ELSEVIER

Journal of Econometrics 79 (1997) 53–81

JOURNAL OF Econometrics

Estimation and inference in nearly unbalanced nearly cointegrated systems

Serena Ng[a,b], Pierre Perron[b]

[a] *Department of Economics, Boston College, Chestnut Hill, MA 02167, USA*
[b] *Département de Sciences Economiques and CRDE, Université de Montréal, CP 6128, Succ. Centre-Ville, Montréal, Québec, Canada H3C 3J7*

Received 1 July 1995; received in revised form 1 July 1996

Abstract

This paper considers the role of normalization in least-squares estimation of cointegrating vectors. It is shown, using an empirical example and Monte-Carlo simulations of bivariate models, that the least-squares estimates can have very poor finite sample properties when normalized in one direction but are well behaved when normalized in the other. This occurs when one of the $I(1)$ variables is a weak random walk or is nearly stationary. The choice of the regressand also has implications for residual based unit root tests for cointegration. We provide theoretical explanations for why the least-squares estimates from one normalization can be outright inconsistent in well-defined local asymptotic frameworks. Ranking the spectral density at frequency zero of the first differenced series is suggested as a practical guide to determining which variable to use as the regressand.

Keywords: Unit root; Cointegration; Normalization; Spectral density function at frequency zero
JEL classification: C22; C51; E31

1. Introduction

Cointegration is an important concept. It provides a tight analytical framework for analyzing the comovements of variables at low frequencies. A convenient way to obtain consistent estimates of cointegrating vectors is least-squares

The authors acknowledge grants from the Social Science and Humanities Research Council of Canada (SSHRC), the Fonds de la Formation de Chercheurs et l'Aide à la Recherche du Québec (FCAR), and the National Science and Engineering Research Council of Canada (NSERC). We thank two anonymous referees for helpful comments.

PII S0304-4076(97)00007-9

estimation. This paper is concerned with the robustness of the static least-squares estimator in single equation estimations of cointegrating vectors and of the role of normalization in estimating these cointegrating relationships.

A static cointegrating regression consists of regressing a variable belonging to the cointegrated system on the contemporaneous values of the remaining variables in the system, where all variables are known to be or can be tested as being $I(1)$. An important finding, due to Stock (1987), is that the least-squares estimator for this regression is super-consistent. That is to say, it converges at rate T, where T is the sample size. However, it has also been shown that serial correlation in some series and/or correlation among the variables in the system will, in general, induce asymptotic bias, asymmetry, and nuisance parameters to the limiting distribution of the least-squares estimator. See Phillips and Hansen (1990) among others. Accordingly, the least-squares estimator is sub-optimal relative to fully efficient estimators such as the FIML estimator of Johansen (1991), and the dynamic ordinary least-squares (DOLS) estimator of Saikkonen (1991) and Stock and Watson (1993).

In spite of the inefficiency of the least-squares estimator, its properties are still worthy of investigation for several reasons. First, the least-squares estimator provides the theoretical basis for the construction of more efficient estimators. Indeed, the fully modified estimator (FM-OLS) of Phillips and Hansen (1990) and the canonical cointegrating regression (CCR) of Park (1992) are built upon consistency of the static OLS estimator. Second, super-consistency of the static OLS estimator implies that the estimates should be reasonably precise. This provides a rationale for using the least-squares residuals from the static regression as the basis of tests for cointegration. See, for example, Phillips and Ouliaris (1990). Properties of the least-squares estimator have direct implications for the size and power of tests for cointegration to the extent that they affect the properties of the least-squares residuals. Third, least-squares estimation of the cointegrating vector provides an estimate of 'equilibrium error' for use in subsequent estimations of error correction models. Least squares estimation of the long-run cointegrating relationships therefore affects the estimated dynamics of the cointegrated system.

Estimations of cointegrating vectors require the practitioner to take a stand on normalization. In the context of least-squares procedures, this means deciding which variable to put on the left-hand side as the regressand. This has not been seen as an issue of much consequence as the conventional wisdom holds that while the normalization is known to imply different point estimates for elements of the cointegrating vector (except in the unrealistic situation when the R^2 of the regression is unity), it is not thought to have implications for the properties of the estimator in large samples. In particular, all normalizations yield super-consistent estimates. Accordingly, the regressions are usually normalized in a way to facilitate economic interpretation. However, few (if any) studies have examined if and when normalization affects the precision of the estimates. Our analysis in subsequent sections suggest that it does. We show, using bivariate models, that the

least-squares estimator can yield drastically different point estimates of the cointegrating vector depending on the normalization. This occurs when the cointegrated system contains variables with different univariate properties, in particular, when one of the variables has a weak random walk component or is nearly stationary. We refer to such a case as a 'nearly unbalanced nearly cointegrated system'.

The rest of the paper is structured as follows. We begin with an empirical example to motivate the analysis, and summarize salient features of the statistics that will be used. Section 3 describes two data-generating processes and presents simulation results which highlight the parameter space where the estimation problem will arise. Section 4 provides a theoretical explanation for the simulation results by means of local asymptotic analyses. Section 5 studies the properties of residual-based unit root tests in this context. Some observations on alternative estimators of cointegrating vectors are discussed in Section 6. We conclude with practical guidelines for inference and estimation.

2. An empirical example: The Fisher equation

We shall use the after-tax Fisher equation to give a synopsis of the problem. The Fisher equation is defined as

$$(1-\tau)i = r + \pi^e,$$

where τ is the average marginal tax rate, i is the nominal interest rate, π^e is the expected rate of inflation, and r is the real interest rate which is assumed to be a constant. The expected rate of inflation is unobserved, and replacing it by the actual rate of inflation will induce an errors-in-variable problem. Least-squares estimation will yield estimates that are inefficient and possibly suffer from simultaneity bias. But as long as it can be shown that i and π are $I(1)$ variables and that they are cointegrated, consistency of the estimates is implied by standard asymptotic results. To do this, we now review procedures which reveal the univariate properties of the data.

2.1. *Univariate statistics*

We use three statistics to test for the presence of a unit root and the null hypothesis of no cointegration. A constant term is included in the regressions where appropriate.

The Said and Dickey (1984) $t_{\rho\hat{\alpha}}$-statistic is constructed from an augmented autoregression with a constant and k lagged first-differences of the data, where k is selected using a general to specific procedure to test for the significance of the last lag beginning from the largest order, $kmax$.[1]

[1] See Ng and Perron (1995) for details concerning the relationship between k and size distortions.

The Phillips and Perron (1988) $Z_{\alpha\mu}$-test is based on the autoregression $y_t = \hat{\mu} + \hat{\alpha} y_{t-1} + \hat{u}_t$. To estimate the spectral density at frequency zero of the residuals u_t, we apply the quadratic kernel to the least-squares residuals $\hat{u}_t$. The bandwidth is selected using the automated procedure described in Andrews (1991).

The $MZ_{\alpha\mu}$ statistic was suggested by Stock (1990) and analyzed in Perron and Ng (1996). The statistic can be written as $MZ_{\alpha\mu} = Z_{\alpha\mu} + T(\hat{\alpha} - 1)^2/2$, hence can be seen as a modified Phillips–Perron test. Importantly, the statistic makes use of an autoregressive spectral density estimator which bypasses the use of the least-squares residuals. This estimator is defined as $s^2_{\mathrm{AR}} = \sigma^2_{ek}/(1 - \sum_{i=1}^{k} \hat{b}_i)^2$, with $\hat{b}_i$ and σ^2_{ek} obtained from the regression $\Delta y_t = c_0 + b_0 y_{t-1} + \sum_{j=1}^{k} b_j \Delta y_{t-j} + e_{tk}$. The advantages of MZ_α over Z_α and the role of the spectral density estimators are detailed in Perron and Ng (1996).

Features of these statistics relevant to subsequent analysis are:

- The $Z_{\alpha\mu}$ and $MZ_{\alpha\mu}$ tests are asymptotically equivalent under standard assumptions. However, when the serial correlation parameter in the errors of the unit root process is modeled as local to -1 and hence that the spectral density at frequency zero of Δy_t is small, $Z_{\alpha\mu}$ diverges to $-\infty$ but $MZ_{\alpha\mu}$ is $O_p(1)$, and the two tests have very different finite sample properties. As well, while the autoregressive spectral density estimator is consistent, the kernel estimator constructed using the least-squares residuals is not, irrespective of the choice of the kernel and/or the truncation lag. The root of the problem is the severe bias of the least-squares estimator for the autoregressive parameter when there is large negative serial correlation in the errors. Of the three tests considered, the size of $Z_{\alpha\mu}$ is most distorted, and the size of $MZ_{\alpha\mu}$ is more robust to a large negative MA component than $t_{\rho\mu}$.
- When the objective is to estimate the spectral density at frequency zero of Δy_t, we consider two estimators that are more efficient than the ones used in the unit root tests. First, we use the demeaned first differences of the data instead of the estimated residuals to construct the kernel-based estimates. Second, we drop the lagged level of y_t from the autoregression used to construct the autoregressive spectral density estimate. These alternative estimates are not guaranteed to be bounded above zero when the series is stationary, and hence will yield unit root tests that are inconsistent. However, as estimates of the spectral density of Δy_t at frequency zero, they are efficient because they do not depend on the least-squares residuals from the first-order autoregression.

As we will see later, these statistic will play an important role in our analysis. With this backdrop, we now return to our empirical example.

2.2. Empirical results

We use monthly observations from 1954:1 to 1993:11 for the US taken from CITIBASE (479 observations). The interest rate is the three month yield on treasury bills (fygm3) at annual rates, inflation is the annualized quarterly change

Table 1
Analysis of the Fisher equation: unit root tests

				$s^2(QS)\ s_u^2$		$s^2(AR)\ s_u^2$	
	$Z_{\gamma\mu}$	$MZ_{\gamma\mu}$	$t_{\rho\mu}$	(1)	(2)	(3)	(4)
i	−9.63	−8.07	−2.10	1.254	1.243	1.014	0.906
π	−40.31	−11.07	−2.54	1.160	0.312	0.306	0.253

Notes: $s^2(QS)$ is the quadratic kernel estimate, $s^2(AR)$ is the autoregressive spectral density estimate, and s_u^2 is the sample variance of the first differences of the data. (1) For $s^2(QS)$, Andrews' (1991) method is used to select the bandwidth. For i, an $AR(1)$ approximation is used and the method selects a bandwidth of 5; for π, an $MA(3)$ approximation is used and the method selects a bandwidth of 11. (1) The kernel estimator uses the estimated residuals from an autoregression with a constant. The estimate is used in $Z_{\gamma\mu}$.
(2) The kernel estimator uses the residuals from the demeaned first differences of the data.
(3) The $s^2(AR)$ is based on the augmented autoregression of $\Delta y_t = c + b_0 y_{t-1} + \sum_{i=1}^{8} \Delta y_{t-i} + e_t$. The estimate is used in $MZ_{\gamma\mu} = [T^{-1} y_T^2 - s^2(AR)](2T^{-2}\sum_1^T y_t^2)^{-1}$, where y is the demeaned series.
(4) The $s^2(AR)$ is based on the autoregression of $\Delta y_t = c - \sum_{i=1}^{8} \Delta y_{t-i} + e_t$.
The critical value for $Z_{\gamma\mu}$ and $MZ_{\gamma\mu}$ is −14.1, and is −2.86 for $t_{\rho\mu}$.

in the consumer price index ($\pi_t = 4\log(punew_t/punew_{t-3})$). The properties of the data are reported in Table 1. All three unit roots tests conclude that there is a unit root in the nominal interest rate (i). However, while $Z_{\gamma\mu}$ strongly rejects the presence of a unit root in the inflation rate (π), $MZ_{\gamma\mu}$ and $t_{\rho\mu}$ cannot reject the unit root hypothesis.

Now consider the estimates of the spectral density at frequency zero used in the unit root tests. For i, both the (normalized) kernel and the autoregressive spectral density estimates are in the range of unity [see columns (1) and (3) in Table 1]. But, for π, there are marked differences. Whereas the kernel estimate based on the least-squares residuals is around one, it is much smaller and is closer to zero according to s_{AR}^2. These discrepancies in the spectral density estimates, together with the observed discrepancy between $Z_{\gamma\mu}$ and $MZ_{\gamma\mu}$, suggest that there is negative serial correlation in errors of the inflation series in view of the discussion of the previous subsection.

Estimates of the spectral density at frequency zero of the first differences of the data are shown in columns 2 and 4. The quadratic and autoregressive spectral estimates are 1.243 and 0.906, respectively, for the interest rate. For the inflation series they are 0.312 and 0.253, respectively. The small values of the latter estimates again suggest negative residual serial correlation. It is of interest to note the large discrepancy between the residuals-based kernel estimate (i.e. those used in the unit root tests) and the kernel estimate based on the first-differenced data for the inflation series. This discrepancy reflects the poor properties of least-squares estimator in the autoregression, of which negative residual serial correlation is a possible cause.

Table 2
Coefficient estimates for the Fisher equation and cointegration tests

	OLS	DOLS(4)	DOLS(8)	DOLS(12)	Z_α	MZ_α	t_ρ
i on π	0.609	0.719	0.739	0.750	33.75	20.90	−2.01
π on i	0.785	0.788	0.806	0.798	−61.02	−33.04	−2.43

Notes: DOLS(k) denotes inclusion of k leads and lags of the first differences of the regressors in the static least squares regression. See (1) and (3) in notes to Table 1 for the spectral density at frequency zero used in the construction of Z_α and MZ_α. The 5% critical value for Z_α and MZ_α is −20.5, and −3.36 for t_ρ.

The conclusion that one would draw from these results is that the interest rate series is unambiguously $I(1)$. The more robust $MZ_{\alpha\mu}$ and $t_{\rho\mu}$ tests suggest that π is also $I(1)$ but there is strong negative serial correlation in the errors which induces a strong force for the series to revert to its mean. Indeed, estimations of ARIMA models for π reveal the significance of negative moving average lags. The best model selected by the AIC criterion suggests π is an $ARMA(1,3)$ with an autoregressive coefficient that is almost one and with the moving-average coefficient at the third lag in the neighborhood of −0.8, far outweighing the sum of the positive moving average coefficients at lags one and two.

Granted the result that both i and π are $I(1)$, we then proceed to estimate the Fisher equation and test if the two series are cointegrated. The Fisher equation makes no suggestion as to whether empirical tests of the relationship should use i or π as the regressand, so that without a strong a priori reasoning, it is equally legitimate to use π as the regressand as it is to use i. Assuming an average marginal tax rate in the US of around 0.3 over the sample, one would expect a regression of i on π to yield an estimate of $1/(1-\tau)>1$. If we regress π on i, we would expect a regression coefficient of 0.7. Furthermore, the estimates from the two equations should be (approximately) the reciprocal of each other. Table 2 reports the estimation results.

The estimated coefficient from the equation with i as the regressand is quite different from our prior as it falls short of unity. The coefficient from the equation with π as the regressand suggests a marginal tax rate of 0.22, which seems plausible. But, the two estimated coefficients are evidently not the reciprocal of one another. One might argue that inefficiency of the static least-squares estimator might be the source of the problem. However, discrepancies remain even when the equations are estimated with an efficient estimator such as the DOLS. The second through fourth columns of Table 2 augment the static least-squares regression with four, eight, and 12 leads and lags of the first differences of the regressor. For the equation with i on the left-hand side, estimation by DOLS raises the point estimate of the coefficient, but it continues to fall short of one. For the equation with π on the left-hand side, the additional regressors made practically no change to the static least-squares estimates. Furthermore, the coefficients from the two regressions look like they are identical rather than being the reciprocal

of one another. Not only do the regressions give puzzling evidence about the empirical Fisher relationship, the results from the various residual based tests for cointegration are just as confusing. The Phillips–Perron statistic always reject the null hypothesis of no-cointegration, the Said–Dickey statistic suggests no-cointegration, while the MZ_α statistic gives mixed results depending on which variable is used as the regressand.

The above example suggests that the choice of normalization can potentially yield dramatically different point estimates on coefficients of economic interest. As will become clear, one of the two estimates has very poor properties. The key to finding the appropriate normalization lies in the spectral density at frequency zero of the first differences of the regressand relative to those of the first differences of the regressors. The rest of this analysis provides a formal framework for analyzing the issues raised, with the aim of providing practical recommendations for which variable to use as the regressand.

3. When might normalization matter?

In this section, we first present the two data-generating processes used, and then report simulations to illustrate the nature of the problem. Unless noted otherwise, all simulations are performed using 1000 replications. The programs are written in C with routines from Press, Toukolsky, Vetterling and Flannery (1992) running under IRIX 5.2 on an SGI system.

3.1. *The data-generating processes*

Consider a bivariate model

$$x_t^* = \gamma_1^* \mu_t + e_{1t}^*, \quad e_{1t}^* \sim (0, \sigma_1^2),$$

$$y_t^* = \gamma_2^* \mu_t + e_{2t}^*, \quad e_{2t}^* \sim (0, \sigma_2^2),$$

$$\mu_t = \mu_{t-1} + v_t, \quad v_t \sim (0, \sigma_v^2).$$

The variables x_t and y_t are driven by a common stochastic trend μ_t, as well as stationary innovations e_{1t}^* and e_{2t}^* that are serially and mutually uncorrelated. Furthermore, e_{1t}^* and e_{2t}^* are uncorrelated with v_t at all leads and lags by assumption. A nonzero drift can be added to μ_t, but is omitted without loss of generality. Letting $x_t = x_t^*/\sigma_1$, $y_t = y_t^*/\sigma_2$, $e_{1t} = e_{1t}^*/\sigma_1$, $e_{2t} = e_{2t}^*/\sigma_2$, $\gamma_1 = \gamma_1^*/\sigma_1$, and $\gamma_2 = \gamma_2^*/\sigma_2$, we have

$$\begin{aligned} \text{DGP1:}\ x_t &= \gamma_1 \mu_t + e_{1t}, \quad e_{1t} \sim (0, 1), \\ y_t &= \gamma_2 \mu_t + e_{2t}, \quad e_{2t} \sim (0, 1), \\ \mu_t &= \mu_{t-1} + v_t, \quad v_t \sim (0, \sigma_v^2). \end{aligned} \tag{1}$$

The variables x_t and y_t are now of unobserved components form as in Clark (1987). The DGPs can be written as

$$\begin{aligned} \Delta x_t &= \gamma_1 v_t + e_{1t} - e_{1t-1} = u_t^x + \theta_x u_{t-1}^x, \\ \Delta y_t &= \gamma_2 v_t + e_{2t} - e_{2t-1} = u_t^y + \theta_y u_{t-1}^y, \end{aligned} \tag{2}$$

where θ_x is such that $\theta_x(1+\theta_x^2)^{-1} = \sigma_1^2[\gamma_1^2\sigma_v^2 + 2\sigma_1^2]^{-1}$ and θ_y is such that $\theta_y(1+\theta_y^2)^{-1} = \sigma_2^2[\gamma_2^2\sigma_v^2 + 2\sigma_2^2]^{-1}$. The model can also be parameterized as

$$\begin{aligned} \Delta x_t &= \gamma_1 v_t + e_{1t} - e_{1t-1} = \tilde{u}_t^x, \\ y_t &= \frac{\gamma_2}{\gamma_1} x_t - \frac{\gamma_2}{\gamma_1} e_{1t} + e_{2t} = \frac{\gamma_2}{\gamma_1} x_t + \tilde{u}_t^y, \end{aligned} \tag{3}$$

where $E(\tilde{u}_t^x \tilde{u}_t^y) \neq 0$. Because of the absence of exogeneity between x_t and y_t, OLS estimation is, in principle, sub-optimal asymptotically regardless of the choice of the normalization.

The above parameterization of DGP1 shows that it belongs to the class of triangular models analyzed in Phillips (1991)

$$\begin{aligned} &\text{DGP2: } \Delta x_t = u_{1t}, \quad u_{1t} = e_{1t} + \theta_x e_{1t-1} + \sigma_{12} e_{2t}, \\ &\qquad y_t = \beta x_t + u_{2t}, \quad u_{2t} = e_{2t} + \theta_y e_{2t-1} + \sigma_{21} e_{1t}, \end{aligned} \tag{4}$$

where $E(u_{1t}u_{2t})$ may or may not be zero.

3.2. Simulation results

We begin our analysis with DGP1. Suppose it is known that there is a cointegrating vector in the bivariate system. Substituting out the common trend in x_t and y_t of DGP1, we can either write

$$y_t = \beta_x x_t + e_{2t} - \beta_x e_{1t}, \quad \beta_x = \gamma_2/\gamma_1, \tag{5}$$

or

$$x_t = \beta_y y_t + e_{1t} - \beta_y e_{2t}, \quad \beta_y = \gamma_1/\gamma_2. \tag{6}$$

Since standard asymptotic results show T consistency for the coefficients of both regressions, (5) and (6) form equally legitimate basis for estimating the cointegrating vector.

We simulate DGP1 with $\sigma_v^2 = 1$. The coefficient γ_2 is set to one and we vary γ_1 over the range 0.01 and 5. Table 3 reports the simulation results for sample sizes of 50, 200 and 500. Turning first to the results from regressing x_t on y_t (Table 3a), we see that the true regression coefficient, $\gamma_1 = \beta_y$, is precisely estimated over the range of parameter values considered and the accuracy of the

Table 3a

Regression: $x_t = c + \beta_y y_t + u_t$, $\beta_y = \gamma_1, \gamma_2$

DGP: $\mu_t = \mu_{t-1} + v_t$,

$x_t = \gamma_1 \mu_t + e_{1t}$,

$y_t = \gamma_2 \mu_t + e_{2t}$,

$(v, e_1, e_2) \sim (0, I)$, $\gamma_2 = 1$.

$\gamma_1 = \beta_y$ / T	0.01	0.05	0.10	0.20	0.50	1.0	1.5	2.0	5.0
Mean OLS estimates (1000 simulations)									
50	0.004	0.043	0.088	0.169	0.418	0.840	1.268	1.686	4.243
200	0.009	0.047	0.095	0.190	0.476	0.953	1.425	1.898	4.755
500	0.010	0.049	0.098	0.196	0.490	0.979	1.469	1.958	4.898
Empirical distribution of t_β									
$T = 50$									
$t_{0.05}$	1.738	1.760	1.756	−2.016	2.859	−3.848	−4.154	4.407	4.864
$t_{0.95}$	1.550	1.554	1.380	1.186	0.533	−0.127	−0.226	0.430	−0.536
$T = 200$									
$t_{0.05}$	−1.611	−1.917	1.857	−2.216	−3.004	−3.961	−4.526	−4.777	−5.103
$t_{0.95}$	1.617	1.409	1.412	1.127	0.534	0.059	−0.260	−0.054	−0.705
$T = 500$									
$t_{0.05}$	−1.763	1.747	−1.892	−2.273	3.112	−4.252	−4.671	5.089	−5.231
$t_{0.95}$	1.562	1.473	1.378	1.090	0.360	−0.120	−0.460	0.516	−0.779

Table 3b
Regression: $y_t = c + \beta_x x_t + u_t$, $\beta_x = \gamma_2/\gamma_1$

γ_1	0.01	0.05	0.10	0.20	0.50	1.0	1.5	2.0	5.0
β_x	100	20	10	5	2	1	0.75	0.5	0.2
T									
Mean OLS estimates (1000 simulations)									
50	0.094	0.423	0.762	1.101	1.219	0.842	0.613	0.477	0.199
200	0.332	1.503	2.189	2.479	1.679	0.948	0.651	0.494	0.200
500	0.858	3.208	3.939	3.483	1.853	0.979	0.660	0.497	0.200
Empirical distribution of t_β									
$T = 50$									
$t_{0.05}$	357.93	73.005	35.902	18.762	7.244	−3.572	−2.667	2.266	−1.724
$t_{0.95}$	75.979	−15.571	7.952	−4.487	1.719	0.001	0.772	1.111	1.528
$T = 200$									
$t_{0.05}$	−385.23	79.416	40.356	−20.474	−8.360	−4.088	−2.744	−2.315	−1.686
$t_{0.95}$	−74.954	−16.280	9.111	−5.515	−1.985	−0.156	0.592	1.071	1.607
$T = 500$									
$t_{0.05}$	−407.73	−78.834	−41.303	−20.529	8.447	−4.390	−2.905	2.314	−1.747
$t_{0.95}$	−77.677	−16.788	−10.571	−6.043	2.110	−0.104	0.676	0.986	1.500

estimates increases as the sample size increases. There is one feature of the results that is noteworthy. Since by construction, y_t is not weakly exogenous for x_t, the least-squares estimator for β_1 and all test statistics associated with it should have nonstandard limiting distributions. Upon examining the upper and lower 5% critical values of the empirical distribution of the t-statistic on $\hat{\beta}_y$, we find that for large values of γ_1, the empirical distribution of the t-statistic is indeed being shifted to the left of the normal distribution. However, as the value of γ_1 falls, the distribution approaches normality. In the extreme case when $\gamma_1 = 0.01$, the upper and lower 5% critical values are practically the same as those from the normal distribution, even with a moderate sample size of 200.

The picture is very different when we regress y_t on x_t (see Table 3b). The true value of β_x is $1/\gamma_1$, which we also report for convenience. For large values of γ_1, $\hat{\beta}_x$ tends to be downward biased when $T = 50$, but the estimates are reasonably accurate. The precision of the estimates starts to deteriorate when γ_1 falls below unity. When $\gamma_1 = 0.2$, the least-squares estimator is severely biased downwards. For example, the mean of $\hat{\beta}_x$ is only 1.101 when $\beta_x = 5$ at $T = 50$. Although the bias is reduced as T increases, there is still a substantial discrepancy between the true value of 5 and the average estimated value of 3.483 at $T = 500$. Even though $\hat{\beta}_x$ is downward biased, one might expect $\hat{\beta}_x$ to at least increase with the true coefficient, β_x. However, the simulations reveal that as $\gamma_1 \to 0$ and hence $\beta_x \to \infty$, $\hat{\beta}_x$ tends towards 0 rather than increases with β_x. Curiously, the t-statistic associated with $\hat{\beta}_x$ appears to diverge to $-\infty$ as $\gamma_1 \to 0$ but approaches normality as γ_1 increases.

The above simulation results clearly illustrate the fact that the choice of the regressand can severely affect the precision of the estimates. How do these results relate to our empirical example of the Fisher equation? We want to suggest that x_t should be treated as inflation and y_t as the interest rate. We also want to suggest that the estimate from a regression of π on i is to be trusted. To justify these interpretations, we first note that regressions of y_t on x_t become problematical when γ_1 is small. Also recall that DGP1 implies that Δx_t is an MA(1) process with parameter θ_x, where

$$\frac{\theta_x}{(1+\theta_x^2)} = \frac{\sigma_1^2}{\gamma_1^2\sigma_\varepsilon^2 + 2\sigma_1^2}. \tag{7}$$

Since $\gamma_1 = \gamma_1^* \sigma_1$, $\gamma_1 \to 0$ if either $\gamma_1^* \to 0$, or if $\sigma_1^2 \to \infty$. The former corresponds to the case when the common trend, μ_t, is a weak driving force of x_t. The latter corresponds to the case of large variability in the idiosyncratic noise of x_t, so large that their impact dominates that of the stochastic trend in x_t.

The value of γ_1 affects the econometric analysis involving x_t because it follows from (7) that as $\gamma_1 \to 0$, $\theta_x \to -1$. Such a process, referred to as *nearly integrated nearly white noise* by Nabeya and Perron (1994), has a strong tendency to be mean reverting. This feature is inherent in our inflation series. As mentioned

earlier, a moving average coefficient at lag 3 of −0.8 is found in the inflation series. The issue, of course, is not so much the size of the MA coefficient on a particular lag, but the sum of the coefficients at all lags. Although DGP1 allows only an MA(1) in the noise function of Δx_t, it nevertheless encompasses the feature that is of interest.[2]

One way to judge if the simulated values of x_t indeed behave like a nearly integrated nearly white noise process is to examine the size of the unit root tests. The parameter space of interest is when γ_1 falls short of unity. A γ_1 of 0.5 translates into a moving-average coefficient in Δx_t of −0.6. Previous work by Schwert (1989) and Agiakloglou and Newbold (1992) have documented that many unit root tests suffer from size distortions even when θ is −0.5. One exception is $MZ_{\alpha\mu}$, which is more robust to negative serial correlation. At $\gamma_1 = 0.5$, the exact size of $Z_{\alpha\mu}$ is 0.58, of $MZ_{\alpha\mu}$ and $t_{\rho\mu}$ is 0.10, when the nominal size is 0.05. When $\gamma_1 = 0.2$, the exact size of the tests are 0.98, 0.19, and 0.26, respectively. The size discrepancy amongst the tests increase as γ_1 becomes smaller.

If large negative residual serial correlation is indeed the reason why the estimates have properties that depend on the normalization, then we should observe similar results from DGP2 given by (4). We simulate DGP2 letting the noise function in Δx_t be a moving average process. Given that $y_t = \beta x_t + u_{2t}$, it follows that $\Delta y_t = \beta \Delta x_t + \Delta u_{2t}$. Since u_{2t} is itself a stationary moving-average process, Δu_{2t} is overdifferenced. Thus, if Δx_t has a negative moving-average component, Δy_t will inherit a moving average component that implies an even stronger tendency for mean reversion. In other words, the moving-average component in y_t is more negative than that in x_t. In light of the results from DGP1, one might then expect that if the data were generated by DGP2, using y_t as the regressand will give more precise estimates because it has a smaller spectral density at frequency zero than x_t.

Fig. 1 provides a summary of the results for DGP2 assuming $\sigma_{12} = \sigma_{21} = 0$ for the case $\beta = 1$. Least-squares regressions of y_t on x_t give very accurate estimates regardless of the values of θ_x and θ_y. However, when we use x_t as the regressand, the estimates, while little affected by the values of θ_y, are severely downward biased when θ_x is negative and the biases are larger the closer θ_x is to −1. For example, when $\theta_x = -0.8$, β is estimated to be below 0.5 when we use x_t as the regressand, half the true value.

The picture that emerges from the simulations of the two DGPs points to the following general observation. If the first difference of one of the series has negative serial correlation (or a small spectral density at frequency zero),

[2] Perron (1994) suggests that a negative moving-average component should be present in the inflation series if the monetary authorities react to offset inflationary/disinflationary pressures that are inconsistent with an inflation target for the path of the price level. This makes inflation strongly mean reverting.

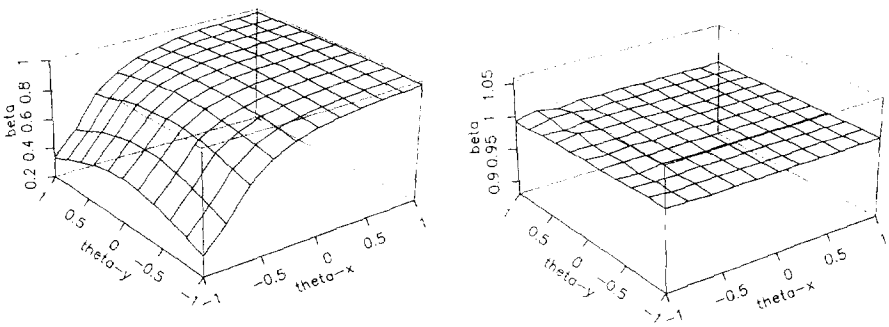

Fig. 1. Average values of the estimates with data from DGP2 with $\sigma_{12} = \sigma_{21} = 0$ and $\beta = 1$: (a) from a regression of x on y; (b) from a regression of y on x.

estimates based upon an equation with that series as the regressand always have better properties than those normalized on the other variable. The next section provides a theoretical rationale for this result.

4. Local asymptotic analyses

This section uses local asymptotic analyses to explain why the static least-squares estimator experiences substantial downward bias for certain normalizations only. The first subsection analyzes DGP1, and the second subsection focuses on DGP2.

4.1. DGP1: γ_1 local to 0

A notable feature of the simulation results reported for DGP1 is that the least squares estimator is severely downward biased when γ_1 is small and when we use y_t as the regressand. We therefore use the following parameterization

$$\gamma_1 = c/\sqrt{T}, \quad c \neq 0,$$

where we recall that DGP1 is the unobserved components model given by (1). The parameter γ_1 tends to zero as $T \to \infty$ at rate $\sqrt{T}$. We also recall that $\gamma_1 \to 0$ if the noise to signal ratio is large ($\sigma_1 \to \infty$) or if the common trend component is small ($\gamma_1^* \to 0$). This in turn implies that the moving average component (θ_x) in (2) is

$$\theta_x = -1 + \delta/\sqrt{T} \to -1 \quad \text{as } T \to \infty,$$

for some noncentrality parameter $\delta > 0$. This local parameterization of θ_x has been used in Nabeya and Perron (1994) and Perron and Ng (1996) to analyze the local asymptotic properties of $I(1)$ processes with MA(1) noise functions. Extending the 'nearly integrated nearly white noise' terminology in the univariate

case to the present multivariate model, y_t and x_t can be said to have a 'nearly cointegrated nearly unbalanced' relationship when γ_1 is local to zero. We begin our analysis with the following lemma.

Lemma 1 (Sample moments). *Let x_t and y_t be generated by* (1) *with $\gamma_1 = c/\sqrt{T}$. Let $W_v(r)$ be a Wiener process defined on $C[0,1]$. Then as $T \to \infty$:*

(i) $T^{-1}\sum_{t=1}^{T} x_t^2 \Rightarrow c^2\sigma_v^2 \int_0^1 W_v(r)^2\,\mathrm{d}r + 1$;

(ii) $T^{-2}\sum_{t=1}^{T} y_t^2 \Rightarrow \gamma_2^2\sigma_v^2 \int_0^1 W_v(r)^2\,\mathrm{d}r$;

(iii) $T^{-3/2}\sum_{t=1}^{T} y_t x_t \Rightarrow c\gamma_2\sigma_v^2 \int_0^1 W_v(r)^2\,\mathrm{d}r$.

The proof to the lemma is standard and is omitted. Part (ii) of the lemma is the usual result for an $I(1)$ process and follows from the fact that y_t is invariant to local variations in γ_1. However, this is not the case with x_t, which is a white noise process in the limit and $\sum_{t=1}^{T} x_t^2$ is $O_p(T)$. A consequence of this slower rate of normalization is that the sample moment for x_t^2 is influenced by σ_1^2 (normalized to 1) in the limit. The properties of the least-squares estimator from the two normalizations are then immediate from Lemma 1. These results are summarized in the following theorem.

Theorem 1. Let x_t and y_t be generated by (1) *with $\gamma_1 = c/\sqrt{T}$. Let $\hat{\beta}_y$ be the least-squares estimate from a regression of x_t on y_t, and let $\hat{\beta}_x$ be the least-squares estimate from a regression of y_t on x_t. Let $W_1(r)$ be a Wiener process independent of $W_v(r)$. As $T \to \infty$.*

From a regression of x_t on y_t with $\beta_y = \gamma_1/\gamma_2$.

1. $T^{1/2}\hat{\beta}_y \Rightarrow c/\gamma_2 = T^{1/2}\beta_y$;
2. $T(\hat{\beta}_y - \beta_y) \Rightarrow \dfrac{\int_0^1 W_v(r)\,\mathrm{d}W_1(r)}{\gamma_2\sigma_v \int_0^1 W_v(r)^2\,\mathrm{d}r}$;
3. $t_{\hat{\beta}_y} \Rightarrow \mathrm{N}(0,1)$.

From a regression of y_t on x_t with $\beta_x = \gamma_2/\gamma_1$.

1. $T^{-1/2}\hat{\beta}_x \Rightarrow \dfrac{c\gamma_2\sigma_v^2 \int_0^1 W_v(r)^2\,\mathrm{d}r}{c^2\sigma_v^2 \int_0^1 W_v(r)^2\,\mathrm{d}r + 1}$;
2. $T^{-1/2}(\hat{\beta}_x - \beta_x) \Rightarrow \dfrac{-\gamma_2/c}{c^2\sigma_v^2 \int_0^1 W_v(r)^2\,\mathrm{d}r + 1}$;
3. $T^{-1/2}t_{\hat{\beta}_x} \Rightarrow \dfrac{-1}{c\sigma_v\left(\int_0^1 W_v(r)^2\,\mathrm{d}r\right)^{1/2}}$.

Remark:

1. The main result of the theorem is that in this local framework, $\hat{\beta}_y$ is consistent but $\hat{\beta}_x$ is not. Absent the influence of σ_1^2 (normalized to 1) on the sample

moment of x_t^2, $T^{-1.2}\hat{\beta}_x$ would have converged to $\gamma_2 c$, the true regression coefficient. However, the presence of $\sigma_1^2 = 1$ in the denominator of the estimator induces a downward bias to the estimator. This explains the simulation results that regressions with y_t as the regressand yield estimates that can be far below the true value.

2. Note that c appears in the numerator and the denominator of the limiting distribution of the normalized least-squares estimator when y_t is the regressand. For this reason, and as seen from Table 3b, $\hat{\beta}_x \to 0$ when c is very large or when c is small.
3. In the standard asymptotic framework with γ_1 fixed, the distribution of $t_{\hat{\beta}_x}$ is nonstandard since weak exogeneity between x_t and y_t is not satisfied in the regressions. Obtaining asymptotic normality of the test statistics is the motivation for the modifications to $t_{\hat{\beta}_x}$ introduced by Phillips and Hansen (1990). However, as stated in the theorem, $\hat{\beta}_x$ is asymptotically mixed normal and $t_{\hat{\beta}_x}$ is asymptotically standard normal even though the exogeneity assumption is violated in finite samples. The reason for this result is that as $\beta_x \to 0$, y_t becomes weakly exogenous for the innovations in x_t in the regression with x_t as the regressand. However, this is not so when y_t is used as the regressand. In such cases, the regression residuals have a nontrivial correlation with the innovations driving x_t, and in consequence, the t statistic has nonstandard properties even when suitably standardized.
4. The theorem suggests that it is desirable from the point of view of both estimation and hypothesis testing to use as regressand the variable that is 'less integrated', in other words, to use the variable whose spectral density at frequency zero is the smallest. There are two intuitive reasons why this works well. The first is that using the 'more integrated' variables as regressors amounts to putting the variables with more variability on the right hand side of the regression. As is well known, the greater are the variations in the regressors, the more precise are the parameter estimates. The second rationale can be seen with reference to the two regression specifications given in (5) and (6). Given that $\gamma_1 \to 0$, $\beta_x \to \infty$ and $\beta_y \to 0$ asymptotically, the regression error in (5) with y_t as the regressand has a variance that diverges at the same rate as γ_1 approaches 0. The regression noise in (6) with x_t as the regressand is, on the other hand, invariant to γ_1 and has finite variance in the limit. The choice of normalization is obvious viewed in this light.

4.2. DGP2: θ_x local to -1

Recall that DGP2 is the triangular representation of a cointegrated system:

$$\Delta x_t = u_{1t}, \quad u_{1t} = e_{1t} - \theta_x e_{1t-1} + \sigma_{12} e_{2t};$$

$$y_t = \beta x_t + u_{2t}, \quad u_{2t} = e_{2t} + \theta_y e_{2t-1} + \sigma_{21} e_{1t}.$$

To rationalize the results of Fig. 1, our local asymptotic framework lets $\theta_x \to -1$ in the limit with

$$\theta_x = -1 + \delta/\sqrt{T}, \quad \delta > 0.$$

A distinct feature of DGP2 is that σ_{12} and σ_{21} may or may not be zero. As we will now show, the properties of the least squares estimator depend on these quantities.

Lemma 2 (Sample moments). *Let x_t and y_t be generated by* (4) *and let $W_1(r)$ and $W_2(r)$ be independent Wiener processes defined on the space $C[0,1]$.*

If $\sigma_{12} \neq 0$:

1. $T^{-2}\sum_{t=1}^{T} x_t^2 \Rightarrow \sigma_{12}^2 \int_0^1 W_2(r)^2\,\mathrm{d}r$;
2. $T^{-2}\sum_{t=1}^{T} y_t^2 \Rightarrow \beta^2\sigma_{12}^2 \int_0^1 W_2(r)^2\,\mathrm{d}r$;
3. $T^{-2}\sum_{t=1}^{T} x_t u_{2t} \to 0$;
4. $T^{-2}\sum_{t=1}^{T} x_t y_t \Rightarrow \beta\sigma_{12}^2 \int_0^1 W_2(r)^2\,\mathrm{d}r$.

If $\sigma_{12} = 0$:

1. $T^{-1}\sum_{t=1}^{T} x_t^2 \Rightarrow \sigma_1^2 + \delta^2\sigma_1^2 \int_0^1 W_1(r)^2\,\mathrm{d}r$;
2. $T^{-1}\sum_{t=1}^{T} y_t^2 \Rightarrow \beta^2\sigma_1^2(1 + \delta^2 \int_0^1 W_1(r)^2\,\mathrm{d}r) + (2\beta\sigma_{21} + \sigma_{21}^2)\sigma_1^2 + (1+\theta_y^2)\sigma_2^2$;
3. $T^{-1}\sum_{t=1}^{T} x_t u_{2t} \to \sigma_{21}\sigma_1^2$;
4. $T^{-1}\sum_{t=1}^{T} x_t y_t \Rightarrow \beta\sigma_1^2(1 + \delta^2 \int_0^1 W_1(r)^2\,\mathrm{d}r) + \sigma_{21}\sigma_1^2$.

Results for the case $\sigma_{12} = 0$ are straightforward applications of the results in Nabeya and Perron (1994). Of note is that the rates of normalization for the case $\sigma_{12} \neq 0$ are higher than for the case $\sigma_{12} = 0$. The reason is that x_t is driven by two partial sums when $\sigma_{12} \neq 0$: one relating to its own innovations, and one relating to the innovations in the y_t process. Therefore, even though the partial sum of innovations in x_t induces mean reversion, this effect is dominated by the stochastic trend consisting of integrated innovations in y_t. Accordingly, x_t behaves like a strictly $I(1)$ variable when $\sigma_{12} \neq 0$. The properties of the least-squares estimator can now be summarized in the following theorem.

Theorem 2. Let x_t and y_t be generated by (4). *Let $\hat{\beta}_y$ be the least-squares estimator of $\beta_y = 1/\beta$ from a regression of x_t on y_t, and let $\hat{\beta}_x$ be the least squares estimator of $\beta_x = \beta$ from a regression of y_t on x_t.*

If $\sigma_{12} \neq 0$:

$$T(\hat{\beta}_x - \beta_x) \Rightarrow$$

$$\frac{\{\sigma_{21}\sigma_1^2 + \sigma_{12}\sigma_2^2(1+\theta_y)[1 + \int_0^1 W_2(r)\,\mathrm{d}W_2(r)] + \sigma_{12}\sigma_{21}\sigma_1\sigma_2 \int_0^1 W_2(r)\,\mathrm{d}W_1(r)\} + [\sigma_2^2(1+\theta_y^2) + \sigma_{21}^2\sigma_1^2]/\beta}{\beta^2\sigma_{12}^2\sigma_2^2 \int_0^1 W_2(r)^2\,\mathrm{d}r}$$

$$T(\hat{\beta}_x - \beta_x) \Rightarrow \frac{\sigma_{21}\sigma_1^2 + \sigma_{12}\sigma_2^2(1+\theta_y)[1+\int_0^1 W_2(r)\,\mathrm{d}W_2(r)] + \sigma_{12}\sigma_{21}\sigma_1\sigma_2\int_0^1 W_2(r)\,\mathrm{d}W_1(r)}{\sigma_{12}^2\sigma_2^2\int_0^1 W_2(r)^2\,\mathrm{d}r}.$$

If $\sigma_{12}=0$:

$$\hat{\beta}_y \Rightarrow \frac{\beta(1+\delta^2\int_0^1 W_1(r)^2\,\mathrm{d}r)+\sigma_{21}}{\beta^2(1+\delta^2\int_0^1 W_1(r)^2\,\mathrm{d}r)+(2\beta\sigma_{21}+\sigma_{21}^2)+(1+\theta_y^2)\sigma_2^2/\sigma_1^2} \neq 1/\beta \equiv \beta_y;$$

$$\hat{\beta}_x \Rightarrow \beta + \frac{\sigma_{21}}{(1+\delta^2\int_0^1 W_1(r)^2\,\mathrm{d}r)} \neq \beta \equiv \beta_x.$$

Remarks:

1. The issue of normalization is irrelevant when $\sigma_{12}\neq 0$ as far as consistency of the least-squares estimator is concerned. The intuition is essentially that both x_t and y_t have the same order of variability as standard $I(1)$ variables in the limit. In particular, both variables are dominated by the stochastic trend built upon innovations in y_t, even though the DGP looks superficially as though innovations in x_t were the source of variation. Although cross correlation among variables of a cointegrated system induces nuisance parameters which invalidate classical inference, $\sigma_{12}\neq 0$ has the unusual but desirable effect of retaining the $I(1)$ nature of the series, allowing rate T convergence to apply to this local asymptotic framework. Note, however, that neither $T(\hat{\beta}_y-\beta_y)$ nor $T(\hat{\beta}_x-\beta_x)$ is mixture normal, both are noncentrally located, and are influenced by all the nuisance parameters in the model. Accordingly, although both estimators are super-consistent, test statistics associated with them cannot be used for inference.
2. In a regression of y_t on x_t with $\sigma_{12}=0$, $\hat{\beta}_y-\beta=\beta_y$ only if $\sigma_{21}=0$. We are back to the result of the standard $I(0)$ asymptotic framework, where the least square estimator is susceptible to simultaneity bias. The result stated in the above theorem generalizes to cases where the innovations are correlated at different lags, and the extent of the bias is independent of the value of β. When $\sigma_{12}=\sigma_{21}=0$, both y_t and x_t behave like stationary processes in the limit. Accordingly, the least-squares estimator does not converge at the fast rate of T as would be the case with strictly integrated variables, but at the slower rate of $\sqrt{T}$ as in the case of stationary variables.
3. From a regression of x_t on y_t, $\hat{\beta}_y$ does not converge to $1/\beta\equiv\beta_y$ even if $\sigma_{21}=0$. The reason is that in this case, y_t inherits the nearly integrated nearly white noise property of x_t, and its sample moments also require a smaller rate of normalization. Stationary innovations in y_t which would otherwise have converged to zero in the standard asymptotic framework have a nontrivial effect in this local asymptotic framework, as seen from the sample moment of y_t^2 in Lemma 2. This in turn generates a downward bias on the least-squares estimator. In general, the bias in $\hat{\beta}_y$ is a function of σ_{21} and the true value of

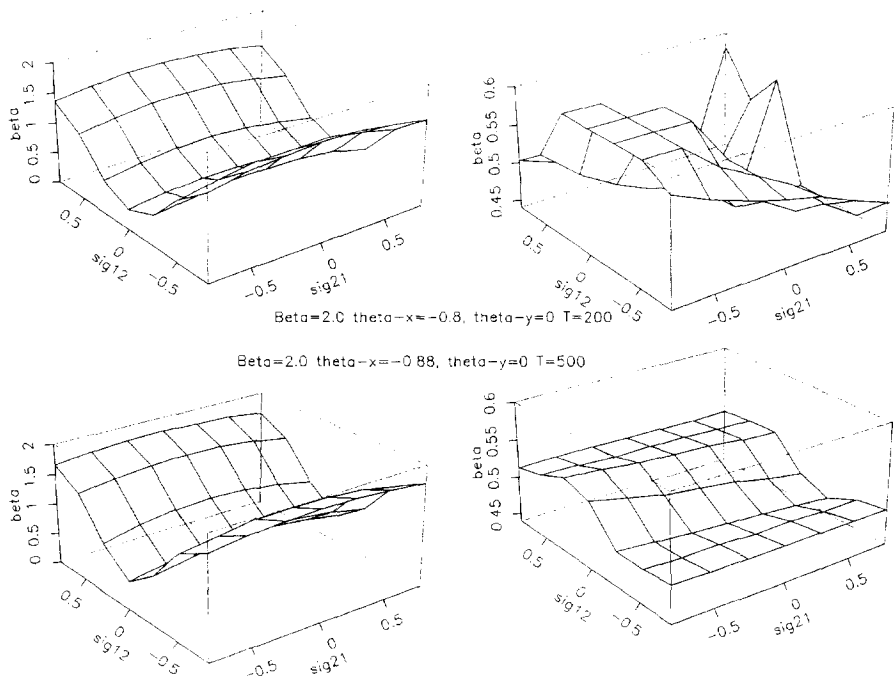

Fig. 2. Average values of the estimates with data from DGP2 with $\beta = 2$: (a) top graphs are with $\theta_x = -0.8$, $\theta_y = 0$ and $T = 200$ with results for a regression of x on y on the left and of y on x on the right; (b) bottom graphs are with $\theta_x = -0.88$, $\theta_y = 0$ and $T = 500$ with results for a regression of x on y on the left and of y on x on the right.

β_x. However, it can be shown that holding σ_{21} fixed, the larger is β, the more precise are the estimates $\hat{\beta}_y$ and $\hat{\beta}_x$ for β_y and β_x irrespective of the value of σ_{21}. Simulations confirm this to be the case.

4. An implication of Theorem 2 is that there is a discontinuity in the limiting distribution of the least-squares estimator at $\sigma_{12} = 0$. However, as we show in Fig. 2, there is substantial leakage around $\sigma_{12} = 0$ in finite samples in the sense that when σ_{12} is greater than but close to zero, the precision of the estimates are still affected. This is so even when the sample size is as large as 500.
5. From a practical standpoint, normalizing the cointegrating regression on the variable with the smallest spectral density at frequency zero (in this case y_t) is still the preferred choice. When $\sigma_{12} \neq 0$, the issue of normalization is irrelevant and hence the prescription can do no worse than alternative normalizations. When $\sigma_{12} = \sigma_{21} = 0$, using y_t as the regressand is the only normalization that can yield asymptotically unbiased estimates. When $\sigma_{12} = 0$ but $\sigma_{21} \neq 0$, the estimates can be asymptotically unbiased and efficient when y_t

is the regressand if the regression error is made uncorrelated with the regressors. Estimators such as the DOLS and the FM-OLS serve this purpose. However, σ_{21} is not the only source of bias in a regression with x_t as the regressand. Thus, even fully efficient estimators cannot remove the bias in the estimates.

6. DGP1 is a special case of DGP2 with $\sigma_{12}=0$ and $\sigma_{21}\neq 0$ for a given sample size. It would seem, by implication of Theorem 2, that the static least-squares estimator will be biased whether normalized on x_t or y_t. However, Theorem 1 suggests that a consistent estimator can be obtained when normalized on x_t. To understand these apparently contradicting results, we need to clarify the assumptions underlying the two local asymptotic models. In the first model, parameterizing γ_1 to be local to zero has two effects. It induces a (noninvertible) negative moving average component to the noise function of x_t, and it changes the correlation between the regression innovation and the regressor. As mentioned earlier, the regression innovations with x_t as the regressand are weakly exogenous for y_t in the limit. Under DGP2, θ_x is parameterized to be local to -1 holding σ_{12} and σ_{21} fixed. The results of the two theorems therefore hold under similar but not identical conditions.

5. Residual-based tests for cointegration

Residual-based tests for cointegration are valid in the standard asymptotic framework because the least-squares estimator is super-consistent. However, to the extent that the choice of normalization affects the properties of the least-squares residuals, the size and power of tests of cointegration tests might also be affected. Using the least-squares residuals implied by the cointegrating vectors reported in Section 3, we construct Z_α, MZ_α, and t_ρ, to test for the presence of a unit root. The overall impression from the simulations, available upon request, is that the statistics generally have good power in rejecting the null hypothesis of no cointegration. The one exception is DGP1 when γ_1 is small. When the residuals are taken from a regression of y_t on x_t, the unit root hypothesis is under rejected, especially when we use the tests more robust to negative moving average errors, namely, MZ_α and t_ρ. When the residuals are taken from a regression of x_t on y_t, all tests have high power but the statistic Z_α, and to some extent t_ρ, suffer from size distortions. Hence, the practical recommendation is to test for cointegration using MZ_α and the residuals from a regression which uses the variable with the smallest spectral density at frequency zero as the regressand. We now provide an intuitive explanation for the results.

The least-squares residuals under DGP1 with y_t as the regressand [see (5)] are

$$\hat{u}_{xt}=(\beta_x-\hat{\beta}_x)x_t+e_{2t}-\beta_x e_{1t}, \tag{8}$$

where $\beta_x = \gamma_2/\gamma_1$. Theorem 1 shows that $\hat{\beta}_x$ is an inconsistent estimator for β_x if $\gamma_1 \to 0$ as T increases. Then

$$\Delta\hat{u}_{xt} = (\beta_x - \hat{\beta}_x)\gamma_1 v_t + \Delta e_{2t} - \hat{\beta}_x \Delta e_{1t}$$

is a stationary and invertible process even as T increases. Accordingly, the least-squares autoregression in $\hat{u}_{xt}$ will yield a regression coefficient of one. This evidence of a unit root will lead us to conclude that there is no cointegration even though x_t and y_t shares the common trend μ_t.

The least-squares residuals from a regression with x_t as the regressand are similarly defined as

$$\hat{u}_{yt} = (\beta_y - \hat{\beta}_y)y_t + e_{1t} - \beta_y e_{2t}, \tag{9}$$

with $\beta_y = \gamma_1/\gamma_2$. Since $\hat{\beta}_y$ is a consistent estimator for β_y, $\sum_{t=1}^{T} \hat{u}_{yt}$ is $O_p(T^{1/2})$, which is indicative of a process that is stationary or $I(0)$. An autoregression in $\hat{u}_{xt}$ will yield a least-squares coefficient that is away from one in the limit, and the null hypothesis of no cointegration is rejected. However, under the null hypothesis of no cointegration, the presence of a large negative moving average error in x_t makes Z_α, and to some extent t_ρ, subject to severe size distortions.

Under DGP2, the regression residuals are

$$\hat{u}_{xt} = (\beta_x - \hat{\beta}_x)x_t + u_{2t} \tag{10}$$

when normalized on y_t, and

$$\hat{u}_{yt} = (\beta_y - \hat{\beta}_y)y_t - \beta_y u_{2t} \tag{11}$$

when normalized on x_t. Given super-consistency of $\hat{\beta}_x$ and $\hat{\beta}_y$ when $\sigma_{12} \neq 0$, both series have partial sums that are $O_p(T^{1/2})$. Unit root tests on these residuals are then consistent following the analysis of Phillips and Ouliaris (1990). When $\sigma_{12} = 0$, the residuals-based cointegration tests will also reject the null hypothesis of no cointegration. The reason, in this case, is that although the least-squares estimator is biased and inconsistent, both x_t and y_t are stationary series in the limit. The two sets of residuals defined above have partial sums that are $O_p(T^{1/2})$, and standard tests once again reject the null hypothesis of a unit root as confirmed by our simulations results. In this setting, the least size distortions occur using MZ_α in a regression of y_t on x_t, and therefore the practical recommendation remains the same as stated above for DGP1.

6. Some observations on alternative estimators

Asymptotically efficient methods of estimating cointegrating vectors have been proposed in the literature. Does least-squares bias of the type examined here extends to these estimators? This section sheds some light on this issue.

6.1. Least-squares-based methods

As the OLS estimator is known to be asymptotically inefficient, we also consider some least-squares-based methods that are asymptotically efficient to ensure that the aforementioned results are not specific to the least-squares estimator. The FM-OLS and CCR estimates of $\partial i/\partial \pi$ from the Fisher equation are 0.61 and 0.97 when normalized on i, but are 0.75 and 0.81 when normalized on π. These are in the same range as the estimates presented in Table 1, suggesting that the normalization problem also applies to the FM-OLS and the CCR.

To further investigate the properties of the estimators, we simulate DGP1 defined in (1) and DGP2 defined in (4) using the same parameterizations as presented in Tables 3a and 3b. We also vary the truncation lag of the kernel estimator in the case of the FM-OLS and the CCR,[3] and the number of leads and lags of first differences of the regressor in the case of the DOLS. While these fully efficient estimates are closer to the true values for both normalizations than the OLS estimates, the estimates remain noticeably inferior when normalized in one direction. We report results in Table 4a for DGP1 with $\gamma_1 = 0.2$, and DGP2 for $\sigma_{12} = 0$ and $\sigma_{12} = 0.5$, both with $\beta = 1$. Evidently, there is a 'good' and a 'bad' normalization, the latter being y_t on x_t in the case of DGP1 and x_t on y_t in the case of DGP2.

An observation of note from Table 4a is that the FM-OLS and the CCR estimates cannot be improved upon by varying the choice of truncation lag and of the kernel itself. However, the DOLS estimates for the 'bad' normalization can be made more precise when the lag length is sufficiently large. It can be shown that increasing k in the DOLS regression 'fixes up' the moment matrix of the regressors as indicated in Lemma 1. Interestingly, the DOLS estimates remain superior to the OLS estimates, but not because of satisfying valid conditioning that the estimator was designed for.

While increasing k removes some of the biases in the DOLS estimator when y_t is used as the regressand, it does not restore normality of the associated t-statistic as can be seen from Table 4b. The upper and lower 5% critical values are still far from those of the normal distribution even as k increases. The departure from normality of the t-statistic associated with the CCR and the FM-OLS from the 'bad' normalization is even more apparent. Thus, from the point of view of inference, it remains preferable to use the variable with the smallest spectral density at frequency zero as the regressand.

Our focus has been on least-squares-based estimators with the regressand expressed in level form. It is important to emphasize that our recommendation of

[3] In these simulations, we use the Coint 2.0 package of Ouliaris and Phillips (1994) running under Gauss 3.2. The Parzen window is used in place of the Quadratic window as it provides the flexibility to set the kernel weights outside the bandwidth to zero. The two kernels give similar results when evaluated at the same bandwidth.

Table 4a
Average estimates of some asymptotically efficient estimators (1000 simulations)

		$k=2$			$k=4$			$k=8$		
		DOLS	CCR	FMOLS	DOLS	CCR	FMOLS	DOLS	CCR	FMOLS
DGP1	$\beta_X=.5$	4.198	3.428	2.713	4.530	3.569	2.581	4.757	3.575	2.529
$\gamma_1=0.2$	$\beta_Y=.2$	0.198	0.196	0.191	0.199	0.198	0.190	0.201	0.199	0.192
DGP2	$\beta_X=1$	0.995	0.994	0.996	0.998	0.999	1.000	0.999	0.998	0.999
$\sigma_{12}=0$	$\beta_Y=1$	0.863	0.766	0.678	0.907	0.785	0.662	0.951	0.796	0.661
$\theta_X=-0.8$										
DGP2	$\beta_X=1$	1.010	1.005	1.007	1.003	1.006	1.008	1.006	1.008	1.011
$\sigma_{12}=0.5$	$\beta_Y=1$	0.962	0.935	0.931	0.931	0.943	0.940	0.996	0.941	0.938
$\theta_X=0.8$										

Subroutines from Coint 2.0 are used to produce the CCR and FMOLS estimates, using the Parzen kernel as the option.

Table 4b
Upper and lower 5% of the t-statistic

		$k=2$						$k=4$						$k=8$					
		DOLS		CCR		FMOLS		DOLS		CCR		FMOLS		DOLS		CCR		FMOLS	
DGP1	$\beta_X=.5$	−15.6	0.35	−12.94	−2.07	−16.17	−2.21	−11.69	2.26	9.12	−1.41	−12.21	−1.34	−9.86	5.47	−7.36	−0.98	−10.30	−0.92
$\gamma_1=0.2$	$\beta_Y=0.2$	−1.80	1.66	−1.94	1.39	1.95	1.39	−1.66	1.61	1.89	1.64	−1.96	1.63	−1.79	1.83	−1.94	1.72	−2.00	1.72
DGP2	$\beta_X=1$	−1.79	1.54	−1.85	1.64	−1.93	1.68	−1.87	1.73	−1.87	1.77	−2.06	1.91	−1.78	1.68	−1.84	1.73	1.98	1.89
$\sigma_{21}=0$	$\beta_Y=1$	−6.49	−0.02	−9.35	−1.82	−12.06	2.45	−4.91	0.87	−8.25	1.31	−11.34	−1.84	3.71	1.66	−7.23	−0.85	10.30	1.30
$\theta_X=-0.8$																			
DGP2	$\beta_X=1$	−0.40	0.98	1.02	1.41	−0.99	1.53	−0.48	0.63	−0.93	1.38	−0.88	1.58	−0.47	0.54	−0.92	1.67	0.84	1.93
$\sigma_{21}=0.5$	$\beta_Y=1$	−4.99	0.12	5.52	−0.57	−6.45	−0.35	−2.26	0.49	−4.40	0.0002	−5.21	−0.19	−1.20	0.78	−4.13	−0.29	5.21	−0.18
$\theta_X=0.8$																			

Notes: Subroutines from Coint 2.0 are used to produce the CCR and FMOLS estimates, using the Parzen kernel as the option. The autoregressive spectral density estimator with two lags is used to construct the long run variance of the DOLS regression error. For DOLS, k is the number of leads and lags of the first differences of the regressors added to the least squares regression. For CCR and FMOLS, k is the bandwidth used in kernel-based estimations.

using as regressand the variable whose spectral density of the differenced series is smallest is based on analyses of regression models of this class. There are evidently other methods of estimating cointegrating vectors. For example, the regressand can be expressed in first-differenced form as in the case of 'nonlinear ECM'.[4] Evaluating the many alternative estimators is outside the scope of the present analysis, and we have not explored whether our criteria for choosing the regressand will generalize. We have nevertheless considered a method of estimating cointegrating vectors outside an OLS based framework, and to this we now turn.

6.2. *The Johansen method*

For most methods of estimating cointegrating vectors, 'normalization' is taken to mean the choice of the regressand. An exception is the reduced-rank regression approach of Johansen (1991), wherein normalization is taken to mean the choice of the length of the eigenvectors. This is sometimes referred to as an empirical normalization. The Johansen procedure does not pretest for the presence of a unit root, and it analyzes all the variables as a system. Because of these fundamental differences with other estimators that belong to the LAMN class, it is of interest to ask if the Johansen approach is rid of the problem being analyzed.

Simulations are again conducted for all parameterizations of DGP1 reported in Table 3 and of DGP2 for combinations of the parameters, while varying the number of lagged first differences that enter the reduced rank regressions. In each experiment, we report the averaged values for the pair of eigenvectors (normalized on x_t), tabulate the frequency distribution of the cointegrating rank (r) chosen by the Trace and λ-max statistics, and record the upper and lower 5 critical values of the Wald-type statistic for testing if the estimated cointegrating vector equals the null value. In Table 5, we only report results with $\gamma_1 = 0.2$ for DGP1, and $\theta_x = -0.8$ for DGP2. Suffice it to mention that results for values of $\gamma_1 > 0.5$ and $\theta_x > -0.5$ accord with theory, in the sense that the tests have exact size close to the nominal size and the cointegrating vector is well estimated.

Of interest from the results is that in all cases, the cointegrating vector is very precisely estimated if we condition on $r = 1$. Phillips (1994) noted some advantages of maximum likelihood estimation of a triangular system over reduced rank regressions. His results, however, are based on the assumption that $\theta_x = \sigma_{12} = \sigma_{21} = 0$ in our DGP2. Compared with the estimates reported earlier, our results suggest that when the error structure is more complex, as are the cases considered here, there may be unexplored benefits in using reduced rank regressions over estimation of the triangular system (such as by FM-OLS). A systematic analysis of the two estimation methods is to be left for future investigations.

[4] See Gonzalo (1994) for a review of these methods.

Table 5
Statistics related to the Johansen procedure (1000 simulations)

	$k=2$	$k=4$	$k=8$
A. The pair of estimated eigenvectors (normalized on x)			
DGP1 $\beta_x=1$; $\gamma_1=\beta_y=0.2$	(1, −0.199); (1, 5.862)	(1, −0.198); (−2.720)	(1, −0.201); (1, −6.424)
DGP2 $\beta_x=\beta_y=1$ $\sigma_{21}=0$; $\theta_x=-0.8$;	(1, −1.1181); (1, 0.004)	(1, −1.016); (1, 0.044)	(1, 1.102); (1, 0.169)
DGP2 $\beta_x=\beta_y=1$ $\sigma_{21}=0.5$; $\theta_x=-0.8$;	(1, −1.0); (1, −0.310)	(1, −1.0); (1, −0.947)	(1, 1.0); (1, 1.547)
B. The 5% and 95% points of the Wald test for estimated $\hat{\beta}_y-\beta_y$			
DGP1 $\beta_x=1$; $\gamma_1=\beta_y=0.2$	(0.0048, 4.52)	(0.0042, 5.53)	(−0.0081, 7.71)
DGP2 $\beta_x=\beta_y=1$ $\sigma_{21}=0$; $\theta_x=-0.8$;	(0.0090, 10.99)	(0.0059, 9.92)	(0.0089, 9.03)
DGP2 $\beta_x=\beta_y=1$ $\sigma_{21}=0.5$; $\theta_x=-0.8$	(0.0013, 2.41)	(0.0018, 2.25)	(0.0034, 3.46)

See Johansen (1991, p. 1564) for a description of the test. The χ^2-statistic has one degree of freedom and lower and upper 5% points of 0.0039 and 3.84.

Table 5 Continued.

	$k=2$		$k=4$		$k=8$	
	Trace	Max λ	Trace	Max λ	Trace	Max λ
C. Frequency that the trace and max λ statistics conclude that $r=1$						
DGP1 $\beta_x=1$; $\gamma_1=\beta_y=0.2$	0.801	0.801	0.883	0.883	0.878	0.882
DGP2 $\beta_x=\beta_y=1$ $\sigma_{21}=0$; $\theta_y=-0.8$;	0.051	0.051	0.362	0.362	0.769	0.770
DGP2 $\beta_x=\beta_y=1$ $\sigma_{21}=0.5$; $\theta_y=0.8$	0.547	0.547	0.814	0.814	0.890	0.890

The 90 percent critical values from Osterwald-Lenum (1992) are 15.66 (for $r=0$) and 6.5 (for $r=1$) for the Trace statistic. The critical values are 12.91 and 6.5 for the Max λ statistic. In no case did the tests conclude $r=0$. Hence the frequency that r is chosen to be 2 can be obtained as one minus the entry.

The results in Table 5 also reveal that the Wald test tends to overreject the hypothesis that the cointegrating vector equals its null value. The problem is reminiscent of the size distortions in the t-test associated with the three fully efficient estimators when the 'bad' normalization is used. Using a test with a nominal size of 10%, the frequency that the Trace statistic concludes $r=1$ is between 0.8 and 0.9 for DGP1, and also for DGP2 with $\sigma_{21}\neq 0$, noting in passing that r is never chosen to be zero. In such cases, the finite sample size of the test seems acceptable. However, for DGP2 with $\sigma_{12}=0$, the size of the statistics is distorted and is more so the shorter the number of lags. With 4 lags, the statistics report $r=1$ at a frequency of only 0.4, and finds $r=2$ in the remaining cases. The exact size of the test remains above 50% for a nominal size of 10% when $\sigma_{12}=0.25$. This is the leakage problem around $\sigma_{12}=0$ discussed earlier.

A finding that $r=2$ in these bivariate DGPs is confounding because if there are n $I(1)$ series, there cannot be more than $n-1$ cointegrating vectors. A finding that $r=2$ implies that both series are stationary. Some intuition for this result can be gained by examining the properties of the second eigenvector. Consider DGP2 with $\sigma_{12}=0$, the case when both series are strongly mean reverting. As seen from Table 5, the first eigenvector is the correct estimate of the cointegrating vector. The second is the unit vector that selects x_t. Since y_t is weakly exogenous for x_t by construction, the eigenvector optimally puts zero weight on the redundant regressor. Simulations of more complex parameterizations of DGP1 also found the statistics to under report $r=1$ when the second eigenvector is $(1,0)$. In such cases, applying a zero weight on the decisively $I(1)$ process will also give a stationary combination of the variables, although the combination is economically uninteresting. Whether the statistics will select one or two cointegrating vectors will depend on how strong are the unit root components relative to the mean-reverting components, and on the causal structure underlying the variables. In general, the stronger is the unit root component in both series, the less chance there is of finding two cointegrating vectors and vice versa.

Do size distortions in testing for the cointegrating rank arise in practice? Returning to the Fisher equation example, the Trace and λ–max statistics are 29.17 and 22.03 for the null hypothesis that $r=0$ against the alternative $r=1$, and 7.15 for both statistics for the hypothesis that $r=1$ against the alternative that $r=2$. While we can decisively reject the absence of any cointegrating vector, we reject the presence of one cointegrating vector in favor of two cointegrating vectors at the 10% level but not at the 5% level. The $r=2$ scenario is therefore of empirical relevance. However, if one had pretested for the presence of a unit root, one would have dismissed the possibility that $r=2$ since it implies both variables are $I(0)$. One would then obtain the unique cointegrating vector (normalized on π) of $[1,-0.87]$, in line with the fully efficient least-squares estimates based on the π normalization. This, however, is based upon results from pretesting for a unit root using a test robust to serial correlation. In this regard, $MZ_{y\mu}$ still plays a useful role.

7. Concluding comments

Estimations of cointegrating vectors are by now standard practice in dynamic analysis of time series data. This paper questions two conventional wisdoms on least-squares estimation of cointegrating vectors. First, we show that although the least-squares estimator achieves strong consistency in the standard asymptotic framework, it is not so when a regressor has a large negative moving-average component that is modeled as local to -1. Second, we show that normalization in one direction can yield estimates that are consistent but not the other. These results extend to least squares based fully efficient estimators.

Using an unobserved components framework, we show that a negative moving-average component can arise when a variable in the cointegrating system has a weak correlation with the common trend, or when the idiosyncratic noise of the variable dominates its total asymptotic variance. Such a variable has properties of an $I(1)$ process, but also exhibits a strong tendency to revert to its mean. For this reason, standard tests have trouble determining whether or not a unit root is present in the series. Gali (1992) presented evidence for a negative moving component in the noise function of many macroeconomic time series, and Vogelsang (1994) discussed how negative moving average errors can arise as a result of additive outliers. The issues analyzed in the present paper are therefore not merely of theoretical interest, but are also issues of practical concern.

Our theoretical and empirical investigations lead to the following practical guideline. The cointegration analyses should begin with tests for the presence of a unit root in the variables, preferably using a test that is robust to negative MA errors such as the MZ_α. Step two is to rank the variables by their estimated normalized spectral density at frequency zero, preferably using estimators that are more efficient than those used in the unit root tests. For instance, one can use a kernel-based estimator with the demeaned first-differences of the data, or an autoregressive spectral density estimator based on the autoregression without the lagged level. Step three is to choose the variable with the smallest estimated normalized spectral density at frequency zero as the regressand. The steps proposed applies to other least-squares-based estimators of the cointegrating vector.

There are other instances when a regression can be nearly unbalanced. For example, if the error process of an $I(1)$ series has a large autoregressive root, so that the series is close to being twice integrated. Accordingly, its spectral density at frequency zero is very large. Although we have not provided a theoretical analysis to such cases, simulations reveal that the least-squares estimator is also more biased when normalized in one direction. However, using the variable with the smallest spectral density at frequency zero as the regressand still gives more precise estimates.

Our focus has been on estimating a single cointegrating vector. As discussed in Hargreaves (1994), ordinary least squares (and its more efficient variants) can also be used to estimate multiple cointegrating vectors from a single equation

with multiple regressors. In that methodology, the second through r cointegrating vectors are constructed to be orthogonal to the first. The choice of the regressand is even more important in that context because all cointegrating vectors are normalized on the same regressand. Our recommendation of comparing the values of the spectral density function at the origin will be especially useful.

References

Agiakloglou, C., Newbold, P., 1992. Empirical evidence on Dickey–Fuller type tests. Journal of Time Series Analysis 13, 471–483.

Andrews, D.W.K., 1991. Heteroskedastic and autocorrelation consistent matrix estimation. Econometrica 59, 817–854.

Clark, P.K., 1987. The cyclical component of U.S. economic activity. Quarterly Journal of Economics 102, 798–814.

Gali, J., 1992. How well does the IS-LM model fit postwar U.S. data? Quarterly Journal of Economics 107, 709–738.

Gonzalo, J., 1994. Five methods to estimate cointegrating vectors. Journal of Econometrics 60, 203–234.

Hargreaves, C., 1994. A review of methods of estimating cointegrating relationships. in: Hargreaves, C. (Ed.), Nonstationary time series analysis and cointegration. Oxford University Press, Oxford.

Johansen, S., 1991. Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models. Econometrica 59, 1551–80.

Nabeya, S., Perron, P., 1994. Local asymptotic distribution related to the AR(1) model with dependent errors. Journal of Econometrics 62, 229–264.

Ng, S., Perron, P., 1995. Unit root tests in ARMA models with data dependent methods for the selection of the truncation lag. Journal of the American Statistical Association 90, 268 281.

Osterwald-Lenum, M., 1992. A note with quantiles of the asymptotic distribution of the maximum likelihood cointegration rank test statistics. Oxford Bulletin of Economics and Statistics 54, 461–472.

Ouliaris, S., Phillips, P.C.B., 1994. COINT 2.0: GAUSS Procedures for Cointegrated Regressions (Predicta Software).

Park, J.Y., 1992. Canonical cointegrating regressions. Econometrica 60, 119–143.

Perron, P., 1994. Nonstationarities and nonlinearities in Canadian inflation. In: Economic behaviour and policy choice under price stability. Bank of Canada, Ottawa.

Perron, P., 1996. The adequacy of asymptotic approximations in the near-integrated autoregressive model with dependent errors. Journal of Econometrics 70, 317–350.

Perron, P., Ng, S., 1996. Useful modifications to unit root tests with dependent errors and their local asymptotic properties. Review of Economic Studies 63, 435–465.

Phillips, P.C.B., 1991. Optimal inference in cointegrated systems. Econometrica 59, 283–306.

Phillips, P.C.B., 1994. Some exact distribution theory for maximum likelihood estimators of cointegrating coefficients in error correction models. Econometrica 62, 73–94.

Phillips, P.C.B., Hansen, B.E., 1990. Statistical inference in instrumental variables regression with $I(1)$ processes. Review of Economic Studies 57, 99–125.

Phillips, P.C.B., Ouliaris, S., 1990. Asymptotic properties of residual based tests for cointegration. Econometrica 58, 165–193.

Phillips, P.C.B., Perron, P., 1988. Testing for a unit root in time series regression. Biometrika 75, 335–346.

Press, W.H., Teukolsky, S., Vetterling, W., Flannery, B., 1992. Numerical recipies in C, 2nd ed. Cambridge University Press, Cambridge.

Said, S.E., Dickey, D.A., 1984. Testing for unit roots in autoregressive-moving average models of unknown order. Biometrika 71, 599–607.

Saikkonen, P., 1991. Asymptotically efficient estimation of cointegrating regressions. Econometric Theory 7, 1–21.

Schwert, G.W., 1989. Tests for unit roots: A Monte Carlo investigation. Journal of Business and Economic Statistics 7, 147–160.

Stock, J.H., 1987. Asymptotic properties of least squares estimators of cointegrating vectors. Econometrica 55, 1035–1056.

Stock, J.H., 1990. A class of tests for integration and cointegration. Manuscript, Kennedy School of Government, Harvard University, Cambridge, MA.

Stock, J.H., Watson, M.W., 1993. A simple estimator of cointegrating vectors in higher order integrated systems. Econometrica 61, 783–820.

Vogelsang, T.J., 1994. On testing for a unit root in the presence of additive outliers. Working paper 94-30, Center for Analytic Economics, Cornell University, Ithaca, NY.