepohl and Saikkonen (1999), we consider the following DGP:

$$y_{1,t} = Ay_{2,t} + u_{1,t}, \qquad \Delta y_{2,t} = u_{2,t}, \tag{13}$$

https://doi.org/10.1017/S0266466607070284 Published online by Cambridge University Press

where $u_{1,t}$ and $u_{2,t}$ are $r \times 1$ and $(n - r) \times 1$ vectors, respectively, and $u_t = [u'_{1,t}, u'_{2,t}]'$ is a weakly stationary process. The system is assumed to have a triangular structure without loss of generality, because the MIC and other statistics considered later are invariant to a rotation of the system. Also, it is assumed that the system contains no deterministic component, which will be relaxed later. The system may also be written as

$$\Delta y_t = C y_{t-1} + v_t, \tag{14}$$

where $v_t = [v'_{1,t}, v'_{2,t}]'$ is a linear transformation of u_t given by

$$v_t = \begin{bmatrix} I_{n_1} & A \\ 0 & I_{n_2} \end{bmatrix} u_t.$$

We assume that the process $\{v_t\}$ has the following representation:

$$\sum_{j=0}^{\infty} G_j v_{t-j} = \varepsilon_t \tag{15}$$

with $\{G_j\}_{j=0}^{\infty}$ and $\{\varepsilon_t\}_{t=1}^{\infty}$ satisfying the following conditions.

Assumption 1. The matrices $\{G_j\}_{j=0}^{\infty}$ are a sequence of constants satisfying $G_0 = I_n$ and $\sum_{j=1}^{\infty} j^a \|G_j\| < \infty$ for some $a > 1$; $\{\varepsilon_t\}_{t=1}^{\infty}$ is an i.i.d. sequence with $\text{var}(\varepsilon_t) = \Sigma$ and $E(\|\varepsilon_t\|^4) < \infty$.[6]

This assumption imposes mild restrictions on the temporal dependence of v_t, which are satisfied by a large family of processes, for example, all stationary and invertible ARMA processes with i.i.d. innovations. However, it excludes processes with long-range dependence. Note that if $G_j = 0$ for $j > k_0$, then v_t is a finite-order VAR(k_0) process. Assumption 1 also implies that the process y_t can be approximated by the following finite-order vector autoregression:

$$\Delta y_t = \Pi y_{t-1} + \sum_{j=1}^{k} \Gamma_j \Delta y_{t-j} + e_t \tag{16}$$

with

$$e_t = \varepsilon_t - \sum_{j=k+1}^{\infty} G_j v_{t-j}. \tag{17}$$

The difference between ε_t and e_t goes to zero if $k \to \infty$. In other words, the process Δy_t can be approximated arbitrarily well by an AR(k) process by choosing a large enough k.

Because the process is cointegrated of rank r, Π has reduced rank, and we now specify it with the following triangular structure, without loss of generality:

$$\Pi = \Phi \Theta' \tag{18}$$

https://doi.org/10.1017/S0266466607070284 Published online by Cambridge University Press

with Φ an $n \times r$ matrix of full column rank and $\Theta' = [I_r : -A]$. To study the properties of the lag order selected by the MIC, denoted k_{MIC}, and its effect on the limit distribution of the cointegration test, we need to specify a maximum order allowed, K_T. We use the same assumption as in Ng and Perron (2001) and Lütkepohl and Saikkonen (1997):

Assumption 2. The order k_{MIC} is selected from a set of values of k ranging from 0 to K_T, with K_T chosen such that $K_T \to \infty$ and $K_T^3/T \to 0$, as $T \to \infty$.

We now study the property of the MIC when the data are generated by (14). We need the following lemma, which is concerned with the behavior of $\tau_k(r)$ when the lag length k varies between 0 and K_T.

LEMMA 1. *Suppose the data are generated by (14) with cointegrating rank r and let $\tau(r)$ denote the limiting distribution of the LR test when the true order is finite and known. Then, under Assumptions 1 and 2,*

(i) if v_t is a VAR process of order k_0, $\tau_k(r) \to^d \tau(r)$ uniformly in $k_0 \leq k \leq K_T$ and $\tau_k(r) = O_p(1)$ uniformly in $0 \leq k < k_0$;
(ii) if v_t is an infinite-order vector autoregression, $\tau_k(r) \to^d \tau(r)$ holds for any sequence of k satisfying $k \to \infty$ and $k \leq K_T$. Also, $\tau_k(r) = O_p(1)$ uniformly in $0 \leq k \leq M$ for any $M < \infty$.

The next result is concerned with the lag order selected by the MIC when the data are generated by a finite-order VAR process, that is, $G_j = 0$ for $j > k_0$ in (15). It establishes that the MIC has the same properties as the usual information criteria, namely, that the lag order is consistently estimated if $C_T \to \infty$ and $C_T/T \to 0$ as the sample size increases and that when C_T is fixed, the probability of underestimation is zero in large samples.

PROPOSITION 1. *Suppose the data are generated by (14) with v_t a VAR process of order k_0. Also assume the cointegrating rank is r. Then, under Assumptions 1 and 2,*

(i) if C_T is a constant, $\Pr(k_{MIC} \geq k_0) \to 1$, as $T \to \infty$;
(ii) if $C_T \to \infty$ with $C_T/T \to 0$ as $T \to \infty$, $k_{MIC} \xrightarrow{p} k_0$.

The next result is concerned with the lag order selected by the MIC when the process is an infinite-order vector autoregression.

PROPOSITION 2. *Suppose the data are generated by (14) with v_t an infinite-order VAR process. Also assume that the cointegrating rank is r. Then, under Assumptions 1 and 2, $k_{MIC} \to \infty$ in probability, as $T \to \infty$.*

The proof follows directly from the proof of Proposition 1 in the Appendix, and hence the details are omitted. The next result states that the LR test for

https://doi.org/10.1017/S0266466607070284 Published online by Cambridge University Press

cointegration constructed from a vector autoregression whose order is selected by the MIC has the same limit distribution as when the order is finite and known.

PROPOSITION 3. *Suppose the data are generated by (14) and assume that the cointegrating rank is r. Then, under Assumptions 1 and 2, the LR test for cointegration, constructed from a vector autoregression whose order is selected by the MIC, has the same limit distribution as when the order is finite and known.*

For the proof of Proposition 3, the reader is referred to Theorem 5.2 of Lütkepohl and Saikkonen (1999). The details are omitted.

4.1. Model with Deterministic Components

The DGP (14) assumes that the process contains no deterministic component. In practice, it is often important to consider processes with a nonzero mean or a linear trend. More precisely, suppose the data are generated by

$$\begin{aligned} y_{1,t} &= u_0 + Ay_{2,t} + u_{1,t}, \\ \Delta y_{2,t} &= \delta_2 + u_{2,t}, \end{aligned} \tag{19}$$

where $u_{1,t}$ and $u_{2,t}$ are as specified previously. We consider the following three possibilities with respect to the data and the regression system used.

Case 1. The DGP contains an intercept but no trend ($\delta_2 = 0$), and the regression system includes a restricted constant, that is, equation (2).

Case 2. The DGP contains a linear trend ($\delta_2 \neq 0$, i.e., at least one variable is trending), and the regression system includes an unrestricted constant, that is, equation (3).

Case 3. The DGP contains an intercept but no trend ($\delta_2 = 0$), and the regression system includes an unrestricted constant, that is, equation (3).

The following proposition establishes that our results carry through in these cases.

PROPOSITION 4. *If the data are generated by (19) with Cases 1–3 applying, then the results of Propositions 1–3 continue to hold.*

To prove this proposition, the key is to establish the appropriate parameterization of the regression. The proof involves finding the limiting distribution of the inverse of the moment matrix of the regressors. If linear trends are present and Case 2 applies, asymptotically this moment matrix may not have full rank

https://doi.org/10.1017/S0266466607070284 Published online by Cambridge University Press

because of the fact that the linear trends dominate and hence the inverse is not well behaved. For example, consider a bivariate unit root process $y_t = (y_{1,t}, y_{2,t})'$ with $y_{1,t}$ and $y_{2,t}$ having a common linear trend; then the moment matrix of y_t is asymptotically of rank 1, after rescaling. This difficulty can be solved using an equivalent parameterization for the regression (see, e.g., Hamilton, 1994, p. 627), which yields a block diagonal matrix of full rank when suitably rescaled. Here, we outline how this is achieved for Case 2. The detailed proof is in the Appendix. Let

$$y_t = (y_{1,t}, y_{2,t}) = (y_{1,t}^{(1)}, y_{1,t}^{(2)}, \ldots, y_{1,t}^{(r)}, y_{2,t}^{(r+1)}, \ldots, y_{2,t}^{(n)})'$$

and

$$\delta = [(A\delta_2)', \delta_2']' = (\delta_1^{(1)}, \ldots, \delta_1^{(r)}, \delta_2^{(r+1)}, \ldots, \delta_2^{(n)})'$$

and suppose that the elements of $y_{2,t}$ are ordered such that $E(\Delta y_{2,t}^{(n)}) = \delta_2^{(n)} \neq 0$, that is, $y_{2,t}^{(n)}$ contains a linear trend. Then, the fitted values from the regression system (3) are equivalent to those obtained with the following regression system:

$$\Delta y_t = u^* + \Pi^* y_{t-1}^d + \gamma^* y_{2,t-1}^{(n)} + \sum_{j=1}^{k} \Gamma_j \Delta y_{t-j}^d + e_t, \tag{20}$$

where

$$y_{t-1}^d = [y_{1,t}^{(1)} - (\delta_1^{(1)}/\delta_2^{(n)}) y_{2,t}^{(n)}, y_{1,t}^{(2)} - (\delta_1^{(2)}/\delta_2^{(n)}) y_{2,t}^{(n)}, \ldots, y_{2,t}^{(n-1)} - (\delta_2^{(n-1)}/\delta_2^{(n)}) y_{2,t}^{(n)}]',$$

$$\gamma^* = \Pi_n + \Pi_1(\delta_1^{(1)}/\delta_2^{(n)}) + \Pi_2(\delta_1^{(2)}/\delta_2^{(n)}) + \cdots + \Pi_{n-1}(\delta_2^{(n-1)}/\delta_2^{(n)}),$$

and $\Delta y_{t-j}^d = \Delta y_{t-j} - \delta$, $u^* = \delta(\sum_{j=1}^k \Gamma_j)$, $\Pi^* = [\Pi_1, \ldots, \Pi_{n-1}]$. In (20), y_{t-1}^d contains no trend, Δy_{t-j}^d is mean zero, and only $y_{2,t}^{(n)}$ contains a linear trend. Hence, the moment matrix of the regressors is asymptotically block diagonal, and the inverse is $O_p(1)$ when appropriately scaled. With this, the proof proceeds as in Lütkepohl and Saikkonen (1999).

Note that when studying the properties of the cointegration tests when the AIC is used, Lütkepohl and Saikkonen (1999) only consider processes with no deterministic trend and conclude that the analysis with trending variables is a nontrivial task. Our results thereby also establish that their conclusions continue to apply with trending variables. In spite of their asymptotic validity, we will show in the next sections that their performance is substantially inferior to that of the MIC in the context of infinite-order VAR processes.

In practice, to apply the MIC the nature of the deterministic components needs to be specified a priori. This is different from Phillips (1996), where the cointegration rank, lag length, and trend degree are determined jointly and where the approach and motivation differ from the present paper, as discussed earlier.

https://doi.org/10.1017/S0266466607070284 Published online by Cambridge University Press

5. THE BEHAVIOR OF $\tau_k(r)$ UNDER A LOCAL ASYMPTOTIC FRAMEWORK

This section examines the behavior of $\tau_k(r)$ under a local asymptotic framework similar to that used in Ng and Perron (2001), which provides a theoretical explanation for the key difference in the finite-sample properties of lag orders selected using the MIC or a standard information criterion when a negative MA component is present. As discussed earlier, $\tau_k(r)$ will influence the choice of k when there is a large negative MA root, a situation for which cointegration tests such as Johansen's LR test exhibit severe size distortions. To understand the issue of lag length selection in this setting, a useful device is to use an asymptotic framework in which the MA coefficients are parameterized as local to -1. So we consider the following process:

$$\begin{aligned} y_{1,t} &= u_{1,t}, \\ \Delta y_{2,t} &= u_{2,t}, \end{aligned} \tag{21}$$

where $y_{1,t}$ and $y_{2,t}$ are of dimensions r and $n - r$, respectively, and

$$u_{2,t} = \varepsilon_t + \theta_T \varepsilon_{t-1},$$

$$\theta_T = (-1 + T^{-1/2}\bar{c})I_{n-r}$$

with $u_{1,t}$ and ε_t i.i.d. sequences with finite fourth-order moments and $\bar{c} > 0$ is a constant. We also suppose that

$$T^{-1}\begin{bmatrix} \sum_{t=1}^{T} u_{1,t}u_{1,t}' & \sum_{t=1}^{T} u_{1,t}\varepsilon_t' \\ \sum_{t=1}^{T} \varepsilon_t u_{1,t}' & \sum_{t=1}^{T} \varepsilon_t \varepsilon_t' \end{bmatrix} \xrightarrow{p} \Sigma.$$

As T increases to infinity, $\theta_T \to -1$, and $y_{2,t}$ becomes a white noise process. However, $y_{2,t}$ is an integrated process that is nearly white noise in finite samples. Nabeya and Perron (1994) used this asymptotic framework to derive an approximation to the finite-sample distribution of the least-squares estimator in a first-order autoregressive model with negative MA errors and showed this approximation not only to be superior to those obtained from standard asymptotic distributions but also to be quite accurate even for small sample sizes. Perron and Ng (1996) used the same setup to analyze the size and power of various unit root tests based upon autoregressions. Our aim here is to use the same device to obtain useful qualitative features about the relative properties of the information criterion and MIC. The results are presented next.

https://doi.org/10.1017/S0266466607070284 Published online by Cambridge University Press

PROPOSITION 5. *Let the data be generated by (21); then under Assumption 2, as* $T \to \infty$,

(i) for any fixed k, the limiting distribution of $\tau_k(r)$ *is unbounded;*
(ii) for any sequence of k satisfying $k \to \infty$ *as* $T \to \infty$ *and* $k^4/T \to 0$,

$$\frac{(k+1)(k+2)}{N}\tau_k(r) \to^d \operatorname{tr}\left\{(\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}) \times \left[\Sigma_{22}^{1/2}\left(\bar{c}^2 \int_0^1 W(s)W(s)'\,ds\right)\Sigma_{22}^{1/2}\right]^{-1}\right\},$$

where $W(s)$ *is an* $(n - r)$*-dimensional vector Wiener process.*

Proposition 5 extends the result of Theorem 2 in Ng and Perron (2001) to the multivariate case.[7] This proposition, along with Lemma 1, explains why the MIC improves upon the original information criterion. Recall that the MIC uses the LR test (as a function of k) as an additional penalty factor. When more lags are added to the regression and the LR test decreases sharply, this is interpreted as a signal that not enough lags are included. When a large negative MA component is present, for which case the local asymptotic framework provides a better approximation, the local limit of $\tau_k(r)$ is unbounded. The slope of $\tau_k(r)$ is approximately $\partial(N/[(k+1)(k+2)])/\partial k \approx -2N/k^3 \to -\infty$, which implies that $\tau_k(r)$ decreases sharply with k. In other words, the MIC prohibits choosing a small lag length when a strong MA component is present. On the contrary, if a strong MA component is not present, which is covered by Lemma 1, the limiting distribution of $\tau_k(r)$ is bounded, uniformly in k. Hence the role played by $\tau_k(r)$ is less significant. The preceding discussion implies that the MIC is able to yield a large lag order when strong MA components are present, whereas it does not yield an overparameterized model when a strong MA component is absent.

6. SIMULATIONS OF THE FINITE-SAMPLE PROPERTIES

We now conduct simulations to assess the finite-sample properties of the lag order selected by the MAIC and the implied size and power of the LR test for cointegration. In the simulations, the lower bound of the order is always zero and the upper bound is $K_T = int(12(T/100)^{1/4})$. All results are obtained using 10,000 replications. We consider only the case where an unrestricted constant is included in the regression, that is, regression (3). The reason is that the size distortions of the cointegration test are highest in this case, and we will clearly show that even in this worst case our method is able to deliver tests with sizes very close to nominal level (the results for the other cases are similar).

We use the DGP (5) considered in Section 2 and report results for systems of dimensions $n = 2, 3$, and 4. Consider first the simulations pertaining to the lag orders selected by the MAIC, the BIC, and the AIC. The median (over all rep-

https://doi.org/10.1017/S0266466607070284 Published online by Cambridge University Press

lications) of the lag selected is reported in Table 3 for a sample of size $T = 200$. The results clearly show that the MAIC selects a lag order that is greater than that selected by the AIC or BIC, especially when the MA coefficient is large and negative. As expected, when the correlation is small all information criteria agree to select a small order. To investigate the implications of these differences, we now turn to the exact size of the LR cointegration test when a 5% nominal size is used. We present results for $T = 100$ and 200 in Figures 2 and 3, respectively. These show not only that the use of the MAIC permits dramatic improvements over the use of either the AIC or BIC, but also that the exact size is nearly 5% over the entire parameter space.

It is also interesting to investigate the exact size when $r > 0$.[8] To this end, we consider the bivariate processes $(y_{1,t}, y_{2,t})$ generated by

$$y_{1,t} = y_{2,t} + u_{1,t}, \qquad \Delta y_{2,t} = u_{2,t},$$

where $u_{1,t}$ and $u_{2,t}$ satisfy

$$u_{1,t} = \varepsilon_{1,t} + \theta_1 \varepsilon_{1,t-1},$$

$$u_{2,t} = \varepsilon_{2,t} + \theta_2 \varepsilon_{2,t-1},$$

TABLE 3. Lag order selected by different information criteria ($T = 200$)

	$n = 2$			$n = 3$			$n = 4$		
φ	MAIC	AIC	BIC	MAIC	AIC	BIC	MAIC	AIC	BIC
−0.8	6	4	1	6	3	1	5	2	0
−0.7	5	3	2	4	3	1	4	2	1
−0.6	3	3	1	3	2	1	3	2	1
−0.5	3	2	1	2	2	1	2	2	1
−0.4	2	2	1	2	1	1	2	1	0
−0.3	1	1	1	1	1	0	1	1	0
−0.2	1	1	0	1	1	0	1	1	0
−0.1	1	0	0	0	0	0	0	0	0
0.0	0	0	0	0	0	0	0	0	0
0.1	0	0	0	0	0	0	0	0	0
0.2	1	1	0	1	1	0	1	1	0
0.3	1	1	1	1	1	1	1	1	0
0.4	2	2	1	1	2	1	2	1	1
0.5	2	2	1	2	2	1	2	2	1
0.6	3	3	2	2	3	1	2	3	1
0.7	4	4	2	3	4	2	3	3	1
0.8	5	5	3	4	5	2	4	4	2

https://doi.org/10.1017/S0266466607070284 Published online by Cambridge University Press

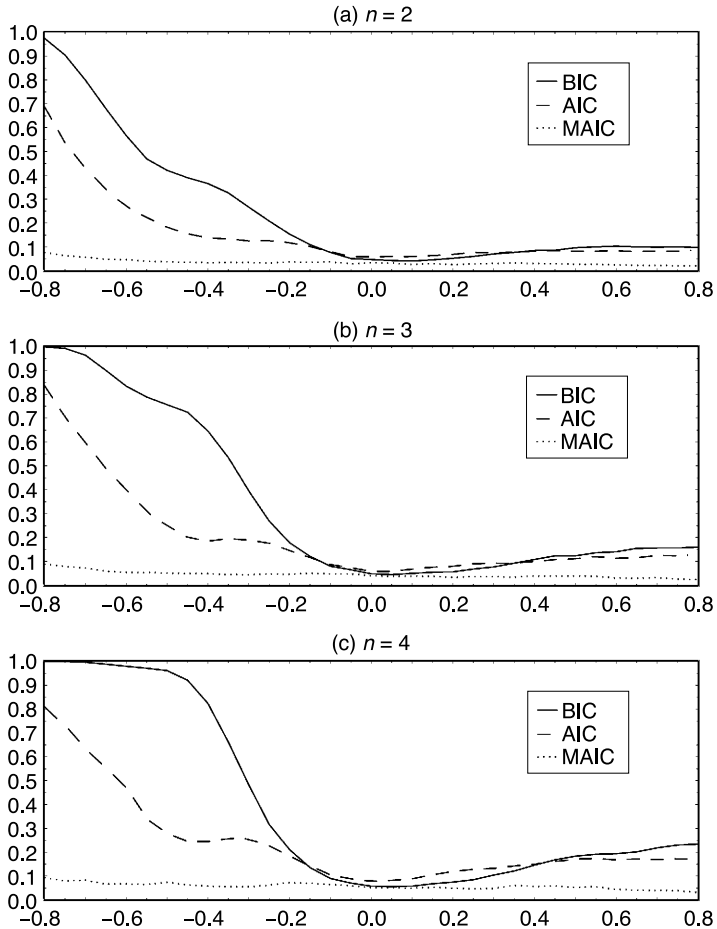


FIGURE 2. Exact size of Johansen's LR test for $r = 0$ (5% nominal level, $T = 100$).

and

$$\begin{pmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{pmatrix} \sim i.i.d.\ N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right).$$

Also, $y_{1,0} = \varepsilon_{1,0} = y_{2,0} = \varepsilon_{2,0} = 0$. We consider the size of the LR test for $r = 1$ at the 5% nominal level with $T = 200$. We are interested in the effects of $(\theta_1, \theta_2, \rho)$ on the size, and accordingly four configurations of parameters are considered:

https://doi.org/10.1017/S0266466607070284 Published online by Cambridge University Press

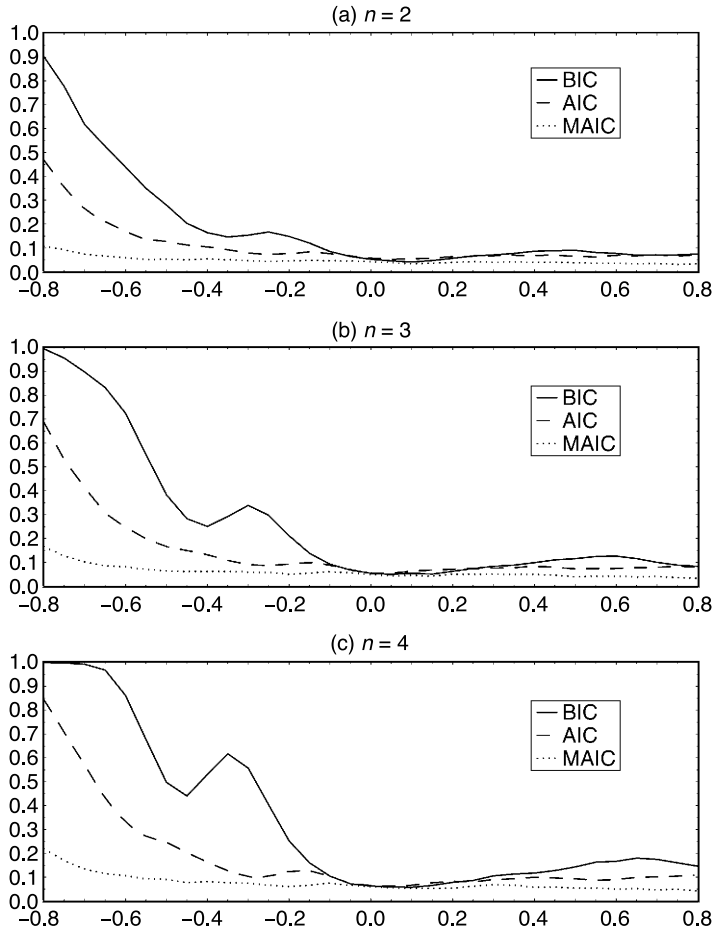


FIGURE 3. Exact size of Johansen's LR test for $r = 0$ (5% nominal level, $T = 200$).

1. $\theta_1 = \rho = 0$, θ_2 varies between -0.8 and 0.8.
2. θ_1 varies between -0.8 and 0.8, $\theta_2 = \rho = 0$.
3. $\theta_1 = \theta_2 = 0$, ρ varies between -0.9 and 0.9.
4. $\theta_1 = 0, \theta_2 = -0.8$, ρ varies between -0.9 and 0.9.

The first parameterization allows an MA component in the stochastic trend, although not in the equilibrium errors. The results are presented in Figure 4a, which again confirms our finding that the MAIC is able to bring the exact size close to the nominal size, whereas using the AIC and the BIC results in serious size distortions. The second parameterization allows an MA component in the equilibrium errors but not in the stochastic trend. Not surprisingly, as shown in

https://doi.org/10.1017/S0266466607070284 Published online by Cambridge University Press

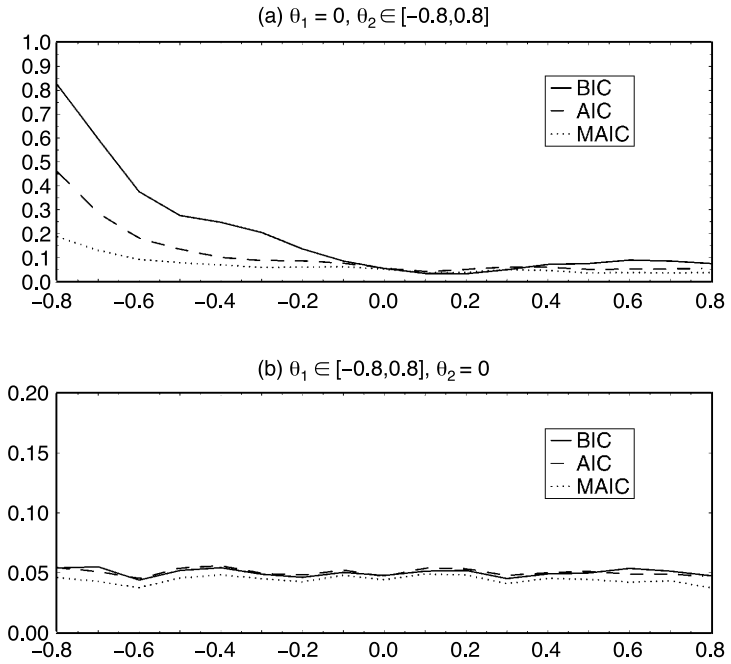


FIGURE 4. Exact size of Johansen's LR test for $r = 1$; uncorrelated shocks (5% nominal level, $T = 200$, $n = 2$).

Figure 4b, size distortions do not occur, and the three criteria perform equally well.

The last two parameterizations examine the effect of cross-equations correlation. To better isolate its effect, we first set $\theta_1 = \theta_2 = 0$. The results are presented in Figure 5a, which shows the information criteria to be robust to cross-equations correlation. We then introduce an MA component, and the picture now changes drastically, as shown in Figure 5b. Interestingly, a strong cross-equation correlation tends to offset the effect of the MA component. This is because when a strong correlation is present, $y_{1,t} - y_{2,t}$ serves as a good proxy to $\varepsilon_{2,t}$. Hence, the lagged values of $y_{1,t}$ and $y_{2,t}$, as a subset of the regressors in the vector autoregression, can effectively reduce the serial correlation, which implies that a long lag is not needed to yield approximately uncorrelated residuals. This is also explained by Proposition 5, which shows that the magnitude of $\tau_k(r)$ is smaller when $\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$ is large. Overall, it is obvious that across all values of $\rho \in [-0.9, 0.9]$, the MAIC performs substantially better than the AIC and the BIC.

It remains to show that this improvement in the exact size of the cointegration test does not come at the expense of a substantial power loss. To that effect, we simulate the size-adjusted power of the test when the lag order is selected

https://doi.org/10.1017/S0266466607070284 Published online by Cambridge University Press

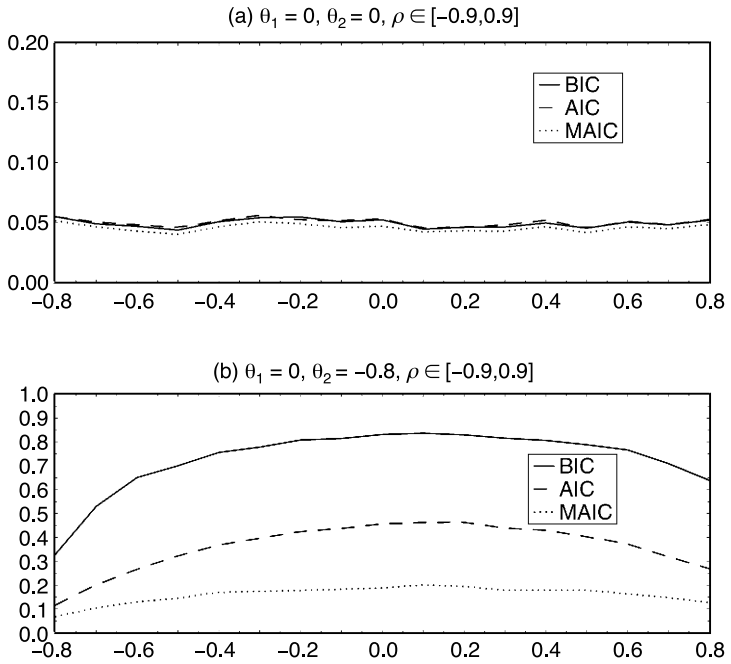


FIGURE 5. Exact size of Johansen's LR test for $r = 1$; correlated shocks (5% nominal level, $T = 200$, $n = 2$).

with the MAIC and the AIC. The DGP used is a simple stationary VARMA(1,1) given by

$$y_t = \rho y_{t-1} + \varepsilon_t + \theta \varepsilon_{t-1}$$

with $\varepsilon_t \sim i.i.d.\ N(0, I_n)$, $y_0 = \varepsilon_0 = 0$, and ρ varying between 1 and 0.7. Hence, under the alternative the cointegrating rank is n. The results are presented in Figures 6–8 for $\theta = 0$, -0.4, and -0.8. They show that there is virtually no power loss when the extent of the correlation is mild, that is, using the AIC or the MAIC leads to tests with the same properties when the extra term in the MAIC is not needed. There is some power loss when the MA component is very negative, which increases as the number of variables in the system increases. This seems a small price to pay to have an exact size close to nominal size. Unreported simulations show that when θ is positive or when the model includes less deterministic components, the power loss is smaller.

6.1. Simulations Calibrated to a Practical Case

Although the previous simulation exercise is useful to assess the general properties of the approximations, it remains to show how the approximations fare

https://doi.org/10.1017/S0266466607070284 Published online by Cambridge University Press

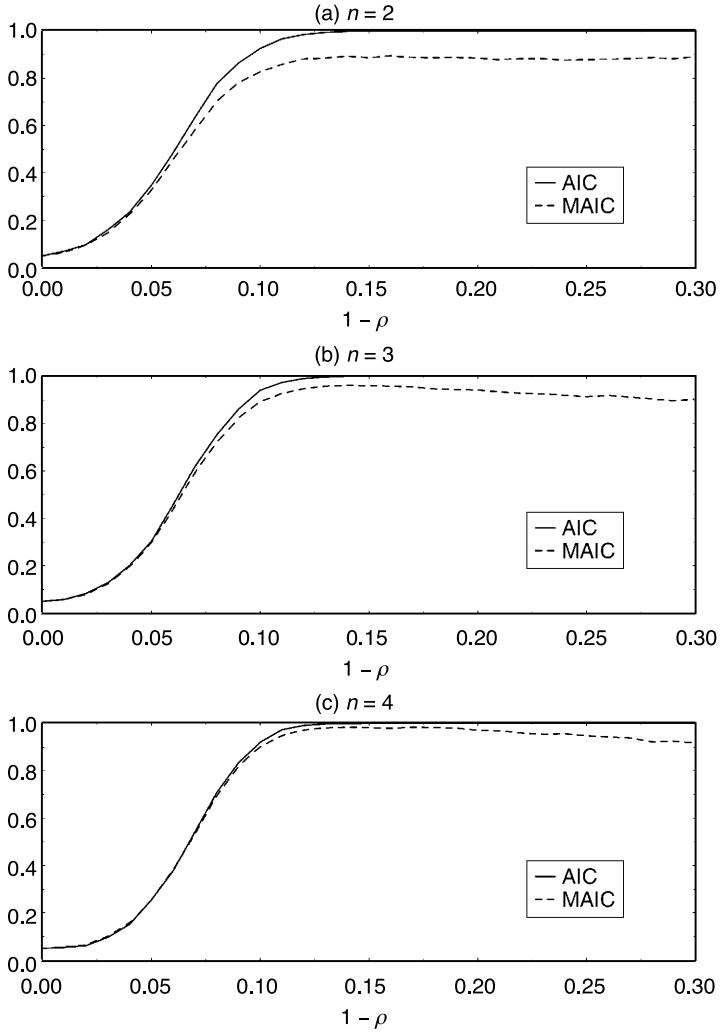


FIGURE 6. Size-adjusted power of Johansen's LR test for $r = 0$ (5% exact size, $T = 200$, $\theta = 0$).

in the context of DGPs that are most likely to occur in practical applications. To that effect, we now consider a simulation design calibrated to the economic application considered by Yap and Reinsel (1995). It is a trivariate system with a VARMA(1,1) structure that is calibrated from a VAR model for three U.S. monthly interest rates.[9] The DGP is

$$\Delta y_t = C y_{t-1} + \varepsilon_t - \Theta_1 \varepsilon_{t-1},$$

https://doi.org/10.1017/S0266466607070284 Published online by Cambridge University Press

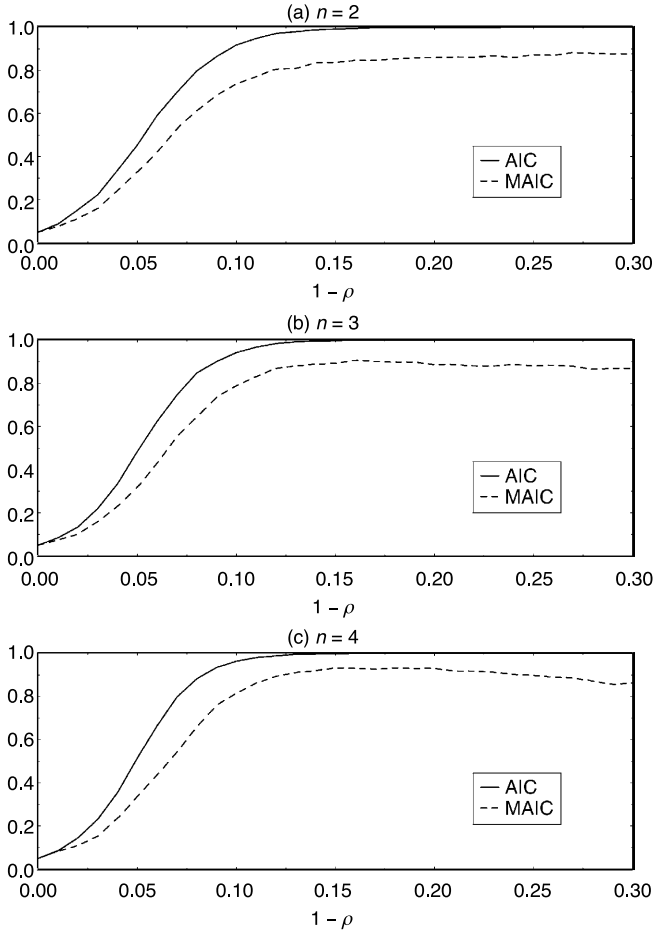


FIGURE 7. Size-adjusted power of Johansen's LR test for $r = 0$ (5% exact size, $T = 200$, $\theta = -0.4$).

where $C = P\,\mathrm{diag}(\lambda_1, \lambda_2, \lambda_3)Q - I_3$, $\Theta_1 = P_\theta\,\mathrm{diag}(-\lambda_\theta, 0.297, -0.202)Q_\theta$, with $Q = P^{-1}, Q_\theta = P_\theta^{-1}$, and

$$Q = \begin{bmatrix} -0.29 & -0.47 & -0.57 \\ -0.01 & -0.85 & 1.00 \\ -0.75 & 1.39 & -0.55 \end{bmatrix}, \qquad P_\theta = \begin{bmatrix} -0.816 & -0.657 & -0.822 \\ -0.624 & -0.785 & 0.566 \\ -0.488 & 0.475 & 0.174 \end{bmatrix},$$

$$\Sigma_\varepsilon = \begin{bmatrix} 0.47 & 0.20 & 0.18 \\ 0.20 & 0.32 & 0.27 \\ 0.18 & 0.27 & 0.30 \end{bmatrix}.$$

https://doi.org/10.1017/S0266466607070284 Published online by Cambridge University Press

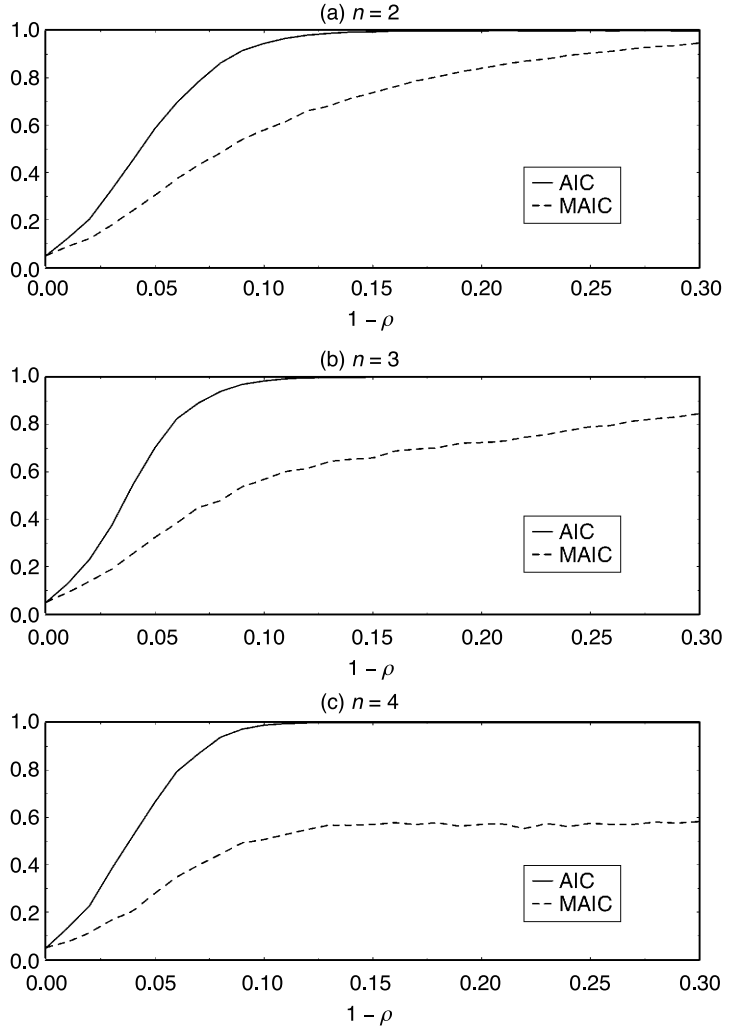


FIGURE 8. Size-adjusted power of Johansen's LR test for $r = 0$ (5% exact size, $T = 200$, $\theta = -0.8$).

Note that the cointegrating rank is determined by the values of $(\lambda_1, \lambda_2, \lambda_3)$. We consider the same three cases as Yap and Reinsel (1995), which are shown in Table 4.

Note that in each case the first row, (a), corresponds to a process satisfying the null hypothesis, and the second and third rows, (b) and (c), correspond to processes satisfying the alternative hypothesis. For each case, we compute

https://doi.org/10.1017/S0266466607070284 Published online by Cambridge University Press

TABLE 4. Three cases for the VARMA(1,1) model

Case	λ_1	λ_2	λ_3	H_0	H_1
M1 (a)	1.0	0.8	0.7	$r = 2$	$r > 2$
M1 (b)	0.9	0.8	0.7	$r = 2$	$r > 2$
M1 (c)	0.8	0.8	0.7	$r = 2$	$r > 2$
M2 (a)	1.0	1.0	0.7	$r = 1$	$r > 1$
M2 (b)	1.0	0.9	0.7	$r = 1$	$r > 1$
M2 (c)	1.0	0.8	0.7	$r = 1$	$r > 1$
M3 (a)	1.0	1.0	1.0	$r = 0$	$r > 0$
M3 (b)	1.0	1.0	0.9	$r = 0$	$r > 0$
M3 (c)	1.0	1.0	0.8	$r = 0$	$r > 0$

Johansen's LR trace statistic from a regression system with a restricted constant, that is, equation (2). The sample size is set to $T = 200$.

To examine the exact size, we let λ_θ vary between -0.7 and 0.7 and compute the empirical rejection frequencies for each λ_θ at the 5% nominal level. The rejection rates, as a function of λ_θ, are presented in Figure 9. Again, the results show that tests based on a VAR approximation with the order selected using the AIC or BIC exhibit severe size distortions. On the other hand, if the lag order is selected with the MAIC, the exact size of the test is close to 5% throughout the parameter space considered. Table 5 presents results for the size-adjusted power. Again, the use of the MAIC implies only a small loss of power relative to the AIC, and the test is more powerful than with the BIC for some configurations.

7. CONCLUSIONS

We considered the properties of cointegration tests constructed from a VAR approximation when the data are generated by a general linear process of weak dependence. We showed that the BIC and AIC are not sufficiently flexible for cointegration tests, in that they lead to the selection of a model that is too parsimonious with the implication that the cointegration tests exhibit substantial size distortions, especially when a negative MA component is present. We extended the work of Ng and Perron (2001), who suggested a modification, the MAIC, which is evidently more robust when there are MA errors negatively correlated and with a finite VAR system when the coefficients of the last few lags are small but have nevertheless an important effect on the overall correlation structure of the series. The idea behind the derivation of the MAIC is to use information in the null hypothesis of the test, as it relates to restrictions on the parameters, to incorporate an extra term in the penalty function. Here, this

https://doi.org/10.1017/S0266466607070284 Published online by Cambridge University Press

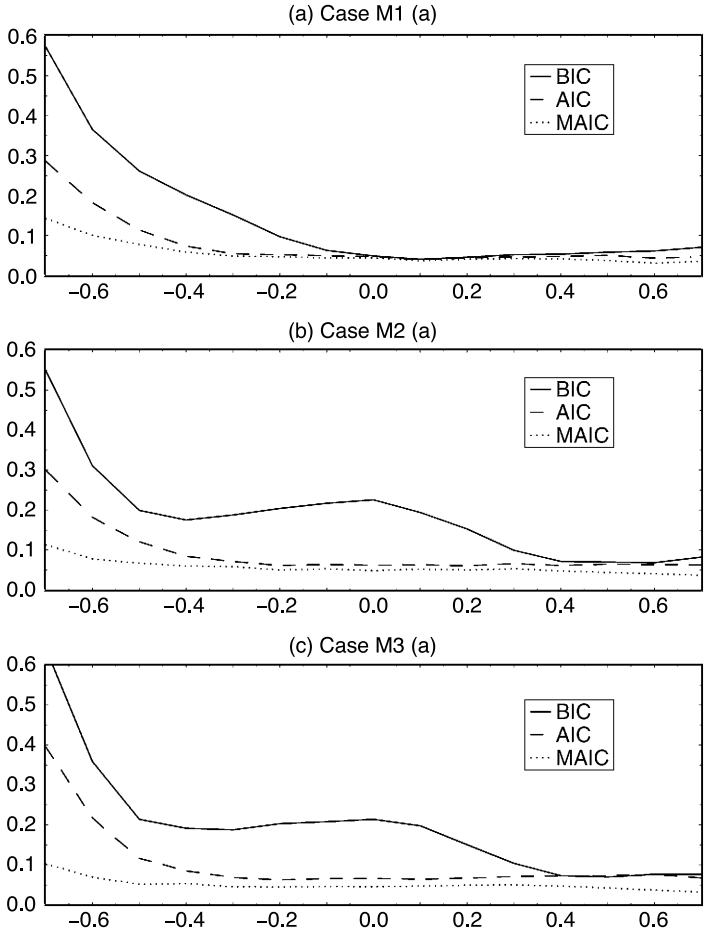


FIGURE 9. Exact size of Johansen's LR test with the empirically calibrated VAR-MA(1,1) model (nominal size 5%, $T = 200$).

extra term is simply the LR test for cointegration itself, and hence the procedure is indeed very easy to use in practice. The use of this MAIC to select the order of the VAR approximation leads to cointegration tests with exact sizes close to nominal size over, basically, the entire parameter space. It should therefore find wide appeal for practical applications. However, it should also be noted that although the suggested information criterion is useful in the context of cointegration tests, as we have shown, it may not be appropriate when selecting the lag order in other contexts, such as selecting a VAR approximation for the purpose of forecasting.

https://doi.org/10.1017/S0266466607070284 Published online by Cambridge University Press

TABLE 5. Size-adjusted power of Johansen's LR test for the VARMA(1,1) model

	Case M1 (b)			Case M1 (c)		
λ_θ	−0.5	0	0.5	−0.5	0	0.5
MAIC	0.714	0.701	0.574	0.911	0.958	0.904
AIC	0.831	0.724	0.663	0.974	0.992	0.949
BIC	0.687	0.781	0.757	0.862	0.999	0.985
	Case M2 (b)			Case M2 (c)		
λ_θ	−0.5	0	0.5	−0.5	0	0.5
MAIC	0.234	0.373	0.315	0.612	0.752	0.685
AIC	0.265	0.449	0.367	0.744	0.936	0.854
BIC	0.227	0.691	0.428	0.577	0.813	0.945
	Case M3 (b)			Case M3 (c)		
λ_θ	−0.5	0	0.5	−0.5	0	0.5
MAIC	0.206	0.299	0.264	0.564	0.678	0.620
AIC	0.231	0.310	0.280	0.640	0.825	0.747
BIC	0.041	0.105	0.300	0.167	0.445	0.813

NOTES

1. Perron and Campbell (1993) consider the case where an unrestricted time trend is also included in the model. Because this case is less often used in practice, we shall not provide explicit results for it, though the MIC to be derived is still applicable.

2. Throughout this paper, we consider the so-called trace test. Note, however, that everything remains valid when considering the maximal eigenvalue test, for which the alternative hypothesis is that the cointegrating rank is $r + 1$.

3. Note that both θ_0 and θ depend on k. This dependence is, however, not explicitly stated to simplify the notation.

4. The appropriate notation should be $\tau_{T,k}(r)$, but we shall omit the dependence on T for simplification.

5. The aim of the finite-sample adjustment is to have a statistic whose asymptotic distribution is a better approximation to the exact distribution. Here, such a concern is not present because the use of the test is to make the penalty factor better approximate the relevant penalty to be applied for the testing procedure.

6. We use the assumptions of Lütkepohl and Saikkonen (1999) because we shall appeal to many of their results. We conjecture that the conditions can be relaxed, in particular to allow for conditional heteroskedasticity, as in, for example, Phillips and Solo (1992) and Chang and Park (2002). Such an extension is, however, outside the scope of our paper.

7. We conjecture, similar to the analysis of Ng and Perron (2001) in the univariate case, that it would be possible to show, using this local asymptotic framework, that setting C_T to a constant is optimal. This requires, however, the derivations of asymptotic expansions, which are very complex in our multivariate framework and beyond the scope of this paper.

https://doi.org/10.1017/S0266466607070284 Published online by Cambridge University Press

8. We thank a referee for suggesting the following simulation experiment.

9. A similar configuration is used by Saikkonen and Luukkonen (1997) and Lütkepohl and Saikkonen (1999).

REFERENCES

Ahn, S.K. & G.C. Reinsel (1990) Estimation for partially non-stationary multivariate autoregressive models. *Journal of the American Statistical Association* 85, 813–823.

Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. In B.N. Petrov & F. Csaki (eds.), *2nd International Symposium on Information Theory*, pp. 267–281. Akademia Kiado.

Akaike, H. (1974) A new look at the statistical model identification. *IEEE Transactions on Automatic Control* AC-10, 716–723.

Anderson, T.W. (1951) Estimating linear restrictions on regression coefficients for multivariate normal distributions. *Annals of Mathematical Statistics* 22, 327–351.

Burnham, K.P. & D.R. Anderson (1998) *Model Selection and Inference: A Practical Information-Theoretic Approach.* Springer.

Cavanaugh, J.E. (1997) Unifying the derivations for the Akaike and corrected Akaike information criteria. *Statistics and Probability Letters* 33, 201–208.

Chang, Y. & J.Y. Park (2002) On the asymptotics of ADF tests for unit roots. *Econometric Reviews* 21, 431–447.

Chao, J.C. & P.C.B. Phillips (1999) Model selection in partially nonstationary vector autoregressive processes with reduced rank structure. *Journal of Econometrics* 91, 227–271.

Cheung, Y.W. & K.S. Lai (1993) Finite-sample sizes of Johansen's likelihood ratio tests for cointegration. *Oxford Bulletin of Economics and Statistics* 55, 313–328.

Engle, R.F. & C.W.J. Granger (1987) Co-integration and error correction: Representation, estimation, and testing. *Econometrica* 55, 251–276.

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

Hamilton, J.D. (1994) *Time Series Analysis*. Princeton University Press.

Hannan, E.J. & B.G. Quinn (1979) The determination of the order of an autoregression. *Journal of the Royal Statistical Society, Series B* 41, 190–195.

Hurvich, C.M. & C.L. Tsai (1989) Regression and time series model selection in small samples. *Biometrika* 76, 297–307.

Johansen, S. (1988) Statistical analysis of cointegration vectors. *Journal of Economic Dynamics and Control* 12, 231–254.

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

Johansen, S. (1996) *Likelihood-Based Inference in Cointegrated Vector Autoregressive Models*. Oxford University Press.

Kuersteiner, G. (2005) Automatic inference for infinite order vector autoregressions. *Econometric Theory* 21, 85–115.

Leeb, H. & B.M. Pötscher (2005) Model selection and inference: Facts and fiction. *Econometric Theory* 21, 21–59.

Lewis, R. & G.C. Reinsel (1985) Prediction of multivariate time series by autoregressive model fitting. *Journal of Multivariate Analysis* 16, 393–411.

Linhart, H. & W. Zucchini (1986) *Information Theory and Statistics*. Dover.

Lütkepohl, H. & P. Saikkonen (1997) Impulse response analysis in infinite order cointegrated vector autoregressive processes. *Journal of Econometrics* 81, 127–157.

Lütkepohl, H. & P. Saikkonen (1999) Order selection in testing for the cointegrating rank of a VAR process. In R.F. Engle & H. White (eds.), *Cointegration, Causality, and Forecasting. A Festschrift in Honour of Clive W.J. Granger*, pp. 168–199. Oxford University Press.

https://doi.org/10.1017/S0266466607070284 Published online by Cambridge University Press

Nabeya, S. & P. Perron (1994) Local asymptotic distribution related to the AR(1) model with dependent errors. *Journal of Econometrics* 62, 229–264.
Ng, S. & P. Perron (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.
Ng, S. & P. Perron (2001) Lag length selection and the construction of unit root tests with good size and power. *Econometrica* 69, 1519–1554.
Ng, S. & P. Perron (2005) A note on the selection of time series models. *Oxford Bulletin of Economics and Statistics* 67, 115–134.
Perron, P. & J.Y. Campbell (1993) A note on Johansen's cointegration procedure when trends are present. *Empirical Economics* 18, 777–789.
Perron, P. & S. Ng (1996) Useful modifications to some unit root tests with dependent errors and their local asymptotic properties. *Review of Economic Studies* 63, 435–463.
Phillips, P.C.B. (1996) Econometric model determination. *Econometrica* 64, 763–812.
Phillips, P.C.B. & W. Ploberger (1994) Posterior odds testing for a unit root with data-based model selection. *Econometric Theory* 10, 774–808.
Phillips, P.C.B. & V. Solo (1992) Asymptotics for linear processes. *Annals of Statistics* 20, 971–1001.
Pötscher, B.M. (1989) Model selection under nonstationarity: Autoregressive models and stochastic linear regression models. *Annals of Statistics* 17, 1257–1274.
Reinsel, G.C. (1993) *Elements of Multivariate Time Series*. Springer-Verlag.
Reinsel, G.C. & S.K. Ahn (1992) Vector AR models with unit roots and reduced rank structure: Estimation, likelihood ratio test, and forecasting. *Journal of Time Series Analysis* 13, 353–375.
Said, S.E. & D.A. Dickey (1984) Testing for unit roots in autoregressive-moving average models of unknown order. *Biometrika* 71, 599–608.
Saikkonen, P. (1992) Estimation and testing of cointegrated systems by an autoregressive approximation. *Econometric Theory* 8, 1–27.
Saikkonen, P. & R. Luukkonen (1997) Testing cointegration in infinite order vector autoregressive processes. *Journal of Econometrics* 81, 93–126.
Schwarz, G. (1978) Estimating the dimension of a model. *Annals of Statistics* 6, 461–464.
Shibata, R. (1980) Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *Annals of Statistics* 8, 147–164.
Wei, C.Z. (1992) On predictive least squares principles. *Annals of Statistics* 20, 1–42.
Yap, S.F. & G.C. Reinsel (1995) Estimation and testing for unit roots in a partially nonstationary vector autoregressive moving average model. *Journal of the American Statistical Association* 90, 253–267.

APPENDIX

Throughout this Appendix, we use the following notation. The euclidean norm of a matrix A is denoted by $\|A\|$, with $\|A\| = [\mathrm{tr}(A'A)]^{1/2}$. The vector induced norm for A is denoted by $\|A\|_1$, with $\|A\|_1 = \sup_{x\neq 0}\|Ax\|/\|x\|$. Note that $\|A\|_1$ is then equivalent to the square of the maximum eigenvalue of $A'A$ and thus $\|A\|_1 \leq \|A\|$. The symbols $\xrightarrow{p}$ and $\Rightarrow$ denote, respectively, convergence in probability and weak convergence under the Skorohod topology, and $O_p(\cdot)$ and $o_p(\cdot)$ are the usual symbols for the order of convergence in probability. For a square matrix A, $|A|$ denotes the determinant of A. Also, $N = T - K_T - 1$ is the effective sample size. Assumptions 1 and 2 are assumed to hold throughout. Lütkepohl and Saikkonen (1999) is denoted as LS in this Appendix due to frequent reference.

Proof of Lemma 1. The first result of part (i) follows immediately from LS (pp. 195–197). To verify the second result, note that for a given k satisfying $0 \leq k < k_0$, Saikkonen and Luukkonen (1997, pp. 113–116, proof of Thm. 3.1) implies $\tau_k(r) = O_p(1)$. (Alternatively, we can follow the argument of Johansen (1996, pp. 158–159), which leads to

https://doi.org/10.1017/S0266466607070284 Published online by Cambridge University Press

the same result.) The order $O_p(1)$ is obviously uniform in k because k_0 is finite. The intuition behind this result is that when $k < k_0$, the eigenvalues corresponding to the cointegrating vectors are still consistently estimated and converge to the true value at rate T. Hence the LR test still has the same order under the null hypothesis, although the limiting distribution is no longer nuisance parameter free.

When the DGP is an infinite-order autoregressive process, then $\tau_k(r) \to^d \tau(r)$, which again follows the proof of LS (pp. 195–197). They also established the uniformity, that is, that the convergence holds for any sequence of k satisfying $k \to \infty$ and $k \leq K_T$. The second result holds following the same argument as in part (i). ■

Proof of Proposition 1. We need the following multivariate extension of Lemma 4.2 of Ng and Perron (1995), due to LS.

LEMMA A.1 (LS, Lem. 5.1). *Suppose the data are generated by (13), satisfying Assumption 1 and having cointegrating rank r. Denote the maximum likelihood estimate of* Σ *from the unrestricted model (16) as* $\hat{\Sigma}_k$ *and that from (16) imposing the restriction* $\Pi y_{t-1} = \Phi u_{1,t-1}$ *as* $\tilde{\Sigma}_k$*; then, uniformly in* $k \leq K_T$*:*

$$\tilde{\Sigma}_k = \hat{\Sigma}_k + o_p\left(\frac{K_T}{T}\right). \tag{A.1}$$

To prove Proposition 1, we first show that $\Pr(k \geq k_0) \to 1$ when either C_T is constant or $C_T \to \infty$ with $C_T/T \to 0$ as $T \to \infty$. First, note that after imposing the restriction $\Pi y_{t-1} = \Phi u_{1,t-1}$, regression (16) only involves stationary regressors. From Lewis and Reinsel (1985), we have that, for any $k < k_0$, $(\tilde{\Sigma}_k - \tilde{\Sigma}_{k_0})$ converges in probability to a positive definite matrix. Hence, using Lemma A.1, $(\hat{\Sigma}_k - \hat{\Sigma}_{k_0})$ also converges in probability to the same positive definite matrix. The difference in the penalty term is, using Lemma 1,

$$\frac{C_T(\tau_{k_0}(r) + k_0 n^2)}{N} - \frac{C_T(\tau_k(r) + kn^2)}{N} = o_p(1).$$

Hence, a $k < k_0$ is never selected asymptotically, and $\Pr(k \geq k_0) \to 1$. To show part (ii), note that $(\ln|\hat{\Sigma}_k| - \ln|\hat{\Sigma}_{k_0}|) = O_p(k/T)$ uniformly in $k_0 < k \leq K_T$ (LS, pp. 193–194). Meanwhile, the penalty term satisfies

$$\begin{aligned}
&\frac{C_T(\tau_k(r) + kn^2)}{N} - \frac{C_T(\tau_{k_0}(r) + k_0 n^2)}{N} \\
&\quad = \frac{C_T((k - k_0)n^2)}{N} + \frac{C_T(\tau_k(r) - \tau_{k_0}(r))}{N} \\
&\quad = \frac{C_T((k - k_0)n^2)}{N} + o_p\left(\frac{C_T}{T}\right)
\end{aligned}$$

with the last equality following from Lemma 1, where it is shown that $\tau_k(r) - \tau_{k_0}(r) = o_p(1)$ uniformly in $k_0 \leq k \leq K_T$. Hence, the penalty term is positive with probability arbitrarily close to 1 for large T and is of higher order than $(\ln|\hat{\Sigma}_k| - \ln|\hat{\Sigma}_{k_0}|)$. This

https://doi.org/10.1017/S0266466607070284 Published online by Cambridge University Press

implies $MIC(k) - MIC(k_0) > 0$ with probability 1 if $C_T \to \infty$, and also $\Pr(k \leq k_0) \to 1$, which proves the result, because we have shown that $\Pr(k \geq k_0) \to 1$ also. ■

Proof of Proposition 4. We first establish some lemmas, which extend Lemmas A1–A4 of LS. The difference is that we allow linear trends, which the results of LS do not cover. Suppose the data are generated by (19) with $\delta_2^{(n)} \neq 0$. We can write (20) as

$$\Delta y_t = \Pi^* \tilde{y}_{t-1}^d + \gamma^* \tilde{y}_{2,t-1}^{(n)} + \sum_{j=1}^{k} \Gamma_j \Delta \tilde{y}_{t-j}^d + e_t,$$

where $\tilde{y}_{t-1}^d$, $\tilde{y}_{2,t-1}^{(n)}$, and $\Delta \tilde{y}_{t-j}^d$ are deviations from their sample means. Rearranging terms,

$$\Delta y_t = \sum_{j=1}^{k} \underline{\Gamma}_j \tilde{u}_{t-1} + \underline{\Gamma}_{k+1,1} \tilde{u}_{1,t-k-1} + \underline{\Pi}_2^* \tilde{y}_{2,t-1}^d + \gamma^* \tilde{y}_{2,t-1}^{(n)} + e_t, \tag{A.2}$$

where $\underline{\Pi}_2^* = 0, \gamma^* = 0$, and $\tilde{y}_{2,t-1}^d$ is an integrated process with zero mean. Define $q_t = [\tilde{u}_{t-1}', \ldots, \tilde{u}_{t-k}', \tilde{u}_{1,t-k-1}']'$, $h_t = [(\tilde{y}_{2,t-1}^d)', \tilde{y}_{2,t-1}^{(n)}]'$, and $p_t = [q_t' : h_t]'$ and also $\Xi = [\underline{\Gamma}_1, \ldots, \underline{\Gamma}_k : \underline{\Gamma}_{k+1,1}]$, $\Upsilon = [\underline{\Pi}_2^* : \gamma^*]$, and $\Lambda = [\Xi : \Upsilon]$. Then, (A.2) can be written as

$$\Delta y_t = \Lambda p_t + e_t \qquad (t = k+2, \ldots, T). \tag{A.3}$$

Let $\hat{\Lambda} = [\hat{\Xi} : \hat{\Upsilon}]$ be the ordinary least squares estimate of Λ from (A.3). Define the scaling matrix $D_T = diag[N^{-1/2} I_{nk+r}, D_{2,T}]$ with $D_{2,T} = diag[N^{-1} I_{n-r-1}, N^{-3/2}]$. Finally, define the moment matrices

$$\hat{R} = D_T \sum_{t=k+2}^{T} p_t p_t' D_T,$$

$$R = diag\left[\Gamma_{qq} : D_{2,T} \sum_{t=k+2}^{T} h_t h_t' D_{2,T} \right],$$

where $\Gamma_{qq} = E(q_t q_t')$, a nonrandom positive definite matrix. We use the partitions $\hat{R}^{-1} = [\hat{R}^{ij}]_{i,j=1,2}$ and $\hat{R}^{22} = [\hat{R}_{ij}^{22}]_{i,j=1,2}$. The following lemma characterizes the rate of convergence of the matrix $\hat{R}^{-1}$ to R^{-1} and extends Lemma A1 of LS to the case with a linear trend.

LEMMA A.2. *Suppose y_t is generated by (19) with $\delta_2^{(n)} \neq 0$; then, as $T \to \infty$, uniformly in $k \leq K_T$,*

(i) $\|\hat{R}^{11}\|_1 = O_p(1)$;
(ii) $\|\hat{R}^{11} - \Gamma_{qq}^{-1}\|_1 = O_p(K_T / N^{1/2})$;
(iii) $\|\hat{R}^{22} - (D_{2,T} \sum_{t=k+2}^{T} h_t h_t' D_{2,T})^{-1}\|_1 = O_p(K_T / N)$;
(iv) $\|\hat{R}^{12}\|_1 = O_p((K_T / N)^{1/2})$;
(v) $\|\hat{R}^{-1} - R^{-1}\|_1 = O_p(K_T / N^{1/2})$.

https://doi.org/10.1017/S0266466607070284 Published online by Cambridge University Press

Proof of Lemma A.2. We first show that the results hold for a given $k \leq K_T$. For part (i),

$$\hat{R}^{11} = \left(N^{-1} \sum_{t=k+2}^{T} q_t q_t' - N^{-1} \sum_{t=k+2}^{T} q_t h_t' D_{2,T} \left(D_{2,T} \sum_{t=k+2}^{T} h_t h_t' D_{2,T} \right)^{-1} \sum_{t=k+2}^{T} D_{2,T} h_t q_t' \right)^{-1}.$$

The first term is such that $N^{-1} \sum_{t=k+2}^{T} q_t q_t' \xrightarrow{p} \Gamma_{qq}$ using standard arguments as in Said and Dickey (1984) and the fact that the mean correction does not affect the order. For the second term,

$$\begin{aligned} \left\| N^{-1/2} \sum_{t=k+2}^{T} q_t h_t' D_{2,T} \right\| &= \left\| N^{-3/2} \sum_{t=k+2}^{T} q_t (\tilde{y}_{2,t-1}^{d})' : N^{-2} \sum_{t=k+2}^{T} q_t \tilde{y}_{2,t-1}^{(n)} \right\| \\ &\leq \left\| N^{-3/2} \sum_{t=k+2}^{T} q_t (\tilde{y}_{2,t-1}^{d})' \right\| + \left\| N^{-2} \sum_{t=k+2}^{T} q_t \tilde{y}_{2,t-1}^{(n)} \right\| \\ &= O_p\left(\left(\frac{K_T}{N} \right)^{1/2} \right) \end{aligned}$$

follows from the property of integrated processes and the fact that $\tilde{y}_{2,t-1}^{(n)}$ is asymptotically equivalent to a linear trend. Also, $\|(D_{2,T} \sum_{t=k+2}^{T} h_t h_t' D_{2,T})^{-1}\| = O_p(1)$, and hence the second term is of order $O_p((K_T/N)) = o_p(1)$, which implies $\|\hat{R}^{11}\|_1 = O_p(1)$. For part (ii),

$$\begin{aligned} &\|(\hat{R}^{11})^{-1} - \Gamma_{qq}\|_1 \\ &\quad \leq \left\| \left((\hat{R}^{11})^{-1} - N^{-1} \sum_{t=k+2}^{T} q_t q_t' \right) + \left(N^{-1} \sum_{t=k+2}^{T} q_t q_t' - \Gamma_{qq} \right) \right\| \\ &\quad \leq \left\| N^{-1} \sum_{t=k+2}^{T} q_t h_t' D_{2,T} \left(D_{2,T} \sum_{t=k+2}^{T} h_t h_t' D_{2,T} \right)^{-1} \sum_{t=k+2}^{T} D_{2,T} h_t q_t' \right\| \\ &\qquad + \left\| \left(N^{-1} \sum_{t=k+2}^{T} q_t q_t' - \Gamma_{qq} \right) \right\| \\ &\quad = O_p\left(\frac{K_T}{N} \right) + O_p\left(\frac{K_T}{N^{1/2}} \right) = O_p\left(\frac{K_T}{N^{1/2}} \right). \end{aligned}$$

Hence,

$$\|\hat{R}^{11} - \Gamma_{qq}^{-1}\|_1 \leq \|\hat{R}^{11}\|_1 * \|(\hat{R}^{11})^{-1} - \Gamma_{qq}\|_1 * \|\Gamma_{qq}^{-1}\|_1 = O_p\left(\frac{K_T}{N^{1/2}} \right).$$

Part (iii) can be proved by changing the role of q_t and h_t and using similar arguments. For part (iv),

$$\begin{aligned} \|\hat{R}^{12}\|_1 &= \left\| \left(N^{-1} \sum_{t=k+2}^{T} q_t q_t' \right)^{-1} \left(N^{-1/2} \sum_{t=k+2}^{T} q_t h_t' D_{2,T} \right) \hat{R}^{22} \right\|_1 \\ &\leq \left\| \left(N^{-1} \sum_{t=k+2}^{T} q_t q_t' \right)^{-1} \right\|_1 * \left\| \left(N^{-1/2} \sum_{t=k+2}^{T} q_t h_t' D_{2,T} \right) \right\|_1 * \|\hat{R}^{22}\|_1 \\ &= O_p(1) * O_p\left(\left(\frac{K_T}{N} \right)^{1/2} \right) * O_p(1) = O_p\left(\left(\frac{K_T}{N} \right)^{1/2} \right). \end{aligned}$$

https://doi.org/10.1017/S0266466607070284 Published online by Cambridge University Press

Part (v) follows automatically from parts (ii)–(iv). The remaining steps to establish that the results hold uniformly in $k = O(K_T)$ follow straightforwardly as in LS (p. 187). ■

The next two lemmas extend Lemmas A2–A4 of LS to allow a linear trend in the DGP with variables replaced by deviations from their sample means. The intent is to show that we can impose the zero restrictions on the coefficients of the nonstationary regressors, h_t, without affecting the asymptotic results. This essentially reduces the problem to that of a regression with only mean-zero stationary regressors, for which results have been established using a predictive regression framework, for example, Lewis and Reinsel (1985). We first need to define residuals from a predictive regression. Rewrite (A.3) as

$$\Delta y_t = \Xi^p q_t + \varepsilon_t^p \qquad (t = k+2, \ldots, T), \tag{A.4}$$

where

$$\Xi^p = E(\Delta y_t q_t')E(q_t q_t')^{-1} = \Gamma_{\Delta q}\Gamma_{qq}^{-1}$$

and ε_t^p is the one-step-ahead prediction error of the best linear predictor of Δy_t based on q_t. Because $\Upsilon = 0$, (A.4) can be augmented as

$$\Delta y_t = \Lambda^p p_t + \varepsilon_t^p \qquad (t = k+2, \ldots, T),$$

where $\Lambda^p = [\Xi^p, \Upsilon] = [\Xi^p, 0]$. Denote the estimates of coefficients from (A.4) as $\tilde{\Xi}^p$ and the residual as $\tilde{\varepsilon}_t^p$. Also, it is convenient to define (refer to (17))

$$e_{1t} = -\sum_{j=k+1}^{\infty} G_j v_{t-j},$$

and we have the following identity:

$$\varepsilon_t^p - \varepsilon_t = e_{1t} - (\Xi^p - \Xi)q_t. \tag{A.5}$$

The next lemma extends Lemmas A2 and A3 of LS.

LEMMA A.3. *Suppose y_t is generated by (19) with $\delta_2^{(n)} \neq 0$; then, as $T \to \infty$, we have, uniformly in $k \leq K_T$,*

(i) $\|\sum_{t=k+2}^{T} e_{1t} h_t' D_{2,T}\| = o_p(k^{-1})$;
(ii) $\|N^{-1}\sum_{t=k+2}^{T} e_{1t} q_t'\| = o_p(k^{-1/2})$; $\|N^{-1}\sum_{t=k+2}^{T} \varepsilon_t^p q_t'\| = O_p(K_T/N^{1/2})$;
(iii) $\|\sum_{t=k+2}^{T} \varepsilon_t^p h_t' D_{2,T} - \sum_{t=k+2}^{T} \varepsilon_t h_t' D_{2,T}\| = (\sum_{j=k+2}^{T} j^a \|G_j\|)O_p(1) + o_p(k^{-1})$; *in particular this term is $o_p(1)$ if $k \to \infty$.*

Proof of Lemma A.3. We first show the results hold for a given $k \leq K_T$. For part (i),

$$\left\| \sum_{t=k+2}^{T} e_{1t} h_t' D_{2,T} \right\| \leq \left\| N^{-1} \sum_{t=k+2}^{T} e_{1t} (\tilde{y}_{2,t-1}^d)' \right\| + \left\| N^{-3/2} \sum_{t=k+2}^{T} e_{1t} \tilde{y}_{2,t-1}^{(n)\prime} \right\|. \tag{A.6}$$

https://doi.org/10.1017/S0266466607070284 Published online by Cambridge University Press

For the first term, let $\bar{y}^d_{2,-1}$ denote the sample mean of $y^d_{2,t-1}$; then

$$\begin{aligned}
&\left\| N^{-1} \sum_{t=k+2}^{T} e_{1t} (\tilde{y}^d_{2,t-1})' \right\| \\
&\quad = \left\| N^{-1} \sum_{t=k+2}^{T} e_{1t} (y^d_{2,t-1})' - \left(N^{-1/2} \sum_{t=k+2}^{T} e_{1t} \right) (N^{-1/2} (\bar{y}^d_{2,-1})') \right\| \\
&\quad \leq \left\| N^{-1} \sum_{t=k+2}^{T} e_{1t} (y^d_{2,t-1})' \right\| + \left\| \left(N^{-1/2} \sum_{t=k+2}^{T} e_{1t} \right) \right\| * \| N^{-1/2} (\bar{y}^d_{2,-1})' \| \\
&\quad = o_p(k^{-1}) + o_p(k^{-1}) * O_p(1) = o_p(k^{-1}),
\end{aligned}$$

where $\| N^{-1} \sum_{t=k+2}^{T} e_{1t} (y^d_{2,t-1})' \| = o_p(k^{-1})$ is obtained using the definition of $(y^d_{2,t-1})'$ (refer to (19)) and proceeding as in LS (pp. 187–188). Note that a similar result is also obtained by Said and Dickey (1984, p. 602). The result $\| (N^{-1/2} \sum_{t=k+2}^{T} e_{1t}) \| = o_p(k^{-1})$ follows by computing its variance, and $\| N^{-1/2} (\bar{y}^d_{2,-1})' \| = O_p(1)$ is obtained by applying the FCLT. Similarly, we can show that the second term in (A.6) is $o_p(1/k)$ using the fact that $\tilde{y}^{(n)\prime}_{2,t-1}$ behaves asymptotically as a linear trend. For part (ii), the first result follows from standard arguments as in Lewis and Reinsel (1985, p. 397), and the second follows from LS. The only difference is to note that the mean correction has no effect on the order of convergence. For part (iii), using (A.5), we have

$$\begin{aligned}
&\left\| \sum_{t=k+2}^{T} \varepsilon_t^p h_t' D_{2,T} - \sum_{t=k+2}^{T} \varepsilon_t h_t' D_{2,T} \right\| \\
&\quad = \left\| \sum_{t=k+2}^{T} e_{1t} h_t' D_{2,T} - \sum_{t=k+2}^{T} (\Xi^p - \Xi) q_t h_t' D_{2,T} \right\| \\
&\quad \leq \left\| \sum_{t=k+2}^{T} e_{1t} h_t' D_{2,T} \right\| + \left\| \sum_{t=k+2}^{T} (\Xi^p - \Xi) q_t h_t' D_{2,T} \right\|.
\end{aligned}$$

The first term is $o_p(1/k)$ by part (i), and for the second term

$$\begin{aligned}
&\left\| \sum_{t=k+2}^{T} (\Xi^p - \Xi) q_t h_t' D_{2,T} \right\| \\
&\quad \leq \left\| N^{-1} \sum_{t=k+2}^{T} (\Xi^p - \Xi) q_t (\tilde{y}^d_{2,t-1})' \right\| + \left\| N^{-3/2} \sum_{t=k+2}^{T} (\Xi^p - \Xi) q_t \tilde{y}^{(n)\prime}_{2,t-1} \right\|.
\end{aligned}$$

Using the result of LS (p. 191), $k^a \| \Xi^p - \Xi \| \leq c k^a \sum_{j=k+2}^{T} \| G_j \|$ for some $c < \infty$, so that

$$\begin{aligned}
&\left\| N^{-1} \sum_{t=k+2}^{T} (\Xi^p - \Xi) q_t (\tilde{y}^d_{2,t-1})' \right\| \\
&\quad \leq \left(c k^a \sum_{j=k+2}^{T} \| G_j \| \right) * \left\| N^{-1} k^{-a} \sum_{t=k+2}^{T} q_t (\tilde{y}^d_{2,t-1})' \right\| = \left(c k^a \sum_{j=k+2}^{T} \| G_j \| \right) * O_p(1)
\end{aligned}$$

https://doi.org/10.1017/S0266466607070284 Published online by Cambridge University Press

because the vector $N^{-1}\sum_{t=k+2}^{T} q_t(\tilde{y}_{2,t-1}^{d})'$ is of dimension $O(k)$ and each is $O_p(1)$. Also,

$$\left\| N^{-3/2} \sum_{t=k+2}^{T} (\Xi^p - \Xi) q_t \tilde{y}_{2,t-1}^{(n)\prime} \right\| = \left(ck^a \sum_{j=k+2}^{T} \|G_j\| \right) * O_p(1)$$

by the same argument and the fact that $\tilde{y}_{2,t-1}^{(n)\prime}$ is dominated by the linear trend. Again, the completion of the proof to show that these results hold uniformly in $k \leq K_T$ follows using the same arguments as in LS (p. 190). ■

The following lemma pertains to the coefficient estimates from regression (A.2), extending Lemma A4 of LS. (Recall that $\tilde{\Xi}^p$ is the parameter estimate in the predictive regression (A.4).)

LEMMA A.4. *Suppose that y_t is generated by (19) with $\delta_2^{(n)} \neq 0$; then, as $T \to \infty$, we have, uniformly in $k \leq K_T$,*

(i) $\|\hat{\underline{\Pi}}_2^*\| = O_p(N^{-1})$; $\|\hat{\gamma}^*\| = O_p(N^{-3/2})$;
(ii) $\|\hat{\Xi} - \tilde{\Xi}^p\| = o_p(K_T/N)$; $\|\tilde{\Xi}^p - \Xi^p\| = o_p(K_T/N^{1/2})$.

Proof of Lemma A.4. For part (i),

$$\hat{\Upsilon} D_{2,T}^{-1} = N^{-1/2} \sum_{t=k+2}^{T} \varepsilon_t^p q_t' \hat{R}^{12} + \left[N^{-1} \sum_{t=k+2}^{T} \varepsilon_t^p (\tilde{y}_{2,t-1}^{d})' : N^{-3/2} \sum_{t=k+2}^{T} \varepsilon_t^p \tilde{y}_{2,t-1}^{(n)\prime} \right] \hat{R}^{22}.$$

We have $\|N^{-1/2}\sum_{t=k+2}^{T} \varepsilon_t^p q_t'\| = O_p(K_T)$ by Lemma A.3(ii), and $\|\hat{R}^{12}\| = O_p((K_T/N)^{1/2})$; hence the first term is $o_p(1)$. Also, $\|N^{-1}\sum_{t=k+2}^{T} \varepsilon_t^p (\tilde{y}_{2,t-1}^{d})'\| = O_p(1)$, $\|N^{-3/2}\sum_{t=k+2}^{T} \varepsilon_t^p \tilde{y}_{2,t-1}^{(n)\prime}\| = O_p(1)$ by Lemma A.3(iii), and $\hat{R}^{22} = O_p(1)$ by Lemma A.1(iii). Hence the second term is $O_p(1)$. For part (ii), the second result follows directly from LS (p. 192). For the first,

$$\begin{aligned}
\|\hat{\Xi} - \tilde{\Xi}^p\| &= \|[\hat{\Xi} - \Xi^p] - [\tilde{\Xi}^p - \Xi^p]\| \\
&\leq \left\| N^{-1} \sum_{t=k+2}^{T} \varepsilon_t^p q_t' \left(\hat{R}^{11} - \left(N^{-1} \sum_{t=k+2}^{T} q_t' q_t \right)^{-1} \right) \right\| \\
&\quad + \left\| N^{-1/2} \sum_{t=k+2}^{T} \varepsilon_t^p h_t' D_{2,T} \hat{R}^{21} \right\|.
\end{aligned}$$

The first term is $o_p(K_T/N)$ from LS (p. 192), whereas

$$\begin{aligned}
&\left\| N^{-1/2} \sum_{t=k+2}^{T} \varepsilon_t^p h_t' D_{2,T} \hat{R}^{21} \right\| \\
&\quad \leq \left\| N^{-1/2} \sum_{t=k+2}^{T} \varepsilon_t^p h_t' D_{2,T} \right\| * \|\hat{R}^{21}\|_1 = O_p(N^{-1/2}) * O_p\left(\left(\frac{K_T}{N} \right)^{1/2} \right) = o_p\left(\frac{K_T}{N} \right),
\end{aligned}$$

which completes the proof. ■

We are now in a position to prove Proposition 4. We only provide a proof for the case where the process contains a linear trend. The proof is similar to that in LS; hence

https://doi.org/10.1017/S0266466607070284 Published online by Cambridge University Press

we only outline the main differences and omit details. We first need to show that Lemma A.1 still holds, which can be shown using arguments as in LS (pp. 192–193), with the term corresponding to $y_{2,t}$ replaced by h_t. Once this is done, the same proof as that of Proposition 1 can be used to show that the results of Propositions 1 and 2 hold when a linear trend is present. Then, with the preceding results, when the order is chosen by the MIC, we have

(a) $\|\sum_{t=k_{MIC}+2}^{T} \varepsilon_t^p h_t' D_{2,T} - \sum_{t=k_{MIC}+2}^{T} \varepsilon_t h_t' D_{2,T}\| = o_p(1)$;
(b) $\hat{\Upsilon} D_{2,T}^{-1} = \sum_{t=K_T+2}^{T} \varepsilon_t h_t' D_{2,T} (D_{2,T} \sum_{t=k_{MIC}+2}^{T} h_t h_t' D_{2,T})^{-1} + o_p(1)$;
(c) $\hat{\Phi} = \Phi_\infty + o_p(1)$, where Φ is defined in (18) and $\Phi_\infty = -(\sum_{j=0}^{\infty} G_j)J$ with $J = [I_r, 0]'$;
(d) $\hat{\Sigma} = \Sigma + o_p(1)$.

These results can be proved by proceeding as in LS (pp. 194–197). Then, following the argument of Saikkonen and Luukkonen (1997, proof of Thm. 3.1), we can show that the LR test for r cointegrating vectors is asymptotically equivalent to the sum of the eigenvalues obtained as the solutions to the following system:

$$\left| D_{2,T}^{-1} \hat{\Upsilon}' \hat{\Sigma}^{-1} \hat{\Upsilon} D_{2,T}^{-1} - D_{2,T}^{-1} \hat{\Upsilon}' \hat{\Sigma}^{-1} \hat{\Phi} (\hat{\Phi}' \hat{\Sigma}^{-1} \hat{\Phi})^{-1} \hat{\Phi}' \hat{\Sigma}^{-1} \hat{\Upsilon} D_{2,T}^{-1} - \lambda \left(D_{2,T} \sum_{t=k_{MIC}+2}^{T} h_t h_t' D_{2,T} \right)^{-1} \right| = 0.$$

From results (a)–(d) asymptotically the solutions of the preceding problem coincide with those of

$$\left| D_{2,T}^{-1} \hat{\Upsilon}' \Sigma^{-1} \hat{\Upsilon} D_{2,T}^{-1} - D_{2,T}^{-1} \hat{\Upsilon}' \Sigma^{-1} \Phi (\Phi' \Sigma^{-1} \Phi)^{-1} \Phi' \Sigma^{-1} \hat{\Upsilon} D_{2,T}^{-1} - \lambda \left(D_{2,T} \sum_{t=K_T+2}^{T} h_t h_t' D_{2,T} \right)^{-1} \right| = 0.$$

Hence, using the MIC leads to a cointegration test with the same asymptotic distribution as in the case with a finite and known lag order. ■

Proof of Proposition 5. Recall that $\tau_k(r)$ is characterized by the following eigenvalue problem:

$$|\lambda S_{11} - S_{10} S_{00}^{-1} S_{01}| = 0, \tag{A.7}$$

where $S_{ij} = N^{-1} \sum_{t=k+2}^{T} R_{it} R_{jt}'$, with R_{0t} the residuals from a regression of Δy_t on $(\Delta y_{t-1}, \ldots, \Delta y_{t-k})$ and R_{1t} the residuals from a regression of y_{t-1} on the same set of regressors. Let the eigenvalues of the system (A.7) be $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_n$; then

$$\tau_k(r) = -N \sum_{j=r+1}^{n} \log(1 - \hat{\lambda}_j).$$

To simplify notation, define $w_t = (\Delta y_{t-1}', \ldots, \Delta y_{t-k}')'$. We first establish a lemma that characterizes the distributions of the components of (A.7).

https://doi.org/10.1017/S0266466607070284 Published online by Cambridge University Press

LEMMA A.5. *Let $W(s)$ be an $n - r$-dimensional vector Wiener process. If the data are generated by (21), then, under Assumption 2, as $T \to \infty$ and uniformly in k,*

(i) $N^{-1}\sum_{t=k+2}^{T} y_{t-1}y'_{t-1} = \Sigma + \begin{bmatrix} 0 & 0 \\ 0 & \Sigma_{22}^{1/2}(\bar{c}^2\int_0^1 W(s)W(s)'\,ds)\Sigma_{22}^{1/2} \end{bmatrix} + O_p(T^{-1/2})$.

(ii) $\|N^{-1}\sum_{t=k+2}^{T} w_t w'_t - \bar{V}\otimes\Sigma\|_1 = O_p(k/T^{1/2})$, *with*

$$\bar{V} = \begin{bmatrix} 2 & -1 & 0s & \dots \\ -1 & 2 & -1 & 0s \\ . & . & . & . \\ 0s & -1 & 2 & -1 \\ \dots & 0s & -1 & 2 \end{bmatrix}.$$

Also, $|\bar{V}| = k+1$, *and* $(\bar{V}^{-1})_{1,1} = k/(k+1)$.

(iii) $\|N^{-1}\sum_{t=k+2}^{T} w_t y'_{t-1} - \iota_1\otimes\Sigma\|_1 = O_p((k/T)^{1/2})$, *with* ι_1 *the* $k\times 1$ *vector* $\iota_1 = (1,0,\dots,0)'$.

(iv) $\|N^{-1}\sum_{t=k+2}^{T} w_t \Delta y'_t + \iota_1\otimes\Sigma\|_1 = O_p((k/T)^{1/2})$.

(v) $N^{-1}\sum_{t=k+2}^{T} y_{t-1}\Delta y'_t = -\Sigma + O_p(T^{-1/2})$.

Proof of Lemma A.5. First, in what follows, all approximation results hold uniformly in $k \le K_T$, unless stated otherwise. For part (i),

$$N^{-1}\sum_{t=k+2}^{T} y_{t-1}y'_{t-1} = N^{-1}\begin{bmatrix} \sum_{t=k+2}^{T} u_{1,t-1}u'_{1,t-1} & \sum_{t=k+2}^{T} u_{1,t-1}y'_{2,t-1} \\ \sum_{t=k+2}^{T} y_{2,t-1}u'_{1,t-1} & \sum_{t=k+2}^{T} y_{2,t-1}y'_{2,t-1} \end{bmatrix}.$$

As in Nabeya and Perron (1994), $y_{2,t}$ can be decomposed as $y_{2,t} = a_T\varepsilon_t + b_T x_t$, with $a_T = -\theta_T$, $b_T = \bar{c}/\sqrt{T}$, and $x_t = x_{t-1} + \varepsilon_t$, an integrated process of dimension $n - r$. Using a standard central limit theorem (CLT),

$$N^{-1}\sum_{t=k+2}^{T} u_{1,t-1}u'_{1,t-1} - \Sigma_{11} = O_p(T^{-1/2}).$$

For the term $N^{-1}\sum_{t=k+2}^{T} y_{2,t-1}y'_{2,t-1}$, we have

$$\begin{aligned} N^{-1}\sum_{t=k+2}^{T} y_{2,t-1}y'_{2,t-1} &= N^{-1}\sum_{t=k+2}^{T}(a_T\varepsilon_{t-1} + b_T x_{t-1})(a_T\varepsilon_{t-1} + b_T x_{t-1})' \\ &= N^{-1}\sum_{t=k+2}^{T} a_T^2\varepsilon_{t-1}\varepsilon'_{t-1} + N^{-1}\sum_{t=k+2}^{T} b_T^2 x_{t-1}x'_{t-1} + O_p(T^{-1/2}) \\ &= \Sigma_{22}^{1/2}\left(I_{n-r} + \bar{c}^2\int_0^1 W(s)W(s)'\,ds\right)\Sigma_{22}^{1/2} + O_p(T^{-1/2}). \end{aligned}$$

https://doi.org/10.1017/S0266466607070284 Published online by Cambridge University Press

For the term $N^{-1}\sum_{t=k+2}^{T} u_{1,t-1}y'_{2,t-1}$, we have

$$N^{-1}\sum_{t=k+2}^{T} u_{1,t-1}y'_{2,t-1} = N^{-1}\sum_{t=k+2}^{T} u_{1,t-1}(a_T\varepsilon_{t-1} + b_T x_{t-1})'$$

$$= N^{-1}\sum_{t=k+2}^{T} u_{1,t-1}a_T\varepsilon'_{t-1} + O_p(T^{-1/2}) = \Sigma_{12} + O_p(T^{-1/2}).$$

The preceding results imply that

$$N^{-1}\sum_{t=k+2}^{T} y_{t-1}y'_{t-1} = \Sigma + \begin{bmatrix} 0 & 0 \\ 0 & \Sigma_{22}^{1/2}\left(\bar{c}^2\int_0^1 W(s)W(s)'\,ds\right)\Sigma_{22}^{1/2} \end{bmatrix} + O_p(T^{-1/2}).$$

For part (ii), by definition $w_t = [\Delta y'_{t-1}, \Delta y'_{t-2}, \ldots, \Delta y'_{t-k}]'$; hence $N^{-1}\sum_{t=k+2}^{T} w_t w'_t$ is a $kn \times kn$ matrix, which can be partitioned into k^2 blocks each of dimension $n \times n$, where the (i,j) block of the partitioned matrix is $N^{-1}\sum_{t=k+2}^{T}\Delta y_{t-i}\Delta y'_{t-j}$. We have the following results for different combinations of i and j. For $i = j$,

$$N^{-1}\sum_{t=k+2}^{T}\Delta y_{t-i}\Delta y'_{t-j} = \begin{bmatrix} N^{-1}\sum_{t=k+2}^{T}\Delta u_{1,t-i}\Delta u'_{1,t-i} & N^{-1}\sum_{t=k+2}^{T}\Delta u_{1,t-i}u'_{2,t-i} \\ N^{-1}\sum_{t=k+2}^{T} u_{2,t-i}\Delta u'_{1,t-i} & N^{-1}\sum_{t=k+2}^{T} u_{2,t-i}u'_{2,t-i} \end{bmatrix}.$$

We have

$$N^{-1}\sum_{t=k+2}^{T}\Delta u_{1,t-i}\Delta u'_{1,t-i} = 2\Sigma_{11} + O_p(T^{-1/2})$$

using a standard CLT, and

$$N^{-1}\sum_{t=k+2}^{T}\Delta u_{1,t-i}u'_{2,t-i} = N^{-1}\sum_{t=k+2}^{T}\Delta u_{1,t-i}\Delta\varepsilon'_{t-i} + \frac{\bar{c}}{\sqrt{T}}\left(N^{-1}\sum_{t=k+2}^{T}\Delta u_{1,t-i}\varepsilon'_{t-i-1}\right)$$

$$= 2\Sigma_{12} + O_p(T^{-1/2}),$$

where we have applied a CLT and a weak law of large numbers. Also,

$$N^{-1}\sum_{t=k+2}^{T} u_{2,t-i}u'_{2,t-i} = N^{-1}\sum_{t=k+2}^{T}\Delta\varepsilon_{t-i}\Delta\varepsilon'_{t-i} + O_p(T^{-1/2}) = 2\Sigma_{22} + O_p(T^{-1/2}).$$

Hence,

$$N^{-1}\sum_{t=k+2}^{T}\Delta y_{t-i}\Delta y'_{t-i} = 2\Sigma + O_p(T^{-1/2}).$$

https://doi.org/10.1017/S0266466607070284 Published online by Cambridge University Press

Applying similar arguments, we have, for $i - j = \pm 1$,

$$N^{-1} \sum_{t=k+2}^{T} \Delta y_{t-i} \Delta y'_{t-j} = -\Sigma + O_p(T^{-1/2});$$

and for $|i - j| > 1$,

$$N^{-1} \sum_{t=k+2}^{T} \Delta y_{t-i} \Delta y'_{t-j} = O_p(T^{-1/2}).$$

Hence,

$$N^{-1} \sum_{t=k+2}^{T} w_t w'_t - \begin{bmatrix} 2 & -1 & 0s & \dots \\ -1 & 2 & -1 & 0s \\ . & . & . & . \\ 0s & -1 & 2 & -1 \\ \dots & 0s & -1 & 2 \end{bmatrix} \otimes \Sigma = O_p(k/T^{1/2}),$$

where the difference is $O_p(k/T^{1/2})$ under the matrix norm. For part (iii), we consider the following two cases. For $j = 1$,

$$N^{-1} \sum_{t=k+2}^{T} \Delta y_{t-1} y'_{t-1} \xrightarrow{p} \Sigma$$

using the fact that $y_{2t} = a_T \varepsilon_t + b_T x_t$; and, for $j \geq 2$,

$$N^{-1} \sum_{t=k+2}^{T} \Delta y_{t-j} y'_{t-1} \xrightarrow{p} 0.$$

Again, using a CLT and applying arguments similar to those in Lewis and Reinsel (1985) yields the desired result. Part (iv) follows from part (ii). Part (v) also follows directly. ■

We now can proceed to prove Proposition 5. First, if k is fixed, all the eigenvalues of (A.7) are bounded away from zero. Accordingly $\tau_k(r)$ diverges at rate $O_p(T)$, which establishes the first result. For the second result, we first analyze the behavior of S_{10}, S_{11}, and S_{00} with $k \to \infty$ and $k^4/T \to 0$. For S_{10},

$$\begin{aligned} S_{10} &= N^{-1} \sum_{t=k+2}^{T} y_{t-1} \Delta y'_t - N^{-1} \sum_{t=k+2}^{T} y_{t-1} w'_t \left(\sum_{t=k+2}^{T} w_t w'_t \right)^{-1} \sum_{t=k+2}^{T} w_t \Delta y'_t \\ &= -\Sigma + (\iota'_1 \otimes \Sigma)(\bar{V}^{-1} \otimes \Sigma^{-1})(\iota_1 \otimes \Sigma) + O_p(k/T^{1/2}) \\ &= -\frac{1}{k+1} \Sigma + O_p(k/T^{1/2}) \end{aligned} \tag{A.8}$$

https://doi.org/10.1017/S0266466607070284 Published online by Cambridge University Press

using Lemma A5(ii)–(v). For S_{11},

$$S_{11} = N^{-1}\sum_{t=k+2}^{T} y_{t-1}y_{t-1}' - N^{-1}\sum_{t=k+2}^{T} y_{t-1}w_t'\left(\sum_{t=k+2}^{T} w_t w_t'\right)^{-1}\sum_{t=k+2}^{T} w_t y_{t-1}'$$

$$= \frac{1}{k+1}\Sigma + \begin{bmatrix} 0 & 0 \\ 0 & \Sigma_{22}^{1/2}\left(\bar{c}^2\int_0^1 W(s)W(s)'\,ds\right)\Sigma_{22}^{1/2} \end{bmatrix} + O_p(k/T^{1/2}) \qquad \textbf{(A.9)}$$

using Lemma A5(i)–(iii). For S_{00},

$$S_{00} = N^{-1}\sum_{t=k+2}^{T} \Delta y_t \Delta y_t' - N^{-1}\sum_{t=k+2}^{T} \Delta y_t w_t'\left(\sum_{t=k+2}^{T} w_t w_t'\right)^{-1}\sum_{t=k+2}^{T} w_t \Delta y_t'$$

$$= 2\Sigma - (\iota_1' \otimes \Sigma)(\bar{V}^{-1} \otimes \Sigma^{-1})(\iota_1 \otimes \Sigma) + O_p(k/T^{1/2})$$

$$= \frac{k+2}{k+1}\Sigma + O_p(k/T^{1/2}) \qquad \textbf{(A.10)}$$

using Lemma A5(ii) and (iv). Hence,

$$S_{00}^{-1} = \frac{k+1}{k+2}\Sigma^{-1} + O_p(k/T^{1/2})$$

using the fact that $S_{00} = O_p(1)$. After some algebra, we can show that the r largest eigenvalues of (A.7) are of order $O_p(1/(k+2))$ and the others are of order $O_p(1/[(k+1)(k+2)])$. To derive the limiting distribution of the smallest $(n-r)$ eigenvalues, we follow the approach of Johansen (1996, p. 159). Define $\beta = (I_r, 0_{r\times(n-r)})'$, $\beta_\perp = (0_{(n-r)\times r}, I_{n-r})'$, and $S(\lambda) = \lambda S_{11} - S_{10}S_{00}^{-1}S_{01}$. Consider the decomposition

$$|(\beta,\beta_\perp)'S(\lambda)(\beta,\beta_\perp)| = |\beta'S(\lambda)\beta|\,|\beta_\perp'\{S(\lambda) - S(\lambda)\beta[\beta'S(\lambda)\beta]^{-1}\beta'S(\lambda)\beta_\perp\}|$$

and let $T \to \infty$ and $\lambda \to 0$, such that $\rho = (k+1)(k+2)\lambda$ is fixed. From (A.8)–(A.10),

$$\beta'S(\lambda)\beta = (k+1)^{-1}(k+2)^{-1}\rho\beta'S_{11}\beta - \beta'S_{10}S_{00}^{-1}S_{01}\beta$$

$$= (k+1)^{-1}(k+2)^{-1}\Sigma_{11}\left(\frac{\rho}{k+1} - 1\right) + o_p(k^{-2})$$

$$= -(k+1)^{-1}(k+2)^{-1}\Sigma_{11} + o_p(k^{-2}).$$

Now,

$$\beta_\perp'\{S(\lambda) - S(\lambda)\beta[\beta'S(\lambda)\beta]^{-1}\beta'S(\lambda)\beta_\perp\}$$

$$= \frac{1}{(k+1)(k+2)}\left\{\rho\Sigma_{22}^{1/2}\left(\bar{c}^2\int_0^1 W(s)W(s)'\,ds\right)\Sigma_{22}^{1/2} - (\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})\right\}$$

$$+ o_p(k^{-2}),$$

https://doi.org/10.1017/S0266466607070284 Published online by Cambridge University Press

again using (A.8)–(A.10). The preceding results imply that the smallest $(n - r)$ eigenvalues, normalized by $(k + 1)(k + 2)$, converge to those of

$$\left| \rho \Sigma_{22}^{1/2} \left(\bar{c}^2 \int_0^1 W(s) W(s)' \, ds \right) \Sigma_{22}^{1/2} - (\Sigma_{22} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12}) \right| = 0, \tag{A.11}$$

and the sum of the eigenvalues of (A.11) converges to

$$\operatorname{tr} \left\{ (\Sigma_{22} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12}) \left(\Sigma_{22}^{1/2} \left(\bar{c}^2 \int_0^1 W(s) W(s)' \, ds \right) \Sigma_{22}^{1/2} \right)^{-1} \right\},$$

which completes the proof. ■

https://doi.org/10.1017/S0266466607070284 Published online by Cambridge University Press