OXFORD BULLETIN OF ECONOMICS AND STATISTICS, 55, 3 (1993)
0305-9049 $3.00

# FINITE-SAMPLE SIZES OF JOHANSEN'S LIKELIHOOD RATIO TESTS FOR COINTEGRATION

*Yin-Wong Cheung and Kon S. Lai*

## 1 INTRODUCTION

Since the work of Engle and Granger (1987) and Granger (1986), there has been a surge of interest in cointegration analyses of equilibrium relationships between non-stationary economic variables. While non-stationary economic series can wander widely through time, economic theory often suggests that specific sets of variables should obey certain long-run equilibrium constraints. If the individual economic series are stationary only after differencing but a linear combination of their levels is stationary, then the series are said to be cointegrated.

An approach to analyze cointegrated systems due to Johansen (1988, 1991) has received much attention recently. Johansen proposes a maximum likelihood (ML) method for estimating long-run equilibrium relationships or cointegrating vectors and derives likelihood ratio (LR) tests for cointegration in a Gaussian vector error correction model.[1] Phillips (1991), examining the distributional properties of the ML estimator of cointegrating vectors, shows that the ML estimator is super-consistent, symmetrically distributed, and median-unbiased asymptotically, and that an optimal theory of inference applies.

An issue concerns the performance of the ML cointegration analysis in finite samples. The LR tests for cointegration is derived from asymptotic results and statistical inferences in finite samples may not be appropriate. In particular, the critical values based on asymptotic distributions can be misleading. Moreover, the LR tests are conducted in a vector autoregressive framework with a given lag specification. Since the 'true' model is generally not known in practice, it is desirable to investigate the sensitivity of the tests

[1]Based on the triangular representation of the error correction model (e.g., Phillips (1991)), Stock and Watson (1991) derive efficient estimators of cointegrating vectors that are computed using the ordinary or generalized least squares method. While Stock and Watson's estimators can apply to systems involving some variables with higher orders of integration, their estimators are asymptotically equivalent to Johansen's (1991) ML estimator in the standard case where variables are all integrated of order one.

to model misspecification. Further, while the theoretical distributions of the LR statistics are obtained under the normality assumption, the assumption seems restrictive for general applicability and little is known about how its violation can affect the distributions of the test statisics.

The purpose of this study is to examine the finite-sample sizes of Johansen's (1991) LR tests for cointegration and their robustness to lag length specification and non-normal errors. In related research, Gonzalo (1990) examines the finite-sample properties of the ML estimator of cointegrating vectors.[2] In contrast to Gonzalo (1990), which deals mainly with the problem of estimation, the present study addresses the problem of hypothesis testing and inference. Accordingly, this study contributes further evidence on the performance of Johansen's ML technique for cointegration analysis. Johansen (1988) derives a LR cointegration test based on a vector autoregressive model without a constant term. Johansen (1991) shows, however, that when a constant term is included in the model, both the LR test statistic and its asymptotic distribution are altered. In addition, the analysis depends crucially on whether or not the underlying processes are allowed to contain a linear trend in the non-stationary part. In this analysis we consider Johansen's (1991) cointegration tests that allow for a time trend, whereas Gonsalo (1990) examines Johansen's (1988) method for estimating cointegrating vectors.

The paper is organized as follows. Section II briefly discusses Johansen's LR tests for cointegration. Section III reports the finite-sample critical values of the LR test statistics based on response surface analysis. Section IV examines the sensitivity of the LR tests for cointegration to lag specification and non-normal errors. Section V concludes.

## II. LIKELIHOOD RATIO TESTS FOR COINTEGRATION

Johansen's (1991) statistical analysis is based on the technique of reduced rank regression (e.g., Ahn and Reinsel (1990) and Velu *et al.* (1986)). Johansen shows that cointegration can be tested as the hypothesis of a reduced rank of a regression coefficient matrix in a vector error correction model (VECM).

Consider in general an $n \times 1$ time series vector $X_t = (x_{1t}, \ldots, x_{nt})'$ represented by

$$C(\mathrm{L})X_t = \mu + \epsilon_t \tag{1}$$

[2]Gonzalo (1990) reports that the finite-sample properties of the Johansen ML estimator of cointegrating vectors are consistent with the asymptotic results in Phillips (1991). This is so even when the errors are non-normal or when the model is over-parameterized. The ML method is also shown to perform better than other methods for estimating cointegrating vectors, including ordinary least squares (Engle and Granger (1987)), non-linear least squares (Stock (1987)), principal components (Stock and Watson (1988)), and canonical correlations (Bossaerts (1988)).

where L is the lag operator, $C(\mathrm{L}) = \mathrm{I} - C_1\mathrm{L} - \ldots - C_k\mathrm{L}^k$, $\mu$ is some constant vector and $\epsilon_t$ is a vector of white Gaussian noises with mean zero and variance matrix $\Omega$. By writing $C(\mathrm{L}) = C(1)\mathrm{L}^k + \Delta C^*(\mathrm{L})$ where $C^*(0) = \mathrm{I}$ and $\Delta = 1 - \mathrm{L}$, the VECM representation for (1) is

$$\Delta X_t = \mu + \Gamma_1 \Delta X_{t-1} + \ldots + \Gamma_{k-1}\Delta X_{t-k+1} + \Gamma_k X_{t-k} + \epsilon_t \tag{2}$$

where $\Gamma_j = -\mathrm{I} + C_1 + \ldots + C_j$, $j = 1, .., k$. It follows that $C(1) = -\Gamma_k$. Under the hypothesis of $r (0 \le r < n)$ cointegrating vectors, $C(1) = \delta\alpha'$ where $\delta$ and $\alpha$ are $n \times r$ matrices of rank $r$ such that $\alpha' X_t$ is stationary. The $r$ columns of $\alpha$ are the cointegrating vectors. Johansen (1991) shows that ML estimation of the cointegration space spanned by $\alpha$ can be based on the least squares residuals from two vector regressions:

$$\Delta X_t = c_1 + \Gamma_1 \Delta X_{t-1} + \ldots + \Gamma_{k-1}\Delta X_{t-k+1} + \xi_{1t} \tag{3}$$

$$X_{t-k} = c_2 + \Gamma_1 \Delta X_{t-1} + \ldots + \Gamma_{k-1}\Delta X_{t-k+1} + \xi_{2t} \tag{4}$$

where $c_1$ and $c_2$ are constant terms. Define the product moment matrices of the residuals as

$$S_{ij} = T^{-1} \sum_{t=1}^{T} \hat{\xi}_{it}\hat{\xi}'_{jt}, \qquad i, j = 1, 2. \tag{5}$$

The LR test statistic for the hypothesis of at most $r$ cointegrating vectors is

$$-2 \ln Q = -T \sum_{j=r+1}^{n} \ln(1 - \lambda_j) \tag{6}$$

with $\lambda_{r+1}, \ldots, \lambda_n$ being the $n - r$ smallest eigenvalues of $S_{21}S_{11}^{-1}S_{12}$ with respect to $S_{22}$. The asymptotic distribution of the $-2 \ln Q$ statistic is given by the trace of a stochastic matrix

$$\int_0^1 (\mathrm{d}W)F' \left[\int_0^1 FF'\,\mathrm{d}t\right]^{-1} \int_0^1 F(\mathrm{d}W)' \tag{7}$$

where $W$ is an $(n-r)$-dimensional standard Brownian motion and $F = \{F_1(t), \ldots, F_{n-r}(t)\}$ is defined by

$$F_j(t) = W_j(t) - \int_0^1 W_j(s)\,\mathrm{d}s, \qquad j = 1, \ldots, n - r.$$

The statistic in (6) is called the trace statistic by Johansen and Juselius (1990). An alternative LR statistic, given by

$$-2 \ln Q_{r|r+1} = -T \ln(1 - \lambda_{r+1}) \tag{8}$$

and called the maximal eigenvalue statistic, examines the null hypothesis of $r$ cointegrating vectors against the alternative of $r+1$ cointegrating vectors. The asymptotic distribution of this statistic is given by the maximum eigenvalue of the stochastic matrix in (7). The percentiles of the asymptotic distribution for both the trace and the maximal eigenvalue statistics are tabulated in Johansen and Juselius (1990; Table A2) using simulation analysis.

## III. FINITE-SAMPLE CRITICAL VALUES

Although the asymptotic critical values of Johansen's LR tests can be obtained from theoretical distributions, empirical analyses necessarily deal with finite samples and the quality of the asymptotic approximations to critical values in finite samples becomes an important issue. To the extent that the appropriate finite-sample critical values are different from the asymptotic ones, finite-sample analyses can bias the LR tests toward finding cointegration either too often or too infrequently. It is thus important to examine the direction and the degree of the finite-sample bias of the LR cointegration tests.

In this study, the finite-sample critical values for both the trace and maximal eigenvalue tests are estimated using response surface analysis in Monte Carlo experiments. Response surface methodology has been used in many fields of applied statistics (see, for example, a review article by Myers, Khuri and Carter (1989)). Early econometric studies that employ response surface analysis include Hendry (1979), Hendry and Harrison (1974), and Hendry and Srba (1977). Recently, MacKinnon (1991) uses response surface analysis to provide finite-sample critical values for the Engle–Granger (1987) two-step cointegration tests. Response surface estimation of critical values is useful. It can yield a simple summarization of the simulation outcomes, and estimates of critical values for not only a few sample sizes but for any finite sample size can be computed easily from response surface equations.

Response surface analysis applies in general to a system where the response of some variable depends on a set of other variables that can be controlled and measured in experiments. The analysis proceeds with simulation experiments evaluating the effects on the response variable of designed changes in the control variables. A response surface describing the response variable as a function of the control variables is then fitted to the experimental data.

In our analysis, the response variable is the finite-sample critical value and the control variables are the sample size ($T$), the number of variables in the estimated system ($n$), and the lag parameter ($k$). A factorial experimental design is employed, covering different possible combinations of $T=\{33, 36, 39, 42, 45, 50, 55, 60, 70, 80, 90, 100, 150, 200, 300, 500\}$, $n=\{2, 3, 4, 5\}$, and $k=\{1, 2, 4, 6\}$. The factorial design can provide a comprehensive coverage of interactions among the control variables. Due to limited degrees of freedom, the experiments for $n=5$, $k=6$, and $T\leq 39$ will not be included.

The design here yields 253 simulation experiments. The data generating process (DGP) is considered, as in Engle and Granger (1987) and Engle and Yoo (1987), to be given by

$$X_t = X_{t-1} + \mu_t \tag{9}$$

with $X_t = (x_{1t}, \ldots, x_{nt})'$ and $u_t$ being an $n \times 1$ vector of independently distributed normal errors of mean zero and covariance matrix $I_n$. In the Monte Carlo experiment sample series of $X_t$ are generated by setting the initial values of $X_0$ equal to zero and creating $T+50$ observations, of which the first 50 observations are discarded to minimize the effect of the initial condition. The Monte Carlo results reported below are all obtained through 20,000 replications. The GAUSS programming language and the subroutine RNDN are used to generate pseudo-random normal innovations.

While the finite-sample critical values are expected to be a function of $T$, $n$, and $k$, finding a proper approximating function for response surface regressions is important for obtaining accurate estimates of the critical values. In general, a response surface can take many possible functional forms. The formulation of the functional form, nonetheless, needs to satisfy some restriction. In our case, specifically, for given values of $n$ and $k$, the finite-sample critical values should equal the asymptotic values at the limit as the value of $T$ approaches infinity. This asymptotic restriction therefore provides a useful criterion for checking the adequacy of the specification of a response surface.

Previous results reported by Reinsel and Ahn (1988) suggest that a scaling factor, which is a simple function of $T$, $n$, and $k$, may be used to obtain approximate finite-sample critical values from their asymptotic counterparts. Reimers (1991), for example, applies the Reinsel–Ahn method to adjust Johansen's test statistics by a factor of $(T-nk)/T$ and compare them with their asymptotic critical values. An equivalent way to make finite-sample corrections is to adjust the critical values and not the test statistics. Accordingly, the implied scaling factor ($SF$) is given by

$$CR_T / CR_\infty = SF = T/(T-nk) \tag{10}$$

where $CR_T$ is the approximate finite-sample critical value, $T$ is the effective number of observations, and $CR_\infty$ is the asymptotic critical value at the corresponding significance level. Estimates of $CR_\infty$ have been provided by Johansen and Juselius (1990). The ratio $CR_T/CR_\infty$ can be viewed as a measure of finite-sample bias: no bias exists only when $CR_T/CR_\infty = 1$. The denominator of the scaling factor can be considered a degree-of-freedom correction term. As the value of $T$ approaches infinity, the scaling factor diminishes to unity and so the value of $CR_T$ converges to its asymptotic value $CR_\infty$, satisfying the asymptotic restriction.

To study the performance of the Reinsel–Ahn approximation method, the following response surface equation is fitted:

$$CR_{Tj} / CR_\infty = \beta_0 + \beta_1 SF_j + \text{errors} \tag{11}$$

where $CR_{Tj}$ is the finite-sample simulation estimate of the critical value, $SF_j$ is the scaling variable defined earlier, and the subscript $j$ refers to the $j$th experiment. The Reinsel–Ahn method implies that $\beta_0 = 0$ and $\beta_1 = 1$, which jointly represent a testable hypothesis in regression analysis. Another interesting but weaker hypothesis that can be tested is $\beta_0 + \beta_1 = 1$. We observe that the asymptotic restriction does not strictly require $\beta_0 = 0$ and $\beta_1 = 1$. As $T \to \infty$, $SF \to 1$ and so $CR_T/CR_\infty \to 1$ as long as $\beta_0 + \beta_1 = 1$. Since the trace test and the maximal eigenvalue test are conducted at both 10 and 5 percent significance levels in the experiments, four response surface regressions are run.

Table 1 contains the response surface regression results. The results suggest that the ability of equation (11) to fit the data is remarkably good, given the high goodness of fit as measured by $R^2$. As reported in Table 1, the $R^2$ values are all greater than 0.997. The sum of squared errors (SSE) is also fairly small for every regression. The standard errors of estimates for the regressions were computed. They are given respectively by 0.01430 and 0.01496 for the 10 and 5 percent trace tests and by 0.01550 and 0.01765 for the 10 and 5 percent maximal eigenvalue tests. The Durbin–Watson (DW) statistics appear satisfactory in all regressions, indicating the presence of no

TABLE 1
*Response Surface Estimation of Critical Values*

| *Regression coeficients and statistics* | *The trace test* | | *The maximal eigenvalue test* | |
|---|---|---|---|---|
| | *10%* | *5%* | *10%* | *5%* |
| $\beta_0$ | 0.11906<br>(0.00383)** | 0.10902<br>(0.03582)** | 0.08566<br>(0.00720)** | 0.07422<br>(0.00889)** |
| $\beta_1$ | 0.88224<br>(0.00308)** | 0.89150<br>(0.02865)** | 0.91284<br>(0.00599)** | 0.92481<br>(0.00740)** |
| $\chi^2$-test ($\beta_0 = 0$ and $\beta_1 = 1$) [$p$-value] | 215.0043<br>[0.0000] | 198.6887<br>[0.0000] | 647.5459<br>[0.0000] | 354.5935<br>[0.0000] |
| $\chi^2$-test ($\beta_0 + \beta_1 = 1$) [$p$-value] | 1.4151<br>[0.2342] | 0.2232<br>[0.6366] | 1.0660<br>[0.3018] | 0.30822<br>[0.57877] |
| $R^2$ | 0.9980 | 0.9978 | 0.9978 | 0.9972 |
| SSE | 0.0499 | 0.0546 | 0.0586 | 0.0760 |
| DW | 1.8496 | 1.7383 | 2.0359 | 1.9816 |

The response surface regression is given by equation (11). The numbers in parentheses are heteroskedasticity-consistent standard errors (White (1980)). Statistical significance is indicated by * at the 10% level and ** at the 5% level. The SSE value gives the sum of squared errors of the corresponding regression.

significant serial correlation. The validity of the asymptotic restriction $\beta_0 + \beta_1 = 1$ is tested, and the restriction is found to be well consistent with the experimental data. In no case can the hypothesis of $\beta_0 + \beta_1 = 1$ be rejected at the usual significance levels. In contrast, the hypothesis of $\beta_0 = 0$ and $\beta_1 = 1$, as implied by the Reinsel–Ahn method, is strongly rejected in all the four regressions. The rejection of the hypothesis of $\beta_0 = 0$ and $\beta_1 = 1$ suggests that the Reinsel–Ahn method does not yield unbiased estimates of the finite-sample critical values for Johansen's tests, and that more accurate estimates of the finite-sample critical values can be provided by the response surface.

The foregoing results support that the finite-sample bias of Johansen's tests is a positive function of $T/(T-nk)$. Since both $n$ and $k$ are of positive values, $T/(T-nk)$ is always greater than one for any finite $T$ value, indicating that the tests are biased toward finding cointegration too often when asymptotic critical values are used. Furthermore, the finite-sample bias toward over-rejection of the no cointegration hypothesis magnifies with increasing values of $n$ and $k$.

In addition to equation (11), a number of different specifications of the response surface were explored. For example, replacing $T-nk$ in $SF$ by $T-nk-1$ was found to reduce the fit of the response surface in all four regressions. Models of general power-series expansions in $T$, $n$, and $k$ yielded much worse fits. Moreover, the fit of the response surface deteriorated when $SF$ was replaced by $1/SF$. Estimation of a log-linear model of the response surface also failed to improve the data fit upon the linear model. On the other hand, including higher order terms in $SF$ in equation (11) could marginally improve the fit of the response surface:

$$CR_{Tj}/CR_{\infty} = \beta_0 + \beta_1 SF_j + \beta_2 SF_j^2 + \ldots + \beta_q SF_j^q + \text{errors}. \qquad (12)$$

However, for such specification, the implied asymptotic restriction was found to be not consistent with the experimental data. Hence, the results are not reported.

## IV. ROBUSTNESS OF THE LIKELIHOOD RATIO TESTS

### *Lag Specifications*

The Johansen analysis assumes that the underlying DGP has a finite order autoregressive representation with some known lag structure. In empirical applications, perfect information about the lag length is generally not available, and data series can exhibit moving-average dependence. Several questions of interest are: How does the presence of autoregressive or moving-average dependence affect the sizes of Johansen's cointegration tests? What is the sensitivity of the tests to under- and over-parametrization? Are standard model selection procedures such as those based on the Akaike information criterion (AIC) and the Schwarz information criterion (SIC) useful for choosing the optimal lag length for Johansen's tests? How do these

lag selection procedures perform in the presence of moving-average dependence?[3]

To study the effects of autoregressive dependence on the empirical test size and the role of lag length selection, Monte Carlo experiments are carried out using the following DGP:

$$\begin{aligned} x_{1t} &= x_{1t-1} + \phi_1(x_{1t-1} - x_{1t-2}) + w_{1t} \\ x_{2t} &= x_{2t-1} + \phi_2(x_{2t-1} - x_{2t-2}) + w_{2t} \end{aligned} \tag{13}$$

where $w_{1t}$ and $w_{2t}$ are independent Gaussian zero-mean white noise processes. The parameters $\phi_1$ and $\phi_2$ are arbitrarily set equal to 0.2 and 0.7, respectively. Johansen's tests with different lags, namely $k = 1, 2, 3$, and 4, are applied to the simulated data series with $T = 200$. Sample sizes of $T = 50$ and 100 were also considered, and they yielded qualitatively similar results. For each lag length, 20,000 replications are simulated. Since the lag length can affect the empirical test size, appropriate critical values corresponding to the individual lags are employed. In additional experiments, the lag length is not preset but estimated by the AIC and the SIC in each replication. Note that since the model in (13) implies $k = 2$, Johansen's tests with $k = 1$ will suffer under-parametrization, but those with $k = 3$ and 4 will suffer over-parametrization.

When compared with an autoregressive model, a moving-average error model analytically implies an infinite order autoregressive structure. An important question is whether a finite order autoregressive model can provide an adequate approximation for a moving average process. To examine how the approximation error can affect Johansen's cointegration tests, Monte Carlo experiments based on the following DGP are conducted:

$$\begin{aligned} x_{1t} &= x_{1t-1} + e_t - 0.3\,e_{t-1} \\ x_{2t} &= x_{2t-1} + w_t - 0.5\,w_{t-1} \end{aligned} \tag{14}$$

where $e_t$ and $w_t$ are independent Gaussian zero-mean white noise innovations. Johansen's tests with lags equal to 1 to 8 are performed on the simulated data series with $T = 200$. As in the experiments with autoregressive dependence, 20,000 replications are simulated and lag-adjusted critical values are employed. In separate experiments, the AIC and the SIC are applied to determine automatically the lag length used in each replication, without presetting the lag length.

Table 2A contains the Monte Carlo results concerning the effects of autoregressive dependence. In general, there is little difference in behaviour between the trace test and the maximal eigenvalue test. Both cointegration

[3]Reimers (1991) has examined the usefulness of standard lag selection criteria in estimating the lag order for Johansen's trace test in simulation experiments. Unlike the present study, Reimers does not consider data processes with moving-average innovations.

TABLE 2
*Effects of Lag Specification*

| Fitted $k$ | *The trace test* | | *The maximal eigenvalue test* | |
|---|---|---|---|---|
| | *SZ10* | *SZ05* | *SZ10* | *SZ05* |
| (A) The true DGP is given by model (13) with autoregressive dependence ($k = 2$): | | | | |
| 1 | 0.280 | 0.228 | 0.304 | 0.237 |
| 2 | 0.105 | 0.053 | 0.107 | 0.050 |
| 3 | 0.108 | 0.057 | 0.108 | 0.052 |
| 4 | 0.111 | 0.057 | 0.111 | 0.055 |
| AIC | 0.107 | 0.055 | 0.107 | 0.054 |
| SIC | 0.107 | 0.055 | 0.107 | 0.054 |
| Lag selected using the AIC and SIC (% of the time in 20,000 replications): | | | | |
| For $k = 1$, 0.000% using the AIC and 0.000% using the SIC. | | | | |
| For $k = 2$, 99.86% using the AIC and 99.96% using the SIC. | | | | |
| For $k = 3$, 0.125% using the AIC and 0.035% using the SIC. | | | | |
| For $k = 4$, 0.015% using the AIC and 0.005% using the SIC. | | | | |
| (B) The true DGP is given by model (14) with moving-average dependence: | | | | |
| 1 | 0.711 | 0.617 | 0.690 | 0.564 |
| 2 | 0.264 | 0.170 | 0.270 | 0.159 |
| 3 | 0.147 | 0.081 | 0.150 | 0.073 |
| 4 | 0.118 | 0.060 | 0.120 | 0.057 |
| 5 | 0.107 | 0.055 | 0.108 | 0.052 |
| 6 | 0.102 | 0.051 | 0.102 | 0.053 |
| 7 | 0.097 | 0.049 | 0.095 | 0.047 |
| 8 | 0.098 | 0.050 | 0.099 | 0.050 |
| AIC | 0.271 | 0.183 | 0.276 | 0.189 |
| SIC | 0.288 | 0.203 | 0.293 | 0.207 |
| Lag selected using the AIC and SIC (% of the time in 20,000 replications): | | | | |
| For $k = 1$, 5.235% using the AIC and 9.500% using the SIC. | | | | |
| For $k = 2$, 86.72% using the AIC and 86.37% using the SIC. | | | | |
| For $k = 3$, 7.905% using the AIC and 4.095% using the SIC. | | | | |
| For $k = 4$, 0.140% using the AIC and 0.040% using the SIC. | | | | |
| For $k = 5$, 0.000% using the AIC and 0.000% using the SIC. | | | | |
| For $k = 6$, 0.000% using the AIC and 0.000% using the SIC. | | | | |
| For $k = 7$, 0.000% using the AIC and 0.000% using the SIC. | | | | |
| For $k = 8$, 0.000% using the AIC and 0.000% using the SIC. | | | | |

All the simulation results are based on 20,000 replications. The columns beneath *SZ10* and *SZ05* give estimates of the empirical sizes of the 10% and 5% tests, respectively. In experiment (A), the maximum lag allowed in lag selection using the AIC and SIC is 4. In experiment (B), the maximum lag allowed in lag selection using the AIC and SIC is 8.

tests appear robust to over-parametrization, given that the biases in test sizes are reasonably small. In the case of under-parametrization ($k = 1$), however, the cointegration tests are severely distorted with the empirical test sizes being far above their nominal levels.[4] With a 5 percent test, for example, the upward bias in the test size is given by $+0.178$ for the trace test and by $+0.187$ for the maximal eigenvalue test. Interestingly, the Monte Carlo results also show that the lag selection criteria are useful for choosing the right lag length. The SIC seems to perform slightly better than the AIC. Still the AIC can pick the correct lag in more than 99.8 percent of all replications in our case.

Table 2B contains the Monte Carlo results for the effects of moving-average dependence. When a relatively low order autoregressive model is used, both the trace tests and the maximal eigenvalues test are seriously biased toward spuriously finding cointegration. The results underscore the fact that the low order autoregressive model provides a poor approximation for the moving-average process. This misspecification problem can be mitigated by estimating a high order autoregressive model. When the number of lags is adequately specified, the results appear encouraging. The bias of either test becomes very small. In general, we expect that the number of lags required to provide an adequate approximation should increase with the strength of the moving-average dependence. Unfortunately, the lag selection criteria cannot help find the proper approximating model in this situation. As shown in Table 2B, both the AIC and the SIC perform very poorly in the presence of moving-average dependence. The disappointing performance suggests that we should not rely solely on the lag selection criteria in choosing the estimated model for Johansen's tests.

### *Non-Normal Innovations*

The critical values in the response surface analysis are obtained under the assumption of normal innovations, as in the formal derivation of Johansen's tests. This assumption may not be an appropriate characterization of the data series empirically. In the analysis below, we examine the potential bias in the test size due to non-normal innovations, including non-symmetric and leptokurtic ones. A commonly known source of leptokurtic innovations is conditional heteroskedasticity, which leads to heavy-tailed distributions.

A simple Monte Carlo experiment is designed to illustrate the potential effects of non-normal innovations on Johansen's cointegration tests. Fleishman's (1978) power transformation method, outlined in the Appendix, is employed to generate non-normal variates with specified first four moments: mean, variance, skewness, and kurtosis. This method imposes little

[4]The results here are comparable to those in Gonzalo (1990), which suggest that the ML estimator of cointegrating vectors are much more sensitive to under-parameterization than to over-parameterization.

structure on simulated innovation processes and reduces dependence of the analysis on a particular specification of the processes. In the Monte Carlo experiment the DGP is considered as follows:

$$x_{1t} = x_{1t-1} + v_t$$
$$x_{2t} = x_{2t-1} + z_t \tag{15}$$

where $v_t$ and $z_t$ are non-normal random variables generated using the Fleishman method. The mean of either variable is set equal to zero, and its variance is specified as unity. For simplicity, $v_t$ and $z_t$ are considered to have the same degree of skewness ($SK$) and excess kurtosis ($KU$) in the simulation experiment.[5] The experiment adopts a factorial design covering all combinations of $KU = \{0.00, 0.75, 1.50, 2.25, 3.00, 3.75\}$ and $SK = \{-1.00, -0.75, -0.5, 0.0, 0.5, 0.75, 1.0\}$.[6] In the case of $KU = SK = 0$, $v_t$ and $z_t$ are normally distributed innovations. The experimental results are based on 20,000 replications. To control for the effects of the sample size and the lag length on Johansen's tests, we set $T = 200$ and $k = 2$ in each replication. Hence, distortions in the test size detected in the experiment will arise from non-normality only.

If deviations from normality can affect Johansen's tests, the bias in the test size will be a function of $SK$ and $KU$. Without any knowledge about the exact functional relationship, we follow the usual response surface analysis and consider a power-series approximating function:

$$BIAS = \alpha_0 + \alpha_1 SK + \alpha_2 KU + \alpha_3 SK^2 + \alpha_4 KU^2 + \alpha_5 SK*KU + \alpha_6 SK^3 + \alpha_7 KU^3 + \alpha_8 SK^{2}*KU + \alpha_9 SK*KU^2 + \ldots + \text{error} \tag{16}$$

where $BIAS$ = the bias in the test size measured by the change in the empirical test size relative to that in the benchmark case where $SK = KU = 0$; $BIAS > 0$ indicates an upward bias and $BIAS < 0$ indicates a downward bias in the empirical test size. A testable restriction is that the constant term $\alpha_0$ equals zero since, by construction, $BIAS = 0$ when $SK = KU = 0$.

The estimation results concerning the effects of non-normality on Johansen's tests are reported in Table 3. After experimenting with different orders of power-series expansions, a second order power-series function was found to fit the data best for the trace test, whereas a third order one was found to fit the data best for the maximal eigenvalue test. The empirical results confirm that in no case can the hypothesis of a zero constant term be rejected at the usual significance levels. The results further show that

[5]The skewness is computed as the third sample moment standardized by the cube of the standard deviation. The excess kurtosis is the fourth sample moment divided by the square of the variance minus three. For a normal distribution both coefficients should be equal to zero.

[6]To provide some idea about the degree of skewness, we note that for a sample statistic of $|SK| \geq 0.5$, the hypothesis of a symmetric distribution can be rejected at the 5 percent significance level, even in samples of size 50.

skewness in innovations has a statistically significant effect on the test sizes of both the trace and the maximal eigenvalue test. On the other hand, the trace test seems to be more robust to excess kurtosis in innovations than the maximal eigenvalue test. For the trace test, all the coefficients on $KU$ and its related terms are not statistically significant, except for a cross-product term $SK*KU$ in the case of a 5 percent trace test. For the maximal eigenvalue test, in contrast, most of the coefficients on $KU$ and its related terms are statistically significant.

Given the possible effects of non-normality, the question next is whether the resulted bias in the test size is substantial in magnitude. The coefficient estimates reported in Table 3 all appear rather small and are of a magnitude of order $10^{-3}$ or smaller. To evaluate quantitatively how sensitive Johansen's tests are to deviations from normality, the estimated equations in Table 3 are used to compute the bias in the test size with a range of values of skewness and excess kurtosis. With $KU = 3$, 5, and 7 (but no skewness), for example, the estimated biases in the test size are given respectively by $+0.0018$, $+0.0028$, and $+0.0124$ for the 10 percent maximal eigenvalue test and by $-0.0002$, $+0.0001$, and $+0.0004$ for the 10 percent trace test. With $SK = 1$, 2, and 3 (but no excess kurtosis), the estimated biases in the test size are given respectively by $-0.0057$, $-0.0314$, and $-0.0894$ for the 10 percent maximal eigenvalue test and by $-0.0015$, $-0.0038$, and $-0.0074$ for the 10 percent trace test. These estimates indicate that while the maximal eigenvalue test is reasonably robust to excess kurtosis, it shows substantial bias in the presence of large skewness. The trace test, in contrast, shows little bias in the presence of either skewness or excess kurtosis. In general, the trace test is found to be more robust to both skewness and excess kurtosis than the maximal eigenvalue test. This finding may possibly reflect the difference in the way the test statistics of the two tests are constructed.

## V. CONCLUSIONS

In this study the finite-sample bias of Johansen's (1991) LR test for cointegration has been examined using the Monte Carlo method. Response surface analysis is employed to obtain approximations to the finite-sample critical values and show the individual roles of the sample size, the dimension of the variable system, and the lag order in determining the finite-sample bias of Johansen's tests. It is found that Johansen's tests are biased toward finding cointegration more often than what asymptotic theory suggests. Moreover, the finite-sample bias magnifies as the dimension of the estimated system or the lag length increases. Proper corrections of the critical values in finite samples are therefore particularly essential when the estimated system contains many variables and/or long lags. Other results obtained are summarized as follows:

TABLE 3
*Effects of Non-Normal Errors on the Test Size*

| *Regressors and statistics* | *The trace test* | | *The maximal eigenvalue test* | |
|---|---|---|---|---|
| | *10%* | *5%* | *10%* | *5%* |
| Constant | −0.00051<br>(0.00031) | −0.00019<br>(0.00023) | 0.00041<br>(0.00032) | 0.00004<br>(0.00023) |
| *SK* | −0.00032<br>(0.00037) | −0.00160<br>(0.00023)** | −0.00056<br>(0.00047) | −0.00293<br>(0.00038)** |
| *KU* | 0.00007<br>(0.00033) | 0.00034<br>(0.00021) | 0.00276<br>(0.00061)** | −0.00106<br>(0.00049)** |
| $SK^2$ | −0.00066<br>(0.00031)** | −0.00045<br>(0.00022)* | −0.00346<br>(0.00078)** | −0.00146<br>(0.00050)** |
| $KU^2$ | 0.00001<br>(0.00008) | −0.00004<br>(−0.00005) | −0.00122<br>(0.00036)** | 0.00076<br>(0.00031)** |
| *SK*KU* | 0.00004<br>(0.00013) | 0.00025<br>(0.00009)** | 0.00097<br>(0.00048)* | 0.00187<br>(0.00036)** |
| $SK^3$ | | | −0.00211<br>(0.00052)** | −0.00022<br>(0.00038) |
| $KU^3$ | | | 0.00015<br>(0.00006)** | −0.00013<br>(0.00005)** |
| *SK*KU* | | | −0.00011<br>(0.00010) | −0.00029<br>(0.00009)** |
| $SK*KU^2$ | | | 0.00096<br>(0.00028)** | 0.00052<br>(0.00021)** |
| $R^2$ | 0.1756 | 0.7727 | 0.8181 | 0.8259 |
| *DW* | 2.1645 | 2.1860 | 1.6229 | 2.3710 |
| Ljung–Box (12)<br>[*p*-value] | 4.7610<br>[0.9066] | 11.3038<br>[0.5031] | 15.4581<br>[0.2173] | 17.3187<br>[0.1380] |
| White's test<br>[*p*-value] | 21.7417<br>[0.1148] | 20.8334<br>[0.1422] | 31.0256<br>[0.2701] | 24.8095<br>[0.5821] |

The dependent variable is the bias in the empirical test size (see equation (16)). The Ljung–Box($r$) test statistic has asymptotically a chi-square distribution with $r$ degrees of freedom under the null hypothesis of no serial correlation in the residual. The numbers in parentheses are heteroskedasticity-consistent standard errors (White (1980)). Statistical significande is indicated by * at the 10% level and ** at the 5% level.

(1) Response surface analysis can provide better approximations for the finite-sample critical values of Johansen's tests than the Reinsel–Ahn (1988) method.

(2) Johansen's cointegration tests are rather sensitive to under-parametrization in the lag length, though not so to over-parametrization. The result points to the importance of proper lag specifications in estimating cointegrated systems.

(3) For autoregressive processes, standard lag selection criteria such as the AIC and the SIC can be useful for choosing the right lag order for Johansen's tests.

(4) The presence of moving-average dependence can seriously bias Johansen's tests toward finding spurious cointegration unless the number of lags is sufficiently specified to provide a good finite order approximation for the moving-average component. In this case, both the AIC and the SIC perform poorly in selecting the proper lag length for Johansen's tests.

(5) Between Johansen's two LR tests for cointegration, the trace test shows more robustness to both skewness and excess kurtosis in innovations than the maximal eigenvalue test.

*University of California, Santa Cruz,*
*California State University, Los Angeles*

*Date of Receipt of Final Manuscript: July 1992*

## REFERENCES

Ahn, S. K. and Reinsel, G. C. (1990). 'Estimation for Partially Nonstationary Multivariate Autoregressive Models', *Journal of the American Statistical Association*, Vol. 85, pp. 813–23.

Bossaerts, P. (1988). 'Common Nonstationary Components of Asset Prices', *Journal of Economic Dynamics and Control*, Vol. 12, pp. 347–64.

Engle, R. F. and Granger, C. W. J. (1987). 'Cointegration and Error Correction: Representation, Estimation, and Testing', *Econometrica*, Vol. 55, pp. 251–76.

Engle, R. F. and Yoo, B. S. (1987). 'Forecasting and Testing in Cointegrated Systems', *Journal of Econometrics*, Vol. 35, pp. 143–59.

Fleishman, A. I. (1978). 'A Method for Simulating Non-Normal Distribution', *Psychometrika*, Vol. 43, pp. 521–32.

Gonzalo, J. (1990). 'Comparison of Five Alternative Methods of Estimating Long Run Equilibrium Relationships', Discussion Paper, University of California at San Diego.

Granger, C. W. J. (1986). 'Developments in the Study of Cointegrated Economic Variables', *BULLETIN*, Vol. 48, pp. 213–28.

Johansen, S. (1988). 'Statistical Analysis of Cointegration Vectors', *Journal of Economic Dynamics and Control*, Vol. 12, pp. 231–54.

Johansen, S. (1991). 'Estimation and Hypothesis Testing of Cointegration Vectors in Gaussian Vector Autoregressive Models', *Econometrica*, Vol. 59, pp. 1551–580.

Johansen, S. and Juselius, K. (1990). 'Maxiumum Likelihood Estimation and Inference on Cointegration — With Applications to the Demand for Money', *BULLETIN*, Vol. 52, pp. 169–210.

Hendry, D. F. (1979). 'The Behaviour of Inconsistent Instrumental Variables Estimators in Dynamic Systems with Autocorrelated Errors', *Journal of Econometrics*, Vol. 9, pp. 295–314.

Hendry, D. F. and Harrison, R. W. (1974). 'Monte Carlo Methodology and the Small Sample Behaviour of Ordinary and Two-Stage Least Squares', *Journal of Econometrics*, Vol. 2, pp. 151–74.

Hendry, D. F. and Srba, F. (1977). 'The Properties of Autoregressive Instrumental Variables Estimators in Dynamic Systems', *Econometrica*, Vol. 45, pp. 969–90.

MacKinnon, J. G. (1991). 'Critical Values for Cointegration Tests', in Engle, R. F. and Granger, C. W. J. (eds), *Long-Run Economic Relationships: Readings in Cointegration*, Oxford University Press, New York.

Myers, R. H., Khuri, A. I. and Carter, W. H. (1989). 'Response Surface Methodology: 1966–1988', *Technometrics*, Vol. 31, pp. 137–57.

Phillips, P. C. B. (1991). 'Optimal Inference in Cointegrated Systems', *Econometrica*, Vol. 59, pp. 283–306.

Reinsel, G. C. and Ahn, S. K. (1988). 'Asymptotic Properties of the Likelihood Ratio Test for Cointegration in the Nonstationary Vector AR Model', Technical Report, Department of Statistics, University of Wisconsin, Madison.

Reimers, H.-E. (1991). 'Comparisons of Tests for Multivariate Cointegration', Working Paper Number 58, Institute of Statistics and Econometrics, University of Kiel.

Stock, J. H. (1987). 'Asymptotic Properties of Least Squares Estimators of Cointegrating Vectors', *Econometrica*, Vol. 55, pp. 1035–56.

Stock, J. H. and Watson, M. W. (1988). 'Testing for Common Trends', *Journal of American Statistical Association*, Vol. 83, pp. 1097–107.

Stock, J. H. and Watson, M. W. (1991). 'A Simple Estimator of Cointegrating Vectors in Higher Order Integrated Systems', Working Paper, University of California at Berkeley.

Velu, R. P., Reinsel, G. C. and Wichern, D. W. (1986). 'Reduced Rank Models for Multiple Time Series', *Biometrika*, Vol. 73, pp. 105–18.

## APPENDIX

### *Simulating Non-normal Innovations*

A procedure to generate non-normal errors is devised by Fleishman (1978). The Fleishman method involves constructing a random variate, $z_t$, as a linear combination of a standard normal variate, $y_t$, and its square and cube:

$$z_t = a + by_t + cy_t^2 + dy_t^3. \tag{A1}$$

The constants $a$, $b$, $c$, and $d$ are chosen to provide $z_t$ with the specified distribution properties. By evaluating the first four moments of $z_t$, a system of four equations with four unknowns can be used to determine $a$, $b$, $c$, and $d$. Fleishman (1978) shows that for a distribution with mean zero and unit

variance, the corresponding constants can be obtained as the solution to the following equations:

$$b^2 + 6bd + 2c^2 + 15d^2 - 1 = 0$$

$$2c(b^2 + 24bd + 105d^2 + 2) - SK = 0$$

$$24[bd + c^2(1 + b^2 + 28bd) + d^2(12 + 48bd + 141c^2 + 225d^2)] - KU = 0$$

$$a = -c$$

where $SK$ is the desired skewness and $KU$ is the desired kurtosis. Since the system of equations are non-linear, it is solved using optimization methods. Using the solution values for $a$, $b$, $c$, and $d$, non-normal variates with different degrees of skewness and excess kurtosis can be generated by equation (A1).