# Is Double Trouble?
# How to Combine Cointegration Tests[*]

Christian Bayer[†] and Christoph Hanck[‡]

IGIER, Università Commerciale L. Bocconi
and
Technische Universität Dortmund

April 1, 2008

**Abstract**

This paper suggests a combination procedure to exploit the imperfect correlation of cointegration tests to develop a more powerful meta test. To exemplify, we combine Engle and Granger (1987) and Johansen (1988) tests. Either of these underlying tests can be more powerful than the other one depending on the nature of the data-generating process. The new meta test is at least as powerful as the more powerful one of the underlying tests irrespective of the very nature of the data generating process. At the same time, our new meta test avoids the size distortion inherent in separately applying multiple tests for cointegration to the same data set.

KEYWORDS: *Cointegration, Meta Test, Multiple Testing*

JEL-Codes: *C12, C22*

## 1 Introduction

Testing for cointegration has become one of the standard tools in applied economic research. Various tests have been suggested for this purpose, most of which are implemented in standard econometric packages and hence are easily available nowadays. Well-known examples include the residual-based tests of Engle and Granger (1987) and

[*] Research supported by DFG under Sonderforschungsbereich 475.

[†] IGIER, Università Commerciale L. Bocconi, Via Salasco 5, 20136 Milano, Italy. Tel.: +39 02 5836 3386. email: christian.bayer@unibocconi.it.

[‡] Universität Dortmund, Vogelpothsweg 78, 44221 Dortmund, Germany. Tel. +49 (0)231 755 3127, christoph.hanck@uni-dortmund.de.

Phillips and Ouliaris (1990), or the system-based tests of Johansen (1988, 1991). This regularly forces the applied researcher to select from the test decisions of the various applicable procedures. Often one test rejects the null hypothesis whereas another test does not, making it unclear how to interpret test outcomes then. More generally speaking, the $p$-values of different tests are typically not perfectly correlated (Gregory *et al.*, 2004).

Because of the imperfect correlation, it is problematic to choose, for example, a testing strategy that relies on the test that achieves the smallest $p$-value. Such strategy is not suitable to decide whether or not the time series under investigation are cointegrated. It will not control the probability of rejecting a true null hypothesis at some chosen level $\alpha$ because it ignores the multiple testing nature of the problem. More specifically, using the test with the smallest $p$-value will lead to an oversized test.

It has thus been suggested that the significance level of the tests should be adjusted downwards when running more than one cointegration test. One classical solution to the problem is the Bonferroni procedure which compares the $p$-values of $N$ tests with the more challenging cut-off value of $\alpha/N$. Unfortunately, this test procedure—while able to remove the size distortion—has low power. From this line of argument one might view the imperfect correlation of different test statistics mostly as a problem.

However, an imperfect correlation of test statistics also implies that one test contains information that the other one is not exploiting. Hence, we may view the imperfect correlation of underlying tests as beneficial instead. This leads us to propose an aggregation procedure to combine different underlying tests in a meta test that potentially yields an improvement in power.

One approach popular in meta analysis to combine tests is Fisher's 1970 approach which found its application in econometrics in panel-unit root tests for example (Maddala and Wu, 1999). Fisher-type tests are traditionally used to combine results from one test on different and independent samples. This corresponds to the fact that the distribution of the Fisher test as originally proposed applies only to independent test statistics. This rules out using a standard Fisher test in a setup where we want to combine correlated underlying tests of one hypothesis on a single sample. Yet, methods to deal with the issue of correlated test statistics have been developed recently (Hartung, 1999), so that correlation itself is no longer an insuperable obstacle to meta testing. The challenge has instead become to estimate the correlation structure of the test statistics. We propose a bootstrap method to carry out this estimation.

We hence exploit recent advances in meta analysis in order to provide valid inference on cointegration when several underlying tests are available. In particular, we exemplify our test using a combination of an Engle and Granger (1987) cointegration test with a Johansen (1988) maximum eigenvalue test for cointegration rank.

The proposed bootstrap method for estimating the correlation structure of the underlying tests also yields a second version of the meta test that relaxes certain assumptions required for Hartung's method. Both versions of our meta test successfully control the level $\alpha$ of the test and are at the same time powerful.

In particular, we demonstrate that our meta test is as powerful as the more powerful one of the underlying tests, which amongst them can each be more powerful than the other one, depending on the true data-generating process. Consequently the meta test provides a test of non cointegration with attractive power properties across a wide range of relevant data-generating processes. The test can hence be viewed as selecting the more powerful of the underlying tests in a fully data driven fashion. At the same time, the test avoids the size distortion associated with multiple testing that arises when separately employing several underlying tests. To the best of our knowledge, this is the first time that a practical approach is put forward to combine different tests of a given hypothesis applied to a single sample.

The remainder of this paper is organized as follows: Section 2 describes our test procedure and Section 3 gives setup and results of our Monte Carlo experiments. Section 4 revisits a set of cointegration studies to provide an empirical application of our cointegration test. Finally Section 5 concludes.

# 2 Test Procedure

## 2.1 Setup

Let $\mathbf{x}_t = (x_{1t}, \ldots, x_{Kt})' \in \mathbb{R}^K$ be a vector of stochastic variables integrated of order one, $I(1)$. The stochastic vector $\mathbf{x}_t$ is said to be cointegrated if there exists at least one $\boldsymbol{\alpha} \in \mathbb{R}^K$, $\boldsymbol{\alpha} \neq \mathbf{0}$, such that $z_t = \boldsymbol{\alpha}'\mathbf{x}_t$ is a stationary $I(0)$ process. Suppose we have observations $\mathbf{x}_0, \ldots, \mathbf{x}_T$.

We are concerned with the following null hypothesis:

$H_0$ : There exists no cointegrating relationship among the variables in $\mathbf{x}_t$.

against the alternative hypothesis

$H_1$ : There exists at least one $\boldsymbol{\alpha} \neq \mathbf{0}$ such that $z_t = \boldsymbol{\alpha}'\mathbf{x}_t$ is $I(0)$.

The literature has suggested various test procedures to discriminate between these two hypotheses. Well-known examples include the residual-based tests of Engle and Granger (1987) and Phillips and Ouliaris (1990), or the system-based tests of Johansen (1988, 1991).

For the Engle-Granger test, one computes the $t$-statistic of $\gamma-1$ in the OLS regression

$$\Delta\hat{u}_t = (\gamma - 1)\hat{u}_{t-1} + \sum_{p=1}^{P} \nu_p \Delta\hat{u}_{t-p} + \epsilon_t. \tag{1}$$

Here, $\hat{u}_t$ is the usual residual from a first stage OLS regression of one of the $x_{kt}$, $k = 1, \ldots, K$, on the remaining elements of $\mathbf{x}_t$ (and appropriate deterministic terms). The sum $\sum_{p=1}^{P} \nu_p \Delta\hat{u}_{t-p}$ captures residual serial correlation. Alternatively, one could control for serial correlation by the semiparametric approach of Phillips and Ouliaris (1990).

The system-based tests of Johansen (1988) test the presence of $h$ cointegrating relationships by estimating the number of significantly non-zero eigenvalues of the matrix $\hat{\mathbf{\Pi}}$ estimated from the Vector Error Correction Model (VECM)

$$\Delta\mathbf{x}_t = \mathbf{\Pi}\mathbf{x}_{t-1} + \sum_{p=1}^{P} \mathbf{\Gamma}_p \Delta\mathbf{x}_{t-p} + \boldsymbol{\mu}_0 + \boldsymbol{\epsilon}_t. \tag{2}$$

The actual tests are either the $\lambda_{\text{trace}}$-test with test statistic

$$\lambda_{\text{trace}}(h) = -T \sum_{j=h+1}^{K} \ln(1 - \hat{\pi}_j) \tag{3}$$

or the $\lambda_{\max}$-test with test statistic

$$\lambda_{\max}(h) = -T \ln(1 - \hat{\pi}_{h+1}). \tag{4}$$

Here, $\hat{\pi}_j$ denotes the $j$th largest eigenvalue of $\hat{\mathbf{\Pi}}$.

## 2.2 Exploiting Imperfect Correlation between Cointegration Tests

As Gregory *et al.* (2004) show, the $p$-values that correspond to the above test statistics are only weakly correlated, in particular when comparing residual-based and system-based tests.[1] As we argued in the introduction, this means that a more powerful test can

[1] In unreported simulations, we find that under the null hypothesis, the correlation of probits of the $p$-values (see below) is about .55 for $\lambda_{\text{trace}}$ and Engle-Granger tests.

in principle be achieved by exploiting the imperfect correlation of suitably transformed test statistics. The actual test that we propose is based on Hartung's (1999) method to combine dependent test statistics.

Let $\xi_i$ be the test statistics of a test $i = 1, \ldots, N$ of a set of cointegration tests (e.g. the ones discussed above) and $\Xi_i$ its asymptotic distribution function under the null hypothesis. Under the null, the integral transformation of the test statistic, $\Xi_i(\xi_i)$, yields a uniformly distributed random variable on the unit interval. This variable closely corresponds to the $p$-values of the test, which are defined as $p_i = \Xi_i(\xi_i)$ if the test rejects for small values of $\xi_i$ and $p_i = 1 - \Xi_i(\xi_i)$ if the test rejects for large values of $\xi_i$. Based on these $p$-values, we can define a probit representation of the test as $\Phi^{-1}(p_i) =: t_i$, where $\Phi$ is the cumulative distribution function of the standard normal distribution.

Let $\mathbf{t} = (t_1, \ldots, t_N)'$. Then, asymptotically, the components of $\mathbf{t}$ are marginally standard normal under the null. Hartung (1999) now, as highlighted by Demetrescu *et al.* (2006), additionally makes the auxiliary assumption that $\mathbf{t}$ is jointly normally distributed, denoted $\mathbf{t} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$. Under this assumption, we have $\boldsymbol{\iota}'\mathbf{t} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\iota}'\mathbf{\Sigma}\boldsymbol{\iota})$, where $\boldsymbol{\iota} = (1, \ldots, 1)'$. This leads to a standardized meta test statistic,

$$\tau = \frac{\boldsymbol{\iota}'\mathbf{t}}{\sqrt{\boldsymbol{\iota}'\mathbf{\Sigma}\boldsymbol{\iota}}}.$$

The statistic $\tau$ follows a standard normal distribution under $H_0$ and the auxiliary assumption of joint normality. Of course, there is no a priori reason to justify joint normality of $\mathbf{t}$ in the case of cointegration tests that we consider. Fortunately, Demetrescu *et al.* (2006) demonstrate that Hartung's procedure can be fruitfully applied even if the assumption is not met.

As a practical requirement, we need a feasible consistent estimator of $\mathbf{\Sigma}$. If the number of tests $N$ is small, there is no hope to estimate $\mathbf{\Sigma}$ meaningfully from the realizations of $\mathbf{t}$. This is so even if one is willing to assume constant correlation of $t_i, t_j$ as in Hartung (1999), which in any case would not be a sensible assumption in our setting. Instead, we rely on a bootstrap method to estimate $\mathbf{\Sigma}$.

Thus, we require a method to bootstrap cointegration tests under the null hypothesis. Such a bootstrap procedure has recently been proposed by Swensen (2006). In brief, Swensen's procedure resamples residuals from an estimated VECM representation of the data-generating process (DGP) to then generate integrated but non-cointegrated time series. From the resulting bootstrap distribution of the test statistic, we estimate the correlation matrix of $\mathbf{t}$.

More specifically, we use the following algorithm.

**Algorithm 1** :

1. *Estimate the unrestricted VECM*

$$\Delta \mathbf{x}_t = \mathbf{\Pi} \mathbf{x}_{t-1} + \sum_{p=1}^{P} \mathbf{\Gamma}_p \Delta \mathbf{x}_{t-p} + \boldsymbol{\mu}_0 + \boldsymbol{\epsilon}_t \tag{5}$$

*to obtain coefficient estimates* $\hat{\boldsymbol{\mu}}_0, \hat{\mathbf{\Pi}}, \hat{\mathbf{\Gamma}}_p$ *and residuals* $\hat{\boldsymbol{\epsilon}}_t$.[2]

2. *Check whether the system has no explosive root, i.e. whether* $\|z\| \geq 1$*, by solving* $\det[\hat{\mathbf{A}}(z)] = 0$*, where*

$$\hat{\mathbf{A}}(z) = (1-z)\mathbf{I}_K - \hat{\mathbf{\Pi}} z - \hat{\mathbf{\Gamma}}_1 (1-z) z - \cdots - \hat{\mathbf{\Gamma}}_P (1-z) z^P.$$

3. *If so, draw B series of pseudo errors* $\left\{\boldsymbol{\epsilon}^*_{t,b}\right\}_{t=P+1,\dots,T}^{b=1,\dots,B}$ *by resampling non-parametrically with replacement from the residuals* $\left\{\hat{\boldsymbol{\epsilon}}_t\right\}_{t=P+1,\dots,T}$.

4. *With these pseudo errors, construct B series of pseudo observations* $\mathbf{x}^*_{t,b}$ *from*

$$\Delta \mathbf{x}^*_{t,b} = \sum_{p=1}^{P} \hat{\mathbf{\Gamma}}_p \Delta \mathbf{x}^*_{t-p,b} + \hat{\boldsymbol{\mu}}_0 + \boldsymbol{\epsilon}^*_{t,b}.$$

*For the initial observations, set* $\mathbf{x}^*_{t,b} = \mathbf{x}_t, t = 0, \dots, P$.[3]

5. *Compute the test statistics* $\xi^*_{i,b}$ *for all pseudo samples* $b = 1, \dots, B$ *and all cointegration tests that are to be combined,* $i = 1, \dots, N$.

6. *Estimate the distribution function of the test statistic of each test as*

$$\Xi^*_i(x) = \frac{\#\left\{\xi^*_{i,h} \leq x | h = 1, \dots, B\right\}}{B}$$

*and calculate the corresponding p-values* $p^*_{i,b} = \Xi^*_i\left(\xi^*_{i,b}\right)$ *or* $1 - \Xi^*_i\left(\xi^*_{i,b}\right)$, *as appropriate. Correspondingly, calculate the p-values,* $p_i$*, of the test statistics on the original data by evaluating* $\Xi^*_i(\xi_i)$ *or* $1 - \Xi^*_i(\xi_i)$.

[2] As pointed out by Swensen (2006) one could alternatively estimate a restricted VAR in first differences to impose the null of no cointegration. However, as Paparoditis and Politis (2003) show for unit-root tests, imposing such a restriction may lead to a power loss.

[3] Since we require pseudo observations that are integrated but non-cointegrated, $\mathbf{\Pi} = \mathbf{0}$ is imposed

7. *Obtain the corresponding probit representation of each test statistic,* $t^*_{i,b} = \Phi^{-1}(p^*_{i,b})$*, stacked in* $\mathbf{t}^*_b = \left(t^*_{1,b}, \ldots, t^*_{N,b}\right)'$ *where* $\Phi^{-1}$ *is the quantile function of the standard normal distribution. Correspondingly, obtain* $t_i = \Phi^{-1}(p_i)$*.*

8. *Estimate the covariance matrix* $\mathbf{\Sigma}$ *of the probits of the tests by*

$$\mathbf{\Sigma}^* = \frac{1}{B} \sum_b \left(\mathbf{t}^*_b - \bar{\mathbf{t}}^*\right) \left(\mathbf{t}^*_b - \bar{\mathbf{t}}^*\right)',$$

*where* $\bar{\mathbf{t}}^* = \frac{1}{B} \sum_b \mathbf{t}^*_b$*.*

This Algorithm provides a feasible version of the test statistic $\tau$,

$$\tau^* = \frac{\boldsymbol{\iota}'\mathbf{t}}{\sqrt{\boldsymbol{\iota}'\mathbf{\Sigma}^*\boldsymbol{\iota}}},$$

where $\mathbf{t}$ is the probit representation of the bootstrap version of the underlying tests (see step 7 of Algorithm 1). We then reject $H_0$ at level $\alpha$ if $\tau^* < \Phi^{-1}(\alpha)$.

Note that $\tau^*$ will reject $H_0$ at least at level $\alpha$ if all underlying tests $t_i$ reject at level $\alpha$. This is so because $t_i < \Phi^{-1}(\alpha)$ for all $i = 1, \ldots, N$ implies

$$\tau^* = \frac{\boldsymbol{\iota}'\mathbf{t}}{\sqrt{\boldsymbol{\iota}'\mathbf{\Sigma}^*\boldsymbol{\iota}}} \leq \frac{\boldsymbol{\iota}'\mathbf{t}}{N} < \Phi^{-1}(\alpha),$$

since the entries of the positive semi-definite correlation matrix $\mathbf{\Sigma}^*$ are bounded by 1 and $-1$.

Swensen (2006) shows that his bootstrap procedure for the Johansen $\lambda_{\text{trace}}$ test, i.e. steps 1-6 in Algorithm 1, delivers a consistent estimate $\Xi^*_i$ of the distribution of the test statistic under the null hypothesis. It hence yields consistent estimates of $p$-values. The key element in Swensen's (2006) proposition is that the above bootstrap algorithm yields pseudo observations which have a representation asymptotically equivalent to the true DGP. Therefore, we expect Swensen's proposition to carry over to other tests for cointegration, in particular the ones we mentioned before. We corroborate this conjecture via extensive simulation in Section 3.

## 2.3 Alternative Formulation

Our Hartung-type test is a modification of the 'inverse normal' meta test (Stouffer *et al.*, 1949) robustified against dependence among the test statistics. Alternatively, we can formulate an analogous test that is more closely related to the meta test of Fisher

(1970), suitably modified to take dependencies between the test statistics into account. The advantage of this second test is that it does not rely on joint normality of $\mathbf{t}$. We keep from the Fisher test the aggregator of $p$-values

$$\chi = -2 \sum_i \ln (p_i) .$$

Of course we cannot invoke a $\chi^2 (2N)$ null distribution of $\chi$ as independency of the aggregated test statistics is necessary for this result. We therefore propose the following modification of Algorithm 1 to estimate the distribution of $\chi$ to account for dependency among the test statistics.

**Algorithm 2** :

*1. - 6. As in Algorithm 1.*

*7. Obtain the corresponding aggregate $\chi$ test statistic*

$$\chi_b^* = -2 \sum_i \ln \left(p_{i,b}^*\right) .$$

*8. Estimate the cumulative distribution function $\Theta$ of the $\chi_b^*$ by*

$$\Theta^* (x) = \frac{\# \{\chi_h^* \leq x | h = 1, \ldots , B\}}{B}.$$

This provides us with a dependency robust version of the Fisher test, where the $p$-values $p_i$ of the underlying tests are obtained as in step 6 of Algorithm 1. We calculate the final test statistic as

$$\chi^* = -2 \sum_i \ln (p_i)$$

and then reject $H_0$ at level $\alpha$ if $\chi^*$ exceeds the $(1 - \alpha)$-quantile of $\Theta^*$. This second test can be viewed as a distribution-free version of the test described in Algorithm 1.

# 3 Monte Carlo Experiments

## 3.1 Setup

Next, we study the properties of the proposed tests in a series of Monte Carlo experiments. As emphasized in the introduction, different tests for cointegration are likely to differ in their power against different points in the space of the alternative hypotheses. For example, Johansen's $\lambda_{\max}$ test can be expected to be relatively more powerful if the data is indeed generated by a finite order VECM with uncorrelated errors. Since our test combines information from tests that are powerful in different directions, a potential advantage of our testing strategy is that it should be more robust across different DGPs.

We therefore consider the following two alternative DGPs

$$\begin{aligned} \text{DGP(A):} \quad & \Delta \mathbf{x}_t = \mathbf{\Pi} \mathbf{x}_{t-1} + \mathbf{\Gamma} \Delta \mathbf{x}_{t-1} + \mathbf{u}_t \\ & \mathbf{\Gamma} = 0.2 \mathbf{I}_2 \end{aligned}$$

$$\begin{aligned} \text{DGP(B):} \quad & x_{1t} + \beta x_{2t} = z_{1t}, \ x_{1t} + \alpha x_{2t} = z_{2t} \\ & \beta = -2, \ \alpha = -1 \\ & z_{1t} = z_{1t-1} + u_{1t}, \ z_{2t} = \rho z_{2t-1} + u_{2t}. \end{aligned}$$

In both DGPs we set

$$\mathbf{u}_t = \begin{pmatrix} u_{1t} \\ u_{2t} \end{pmatrix} \overset{iid}{\sim} \mathcal{N}\left(0, \mathbf{I}_2\right).$$

These designs are widely used in Monte Carlo studies of cointegration tests. See for instance Engle and Granger (1987), Gonzalo (1994), Gregory *et al.* (2004), or Swensen (2006).

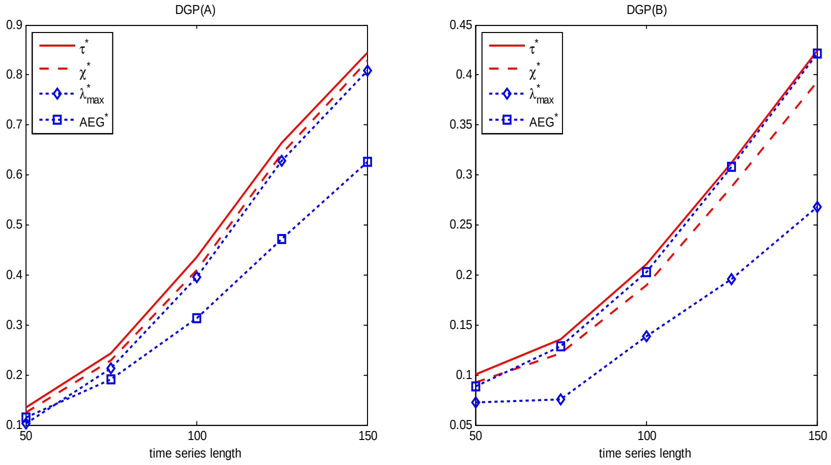
For DGP(A) the null hypothesis of no cointegration is obtained when $\mathbf{\Pi} = \mathbf{0}$, whereas we parameterize the alternative hypothesis of cointegration by $\mathbf{\Pi} = (1 \quad 0)' (.15 \quad -.15)$. For DGP(B), the null hypothesis is obtained when $\rho = 1$, whereas we parameterize the alternative hypothesis of cointegration by $\rho = 0.85$.[4]

For each DGP, we draw 5,000 replications under both the null and the alternative. We choose $T \in \{50, 75, 100, 125, 150\}$ as lengths of the time series. To mitigate the effect of initial conditions, we simulate each DGP for $T + 30$ time periods and discard the first 30 observations. For each replication, we compute the $\tau^*$ and the $\chi^*$ tests based on $B = 10,000$ bootstrap replications. As underlying tests we select Johansen's (1988) $\lambda_{\max}$ test and the augmented Engle and Granger (1987) residual-based test (AEG).

To investigate the relative performance of the new tests, we compare them to following alternative possibilities to test for cointegration: First, the standard augmented Engle and Granger (AEG) and Johansen $\lambda_{\max}$ tests, where we reject the null hypothesis

[4] Of course, Granger's representation theorem would allow us to write DGP(B) in a VECM form. However, error terms would be correlated, the matrix $\mathbf{\Pi}$ would have no rows of zeros under the alternative and $\mathbf{\Gamma}$ would equal $\mathbf{0}$.

Figure 1: Empirical power, DGP(A) and DGP(B), various $T$



See notes to Table 1

if the test statistics fall short of (respectively exceed) the level $\alpha$ critical value computed from the appropriate distribution of the tests.[5] Second, we investigate bootstrap versions of both tests (denoted in the following by AEG* and $\lambda^*_{\text{max}}$), which are by-products of our $\tau^*$ and $\chi^*$ tests. Third, we compute a 'naive' meta test based on the bootstrapped versions of the two underlying tests. This test rejects whenever at least one of the tests rejects. We call this test 'naive' because it ignores the multiple-testing nature of the problem. Studying this test hence reveals the size distortion incurred by selecting the most rejective test from a set of cointegration tests.

Implementation of the cointegration tests typically requires to select an order $\hat{P}$ of lagged differences to account for auto-correlation. In practice this is often done via some lag-length selection criterion, see e.g. Lütkepohl (2005). To reduce the computational burden we waive this option and use a constant order of $\hat{P} = 2$ throughout.

### 3.2 Results

Table 1 reports the empirical size of all tests at the level $\alpha$ of 5%.[6] The main findings may be summarized as follows. As expected, the 'naive' test is oversized and its size

[5] In the case of the AEG test we follow the standard practice of using MacKinnon (1996)-type critical values, which control for number of observations.

[6] We also ran all simulations described above at the 1% and 10% level, with qualitatively similar results. We also experimented with a version of DGP(A) with AR(1) error terms instead of white noise $\mathbf{u}_t$. Again, results are qualitatively similar. Tables with the additional results are available upon request.

Table 1: Empirical size

| | | Bootstrap tests | | | | | asymptotic tests | |
|---|---|---|---|---|---|---|---|---|
| DGP | $T$ | $\chi^*$ | $\tau^*$ | 'naive' | $\lambda^*_{\max}$ | AEG* | $\lambda_{\max}$ | AEG |
| (A) | 50 | 0.0684 | 0.0748 | 0.1026 | 0.0684 | 0.0534 | 0.093 | 0.0352 |
| | 75 | 0.0570 | 0.0636 | 0.0876 | 0.0530 | 0.0530 | 0.0888 | 0.0374 |
| | 100 | 0.0520 | 0.0562 | 0.0822 | 0.0546 | 0.0486 | 0.0882 | 0.0366 |
| | 125 | 0.0518 | 0.0598 | 0.0796 | 0.0490 | 0.0522 | 0.0854 | 0.0336 |
| | 150 | 0.0496 | 0.0564 | 0.0750 | 0.0492 | 0.0464 | 0.0806 | 0.0314 |
| (B) | 50 | 0.0598 | 0.0656 | 0.0932 | 0.0594 | 0.0542 | 0.0876 | 0.0346 |
| | 75 | 0.0558 | 0.0608 | 0.0830 | 0.0488 | 0.0536 | 0.0848 | 0.0356 |
| | 100 | 0.0492 | 0.0524 | 0.0816 | 0.0538 | 0.0490 | 0.0862 | 0.0322 |
| | 125 | 0.0466 | 0.0494 | 0.0746 | 0.0478 | 0.0450 | 0.0836 | 0.0294 |
| | 150 | 0.0482 | 0.052 | 0.0738 | 0.0504 | 0.0464 | 0.0798 | 0.0332 |

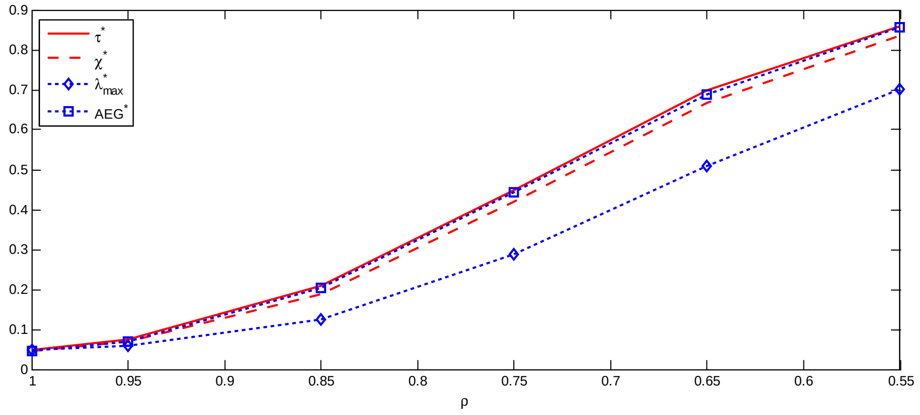
Average rejection rates at nominal level of 5%. 5,000 replications and 10,000 bootstrap replications. The tests $\tau^*$ and $\chi^*$ are described in Algorithms 1 and 2 respectively. AEG and $\lambda_{\max}$ refer to Engle and Granger (1987) and Johansen (1988) tests, AEG* and $\lambda^*_{\max}$ are their bootstrap counterparts. 'Naive' rejects when AEG* or $\lambda^*_{\max}$ rejects.

Table 2: Empirical power

| | | Bootstrap tests | | | | | asymptotic tests | |
|---|---|---|---|---|---|---|---|---|
| DGP | $T$ | $\chi^*$ | $\tau^*$ | 'naive' | $\lambda^*_{\max}$ | AEG* | $\lambda_{\max}$ | AEG |
| (A) | 0.1244 | 0.1342 | 0.1836 | 0.1032 | 0.1154 | 0.1074 | 0.1704 | 0.0712 |
| | 0.2284 | 0.2426 | 0.3106 | 0.2138 | 0.1908 | 0.1884 | 0.3116 | 0.1388 |
| | 0.4094 | 0.4352 | 0.5106 | 0.3964 | 0.3142 | 0.3182 | 0.5258 | 0.2436 |
| | 0.6424 | 0.6644 | 0.7174 | 0.6274 | 0.4714 | 0.4726 | 0.7408 | 0.3954 |
| | 0.8286 | 0.8458 | 0.8690 | 0.809 | 0.6264 | 0.6312 | 0.8884 | 0.5516 |
| (B) | 0.0922 | 0.101 | 0.1274 | 0.0722 | 0.0882 | 0.0746 | 0.1166 | 0.0596 |
| | 0.1216 | 0.1356 | 0.1586 | 0.0760 | 0.1288 | 0.1288 | 0.1414 | 0.0904 |
| | 0.1900 | 0.2108 | 0.2442 | 0.1390 | 0.2032 | 0.2006 | 0.2064 | 0.1500 |
| | 0.2884 | 0.3122 | 0.3442 | 0.1958 | 0.308 | 0.3068 | 0.2816 | 0.2382 |
| | 0.3932 | 0.4248 | 0.4558 | 0.2676 | 0.4212 | 0.4122 | 0.3872 | 0.3402 |

See notes to Table 1

Figure 2: Empirical power, DGP(B), $T = 100$, various $\rho$



See notes to Table 1

exceeds the nominal level by approximately 3 - 4 percentage points.[7] All other bootstrap tests control size reasonably well. The $\tau^*$ test (and to a lesser extent also the $\chi^*$ test) exhibits a slight upward size distortion for small $T$, partly due to a distortion of $\lambda^*_{\max}$ for small $T$. However, this size distortion vanishes for $T \geq 100$. In line with e.g. Swensen (2006), we find a more pronounced upward size distortion of the asymptotic $\lambda_{\max}$ test. By contrast, the asymptotic AEG test is slightly undersized.

Table 2 now reports the empirical power of all tests at the level $\alpha$ of 5%. Figure 1 summarizes the main information of Table 2 graphically. As expected, power increases in $T$ for all tests. While of the single tests the AEG* test is the most powerful single test for DGP (B), the $\lambda_{\max}$ and $\lambda^*_{\max}$ tests are most powerful for DGP(A).[8] This result may not entirely surprising, as both tests were originally designed having DGPs of type (A) and (B) respectively in mind.

The meta tests $\chi^*$ and $\tau^*$ both perform similarly and well, though $\tau^*$ has slightly higher power throughout. In particular, the $\tau^*$ test is somewhat more powerful than the most powerful single test for either DGP.

In addition to the Monte-Carlo experiments displayed in Tables 1 and 2, we run further experiments varying the degree of mean reversion of the cointegration error in DGP(B), i.e. distance of the alternative from the null. That is we choose $\rho \in \{0.95, 0.85, 0.75, 0.65, 0.55\}$. We focus on DGP(B) as the underlying tests are the more

[7] Note that this size distortion is very close to the one that can be inferred from Table I in Gregory *et al.* (2004).

[8] In judging the power of the asymptotic $\lambda_{\max}$ test versus its bootstrap counterpart, one needs to take into account the test's upwards size distortion. Consequently, its size-adjusted power is lower.

Table 3: Rejection rates when combining $N > 2$ tests

| | | Size | | | | Power | | | |
|---|---|---|---|---|---|---|---|---|---|
| DGP | $T$ | $\tau^*(2)$ | $\tau^*(4)$ | $\chi^*(2)$ | $\chi^*(4)$ | $\tau^*(2)$ | $\tau^*(4)$ | $\chi^*(2)$ | $\chi^*(4)$ |
| (A) | 50 | 0.0748 | 0.0750 | 0.0684 | 0.0706 | 0.1342 | 0.1336 | 0.1244 | 0.1272 |
| | 75 | 0.0636 | 0.0598 | 0.0570 | 0.0546 | 0.2426 | 0.2428 | 0.2284 | 0.2336 |
| | 100 | 0.0562 | 0.0526 | 0.0520 | 0.0504 | 0.4352 | 0.4376 | 0.4094 | 0.4216 |
| | 125 | 0.0598 | 0.0562 | 0.0518 | 0.0500 | 0.6644 | 0.6612 | 0.6424 | 0.6524 |
| | 150 | 0.0564 | 0.0536 | 0.0496 | 0.0502 | 0.8458 | 0.8428 | 0.8286 | 0.8340 |
| (B) | 50 | 0.0656 | 0.0650 | 0.0598 | 0.0598 | 0.1010 | 0.1042 | 0.0922 | 0.0942 |
| | 75 | 0.0608 | 0.0582 | 0.0558 | 0.0518 | 0.1356 | 0.1372 | 0.1216 | 0.1268 |
| | 100 | 0.0524 | 0.0502 | 0.0492 | 0.0484 | 0.2108 | 0.2200 | 0.1900 | 0.2028 |
| | 125 | 0.0494 | 0.0502 | 0.0466 | 0.0462 | 0.3122 | 0.3228 | 0.2884 | 0.3022 |
| | 150 | 0.0520 | 0.0506 | 0.0482 | 0.0472 | 0.4248 | 0.4338 | 0.3932 | 0.4098 |

Average rejection rates at nominal level of 5%. 5,000 replications and 10,000 bootstrap replications. The tests $\tau^*$ and $\chi^*$ are described in Algorithms 1 and 2 respectively. $\tau^*(N)$ and $\chi^*(N)$ combine $N$ tests.

challenging competitors in that case. We fix $T = 100$, which corresponds to a typical sample size in applications. Figure 2 summarizes the results. Throughout, the $\tau^*$ test outperforms the AEG$^*$ test marginally, which is itself substantially more powerful than the $\lambda^*_{\max}$ test.

To summarize, both $\tau^*$ and $\chi^*$ control the size of the test and yet provide a powerful and flexible alternative to traditional cointegration tests.

### 3.3 Extension to more than two tests

We combined AEG and $\lambda_{\max}$ tests to illustrate our approach with two widely applied cointegration tests. Of course, as the discussion in Section 2 makes clear, our approach is not restricted to combining two tests. The procedures can accommodate other and more tests as well. Potentially, this could yield further gains in power if the additional tests added extra information.

We therefore ran some extra simulations, where we additionally include the semi-parametric $t$-test of Phillips and Ouliaris (1990) (in their notation $\hat{Z}_t$) and the $\lambda_{\text{trace}}$ test of Johansen (1988). From the work of Gregory *et al.* (2004) we know that the correlation of tests within a group of tests, i.e. among residual-based and among system-based tests, is fairly high. Therefore we expect no large gain in power. Our exercise serves to check whether this intuition is correct; and more importantly it serves to check whether the meta test is still able to control size for $N > 2$. Both questions can be answered affirmatively, see Table 3. For comparison, we report the results for the combination of two tests ($\tau^*(2)$, $\chi^*(2)$) from Tables 1 and 2. There is a small—but insignificant—improvement in both size and power by moving to a version of the meta test that uses four underlying tests.

# 4 Empirical Application

## 4.1 Setup

Naturally we are interested in the applicability and the relevance of our testing strategy in practice. To shed light on this question, we revisit the studies which Gregory *et al.* (2004) investigated for 'mixed signals', i.e. conflicting test results from cointegration tests. Gregory *et al.* (2004) analyze the cointegration tests reported in 34 studies dealing with cointegration which were published in the *Journal of Applied Econometrics (JAE)* from 1994 to March/April 2001.[9] From these studies we construct 161 data sets in which we test for cointegration. The data sets exhibit large differences in sample size, which ranges from 27 to 7693 with a median size of 73. Similarly the number of variables differs across studies and ranges from 2 to 11.

Our goal is to document the extent to which conflicting test results arise in actual applications and how our proposed meta test is able to heal this problem. As Gregory *et al.* (2004), we do not intend to suggest that the authors of the original studies have been in any way strategic in their choice of which test for cointegration to apply. Most applied researchers tend to view the different tests as rather interchangeable, with the choice more dependent on the nature of the investigation.

We follow Gregory *et al.* (2004) closely in their setup. The original published studies employ different methods to test their specifications. To make the results comparable, we impose a unifying but standard methodology. For the residual-based tests where a dependent variable is required, we follow the choice in the original paper if possible. If there is no obvious dependent variable, we choose it on the basis of the highest $R^2$. Additionally we need to allow for variation in lag lengths across data sets. The literature discusses a number of different methods for choosing the number of lags. We have chosen a fairly standard one and determine the lag length $\hat{P}$ for the VECM estimation of our algorithm

[9] The data sets are available online through the *Journal of Applied Econometrics*' website (http://qed.econ.queensu.ca/jae/2004-v19.1/gregory-haug-lomuto/).

using a Schwarz Information Criterion (BIC) as described e.g. in Lütkepohl (2005, Sections 4.3.2 and 8.1). We search over the range $1 \leq \hat{P} \leq \min\left(8\left(\frac{T}{100}\right)^{1/5}, \frac{T-2}{2(K+1)}\right)$, where $K$ is the number of variables, and impose the same number of lags for the two Johansen tests and the Engle-Granger test. Our qualitative conclusions would not be different if alternative selection methods were employed. All tests include a constant and a trend.

## 4.2 Results

We compare the test results of (bootstrap versions of) $\lambda_{\text{trace}}$, $\lambda_{\max}$, Phillips and Ouliaris (1990), AEG tests as underlying tests with the $\tau^*(4)$ test.[10] To see how the $\tau^*$ test had worked in practice, we proceed as follows. We first check whether all single test agree or not in their testing decision at the 5% level, see left panel of Table 4. In those cases where conflicting test results occurred we check what the test used in the original paper had suggested as a test result (more precisely what would have been the outcome of our bootstrap version with the chosen lag-length criterion), see the right panel of Table 4. In all cases we compute and compare to the test result according to the $\tau^*(4)$ test.

Table 4 reports the frequencies for all possible pairs of outcomes.[11] As we argued at the end of Section 2, a feature of our test is that whenever all underlying bootstrap tests reject, so will the $\tau^*$. This theoretical result is confirmed. Moreover, we also see that when all tests do not reject the null, the meta test typically does not reject either. However, such cases of agreeing tests make up only 68% of all data sets (tests).

For the remaining 32% of data sets we have conflicting single tests and here our test turns out to be most useful. It allows the researcher to arrive at a definite conclusion. We find in 60% (=27/45) of the conflicting cases that the meta test does not reject the null. In the remaing 40% of the conflicting cases, however, the $\tau^*$ test leads to a rejection of the null of no cointegration. Moreover, we note the following.

First, rejecting whenever at least one (but not all) of the tests rejected would have lead to a substantial overstatement of cointegration (45 vs. 18 cases according to the $\tau^*$ test). Similarly, not rejecting whenever one test did not reject would have lead to an understatement of cointegration.

Second, the tests that have been ‘preferred’ in the actual studies tend to be more

[10] We performed the analogous exercise with asymptotic single tests. One might view this alternative as closer to conventional empirical practice. On the other hand, using the bootstrap tests as in Table 4 avoids the size distortion of the Johansen test in small samples (see Table 1). In any case, results are very similar. Tables are available upon request.

[11] For 18 data sets step 2 of our Algorithm finds an explosive root and hence we cannot calculate our bootstrap tests.

Table 4: Test results in applied studies and the $\tau^*$ test

| number of cases in which ... | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | single test results ... | | | | ... in case of conflicting | | | |
| | agree | | conflict | | results: 'preferred' test* | | | |
| | $r$ | $\neg r$ | | $\sum$ | | $r$ | $\neg r$ | $\sum$ |
| $\tau^*(4) : r$ | 65 | 1 | 18 | 84 | $\tau^*(4) : r$ | 11 | 5 | 16 |
| $\tau^*(4) : \neg r$ | 0 | 32 | 27 | 59 | $\tau^*(4) : \neg r$ | 14 | 11 | 25 |
| $\sum$ | 65 | 33 | 45 | 143 | $\sum$ | 25 | 16 | 41 |

$r$ : test rejects; $\neg r$ : test does not reject

* : Test type on which conclusions in the original study were based. For four data sets where we obtain conflicting test results no cointegration testing was reported.

Absolute frequencies of cointegration-test results for data from Gregory *et al.* (2004). Single tests include bootstrapped Engle-Granger, Phillips-Ouliaris and Johansen tests. The $\tau^*(4)$ combines these tests as described in Section 2. All bootstrap tests are constructed using 10,000 bootstrap resamples.

rejective than our meta test (25 vs. 16 rejections in 41 tests).[12] This suggests that the evidence in favor of cointegration would have been somewhat less pronounced if the studies could have relied on a suitable meta test for cointegration.

Third, whether or not the preferred test rejected the null does not seem to be informative on whether or not $\tau^*$ rejects conditional on observing conflicting test results. This is reflected by approximately equal conditional probabilities: $27/45 \simeq 14/25 \simeq 11/16$. In other words, we cannot conclude from a published test result what the $\tau^*$ test would indicate, conditional on the fact that a further single test leads to a conflicting test result.

## 5 Conclusion

This paper proposes a meta test that combines information from different underlying tests for cointegration. To the best of our knowledge, this is the first time that a practical approach has been put forward to combine different tests of one hypothesis applied to a

[12] In four cases of conflicting test results, the original study did not report a cointegration test but was rather concerned with e.g. estimating cointegration vectors.

single sample. The test takes into account the multiple testing nature of running more than one underlying test and hence controls size. By contrast, running more than one test and then simply inferring about the hypothesis from the most rejective test does not achieve this goal but leads to a significantly oversized test, as we have shown. While controlling size, the proposed meta test is powerful, and certainly more powerful than traditional methods to account for multiplicity like for example the Bonferroni method.

Extensive Monte Carlo simulations demonstrate the effectiveness of our approach. An application of our test to a set of cointegration studies confirms its practical value. It allows the applied researcher arrive at an unambiguous test decision in cases of conflicting single test results.

The setup we put forward is fairly general and hence can be adopted to other testing problems for which several (imperfectly correlated) tests have been developed. Examples include testing for unit roots or heteroscedasticity. Essentially, what is needed is a bootstrap method suitable for the phenomenon of interest. For the above mentioned testing problems such bootstrap methods would be the sieve and the wild bootstrap, respectively.

In practice, a major advantage of our proposed test should be that it relieves the applied researcher from the discretionary and sometimes arbitrary choice of the cointegration test(s) she wants to rely on to reach a test decision.

# References

Demetrescu M, Hassler U, Tarcolea AI. 2006. Combining significance of correlated statistics with application to panel data. *Oxford Bulletin of Economics and Statistics* **68**: 647–663.

Engle R, Granger C. 1987. Co-integration and error correction: Representation, estimation, and testing. *Econometrica* **55**: 251–76.

Fisher R. 1970. *Statistical Methods for Research Workers*. London: Oliver and Boyd, 14th edn.

Gonzalo J. 1994. Five alternative methods of estimating long-run equilibrium relationships. *Journal of Econometrics* **60**: 203–233.

Gregory AW, Haug AA, Lomuto N. 2004. Mixed signals among tests for cointegration. *Journal of Applied Econometrics* **19**: 89–98.

Hartung J. 1999. A note on combining dependent tests of significance. *Biometrical Journal* **41**: 849–855.

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

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

Lütkepohl H. 2005. *New Introduction to Multiple Time Series Analysis*. Berlin: Springer.

MacKinnon JG. 1996. Numerical distribution functions for unit root and cointegration tests. *Journal of Applied Econometrics* **11**: 601–618.

Maddala G, Wu S. 1999. A comparative study of unit root tests with panel data and a new simple test. *Oxford Bulletin of Economics and Statistics* **61**: 631–652.

Paparoditis E, Politis DN. 2003. Residual-based block bootstrap for unit root testing. *Econometrica* **71**: 813–855.

Phillips PCB, Ouliaris S. 1990. Asymptotic properties of residual based tests for cointegration. *Econometrica* **58**: 165–193.

Stouffer S, Suchman E, DeVinney L, Star S, Williams R. 1949. *The American Soldier*. Princeton: Princeton University Press.

Swensen AR. 2006. Bootstrap algorithms for testing and determining the cointegration rank in VAR models. *Econometrica* **74**: 1699–1714.

## Appendix for the referee

The following provides size and power tables for the 1% and 10% nominal level for DGP(A) and (B). Furthermore it provides size and power results for a third DGP with autocorrelated errors. This DGP has the same parameterization as DGP(A) except for that

$$u_{jt} = \rho u_{jt-1} + \varepsilon_{jt}, \;\; \rho = 0.33.$$

Table 5: Alternative DGP(C): Empirical size of the cointegration tests at nominal level of 5%

| | | Bootstrap tests | | | | | asymptotic tests | |
|---|---|---|---|---|---|---|---|---|
| DGP | $T$ | $\chi^*$ | $\tau^*$ | naive | $\lambda^*_{\max}$ | AEG* | $\lambda_{\max}$ | AEG |
| (C) | 50 | 0.0716 | 0.0814 | 0.1124 | 0.0734 | 0.0594 | 0.1046 | 0.0392 |
| AR(1) Shocks | 75 | 0.0572 | 0.0652 | 0.0942 | 0.058 | 0.0562 | 0.0938 | 0.0398 |
| | 100 | 0.0508 | 0.0564 | 0.0816 | 0.0544 | 0.0456 | 0.0954 | 0.0376 |
| | 125 | 0.0516 | 0.0582 | 0.0830 | 0.0500 | 0.0542 | 0.0920 | 0.0366 |
| | 150 | 0.0518 | 0.0558 | 0.0784 | 0.0488 | 0.0496 | 0.0856 | 0.0316 |

See notes to Table 1

Table 6: Alternative DGP(C): Empirical power of the cointegration tests at nominal level of 5%

| | | Bootstrap tests | | | | | asymptotic tests | |
|---|---|---|---|---|---|---|---|---|
| DGP | $T$ | $\chi^*$ | $\tau^*$ | naive | $\lambda^*_{\max}$ | AEG* | $\lambda_{\max}$ | AEG |
| (C) | 50 | 0.1144 | 0.1220 | 0.1756 | 0.0982 | 0.1092 | 0.1660 | 0.0712 |
| AR(1) Shocks | 75 | 0.1850 | 0.1970 | 0.2634 | 0.1688 | 0.1670 | 0.2654 | 0.1206 |
| | 100 | 0.3300 | 0.3518 | 0.4192 | 0.3042 | 0.2646 | 0.4346 | 0.2074 |
| | 125 | 0.5282 | 0.5544 | 0.6140 | 0.4978 | 0.4122 | 0.6282 | 0.3388 |
| | 150 | 0.7174 | 0.7372 | 0.7808 | 0.6898 | 0.5488 | 0.7950 | 0.4716 |

See notes to Table 1

Table 7: Empirical size of the cointegration tests at nominal level of 1%

| | | Bootstrap tests | | | | | asymptotic tests | |
|---|---|---|---|---|---|---|---|---|
| DGP | $T$ | $\chi^*$ | $\tau^*$ | 'naive' | $\lambda^*_{\max}$ | AEG$^*$ | $\lambda_{\max}$ | AEG |
| (A) | 50 | 0.0142 | 0.0210 | 0.0236 | 0.0160 | 0.0106 | 0.0200 | 0.0064 |
| | 75 | 0.0106 | 0.0144 | 0.0182 | 0.0100 | 0.0120 | 0.0168 | 0.0080 |
| | 100 | 0.0104 | 0.0148 | 0.0186 | 0.0122 | 0.0090 | 0.0186 | 0.0064 |
| | 125 | 0.0090 | 0.0142 | 0.0126 | 0.0086 | 0.0082 | 0.0144 | 0.0052 |
| | 150 | 0.0104 | 0.0128 | 0.0174 | 0.0128 | 0.0088 | 0.0210 | 0.0070 |
| (B) | 50 | 0.0170 | 0.0210 | 0.0236 | 0.0160 | 0.0106 | 0.0200 | 0.0064 |
| | 75 | 0.0116 | 0.0144 | 0.0182 | 0.0100 | 0.0120 | 0.0168 | 0.0080 |
| | 100 | 0.0112 | 0.0148 | 0.0186 | 0.0122 | 0.0090 | 0.0186 | 0.0064 |
| | 125 | 0.0090 | 0.0142 | 0.0126 | 0.0086 | 0.0082 | 0.0144 | 0.0052 |
| | 150 | 0.0122 | 0.0128 | 0.0174 | 0.0128 | 0.0088 | 0.0210 | 0.0070 |

See notes to Table 1

Table 8: Empirical size of the cointegration tests at nominal level of 10%

| | | Bootstrap tests | | | | | asymptotic tests | |
|---|---|---|---|---|---|---|---|---|
| DGP | $T$ | $\chi^*$ | $\tau^*$ | 'naive' | $\lambda^*_{\max}$ | AEG$^*$ | $\lambda_{\max}$ | AEG |
| (A) | 50 | 0.1212 | 0.1262 | 0.1822 | 0.1238 | 0.1046 | 0.1738 | 0.0706 |
| | 75 | 0.1078 | 0.1118 | 0.1598 | 0.1026 | 0.1022 | 0.1754 | 0.0764 |
| | 100 | 0.0984 | 0.1022 | 0.1524 | 0.1010 | 0.0952 | 0.1620 | 0.0726 |
| | 125 | 0.1026 | 0.1052 | 0.1560 | 0.1002 | 0.1058 | 0.1664 | 0.0800 |
| | 150 | 0.1012 | 0.1036 | 0.1510 | 0.0978 | 0.0994 | 0.1566 | 0.0738 |
| (B) | 50 | 0.1106 | 0.115 | 0.1702 | 0.1162 | 0.1008 | 0.1688 | 0.0684 |
| | 75 | 0.1060 | 0.1116 | 0.1582 | 0.0996 | 0.1052 | 0.1690 | 0.0752 |
| | 100 | 0.1022 | 0.1034 | 0.1574 | 0.1028 | 0.0992 | 0.1572 | 0.0728 |
| | 125 | 0.0962 | 0.1006 | 0.1486 | 0.0980 | 0.0950 | 0.1606 | 0.0682 |
| | 150 | 0.0940 | 0.0982 | 0.1426 | 0.0944 | 0.0934 | 0.1556 | 0.0712 |

See notes to Table 1

Table 9: Empirical power of the cointegration tests at nominal level of 1%

| | | Bootstrap tests | | | | | asymptotic tests | |
|---|---|---|---|---|---|---|---|---|
| DGP | $T$ | $\chi^*$ | $\tau^*$ | naive | $\lambda^*_{\max}$ | AEG* | $\lambda_{\max}$ | AEG |
| (A) | 50 | 0.0302 | 0.0378 | 0.0472 | 0.0268 | 0.0264 | 0.0444 | 0.0166 |
| | 75 | 0.0660 | 0.0802 | 0.0978 | 0.0614 | 0.0546 | 0.1034 | 0.0374 |
| | 100 | 0.1402 | 0.1664 | 0.1990 | 0.1486 | 0.0986 | 0.2224 | 0.0694 |
| | 125 | 0.2884 | 0.3362 | 0.3790 | 0.3126 | 0.1892 | 0.4164 | 0.1370 |
| | 150 | 0.4874 | 0.5456 | 0.5796 | 0.5190 | 0.2888 | 0.6242 | 0.2144 |
| (B) | 50 | 0.0240 | 0.0334 | 0.0344 | 0.0178 | 0.0228 | 0.0270 | 0.0148 |
| | 75 | 0.0288 | 0.0360 | 0.0384 | 0.0192 | 0.0308 | 0.0320 | 0.0210 |
| | 100 | 0.0582 | 0.0750 | 0.0748 | 0.0344 | 0.0608 | 0.0594 | 0.0402 |
| | 125 | 0.0906 | 0.1168 | 0.1140 | 0.0570 | 0.0964 | 0.0896 | 0.0616 |
| | 150 | 0.1378 | 0.1722 | 0.1620 | 0.0804 | 0.1446 | 0.1282 | 0.1006 |

See notes to Table 1

Table 10: Empirical power of the cointegration tests at nominal level of 10%

| | | Bootstrap tests | | | | | asymptotic tests | |
|---|---|---|---|---|---|---|---|---|
| DGP | $T$ | $\chi^*$ | $\tau^*$ | naive | $\lambda^*_{\max}$ | AEG* | $\lambda_{\max}$ | AEG |
| (A) | 50 | 0.2188 | 0.2258 | 0.3132 | 0.1894 | 0.2074 | 0.2910 | 0.1456 |
| | 75 | 0.3748 | 0.3834 | 0.4762 | 0.3348 | 0.3212 | 0.4746 | 0.2548 |
| | 100 | 0.6044 | 0.6120 | 0.6942 | 0.5628 | 0.4776 | 0.6944 | 0.3990 |
| | 125 | 0.8010 | 0.8114 | 0.854 | 0.7740 | 0.6446 | 0.8654 | 0.5756 |
| | 150 | 0.9306 | 0.9370 | 0.9484 | 0.9066 | 0.7874 | 0.9544 | 0.7324 |
| (B) | 50 | 0.1618 | 0.1698 | 0.2278 | 0.1344 | 0.1624 | 0.2166 | 0.1116 |
| | 75 | 0.2150 | 0.2280 | 0.2912 | 0.1624 | 0.234 | 0.2610 | 0.1778 |
| | 100 | 0.3126 | 0.3286 | 0.3926 | 0.2304 | 0.3366 | 0.3478 | 0.2690 |
| | 125 | 0.4364 | 0.4582 | 0.5172 | 0.3214 | 0.4704 | 0.4480 | 0.3944 |
| | 150 | 0.5626 | 0.5830 | 0.6456 | 0.433 | 0.605 | 0.5580 | 0.5242 |

See notes to Table 1

Table 11: Frequencies of test results in applied studies and the $\tau^*$ test: Comparison to asymptotic tests

| | results | | | |
|---|---|---|---|---|
| | conflict | agree | | |
| | | $r$ | $\neg r$ | $\sum$ |
| $\tau^* : r$ | 12 | 70 | 2 | 84 |
| $\tau^* : \neg r$ | 18 | 4 | 37 | 59 |
| $\sum$ | 30 | 74 | 33 | 143 |

$r$ : test rejects; $\neg r$ : test does not reject
$^*$ : Test type on which conclusions in the original study were based. For four data sets where we obtain conflicting test results no cointegration testing was reported.

Absolute frequencies of cointegration-test results for data from Gregory *et al.* (2004). Single tests include asymptotic Engle-Granger, Phillips-Ouliaris and Johansen tests. The $\tau^*$ (4) combines bootstrtapped versions of these tests as described in Section 2. All bootstrap tests are constructed using 10,000 bootstrap resamples.