# Combining Non-Cointegration Tests*

Christian Bayer† Christoph Hanck‡
Universität Bonn Rijksuniversiteit Groningen

First Version: May 2008
This Version: June 2012

**Abstract**

The local power of many popular non-cointegration tests has recently been shown to depend on a certain nuisance parameter. Depending on the value of that parameter, different tests perform best. This paper suggests combination procedures with the aim of providing meta tests that maintain high power across the range of the nuisance parameter.[1] We demonstrate the local power of the new meta tests to be in general almost as high as that of the most powerful of the underlying tests. When the underlying tests have similar power, the meta tests even appear more powerful than the best underlying test. At the same time, our new meta tests avoid the arbitrary decision which test to use if individual test results conflict. Moreover it avoids the size distortion inherent in separately applying multiple tests for cointegration to the same data set. We use the new tests to investigate 286 data sets from published cointegration studies. There, in one third of all cases individual tests give conflicting results whereas our meta tests provide an unambiguous test decision.

*Keywords*: Cointegration, Meta Test, Multiple Testing
*JEL-Codes*: C12, C22

*We are grateful to two anonymous referees whose comments helped to substantially improve the paper. All errors are ours. Part of the research was carried out when the authors were at Technische Universität Dortmund with support by DFG under Sonderforschungsbereich 475. We are grateful to Jörg Breitung, Carsten Burhop, Carlo Favero, Christian Gengenbach, Alain Hecq, Walter Krämer, Franz Palm and Jean-Pierre Urbain as well as conference and seminar participants in Neuchâtel, Bonn, Maastricht, Groningen, Amsterdam, Aachen, Münster and Magdeburg for valuable comments and suggestions.

†IIW, Lennéstr. 37, 53113 Bonn, Germany. Tel.: +49 (0)228 73 4073. email: christian.bayer@uni-bonn.de.

‡Department of Economics and Econometrics, Nettelbosje 2, 9747AE Groningen, Netherlands. Tel.: +31 (0)50 3633836, Fax +31 (0)50 3637337. email: c.h.hanck@rug.nl.

# 1 Introduction

Cointegration testing is a standard tool in applied economics. Various tests have been suggested, most of which are implemented in econometric software packages. Well-known examples include the residual-based test of Engle and Granger (1987), or the system-based tests of Johansen (1988). Boswijk (1994) and Banerjee *et al.* (1998) suggest error-correction-based tests. This regularly forces the practitioner to select from various test decisions. This choice is difficult as there is no uniformly most powerful test, even asymptotically (e.g. Elliott *et al.*, 2005). Often one test rejects while another test does not, complicating interpretation of test outcomes.

More generally, the $p$-values of the tests are not perfectly correlated (Gregory *et al.*, 2004), which rules out relying, e.g., on the test with the smallest $p$-value. Doing so would lead to an oversized test as it ignores the multiple testing nature of such procedure. The imperfect correlation reflects that the tests are not equivalent, which also has implications for their power: Pesavento (2004) shows that the power ranking of cointegration tests depends crucially on the value of a nuisance parameter, viz. the squared long-run correlations of error terms driving the variables, cf. Sec. 2.

This suggests that suitable combinations of tests might yield a more robust power performance, and possibly even power gains, relative to individual tests. Combining the above individual tests in the spirit of Fisher's (1932) famous test, Section 3 develops such devices. As the individual test statistics have nonstandard distributions and are correlated, the distribution function of the combination statistic is analytically intractable. However, as is typical for individual cointegration statistics, we can simulate its asymptotic distribution, exploiting Pesavento's (2004) results. Similarly, Section 3 modifies the Union-of-Rejections ($UR$) approach of Harvey *et al.* (2009) and apply the generalized $UR$ test to the present testing problem.

Section 4 shows our Fisher-type test to perform well asymptotically. Its local power is close to that of the best of the individual tests for different values of the nuisance parameter, and even exceeds it when the individual tests have similar power. The $UR$ test is most useful when the individual tests have strongly different power; its power is always close to that of the better underlying test.

Section 5 proposes bootstrap analogs of our tests. Section 6 presents finite-sample experiments of the asymptotic and bootstrap combination tests. The asymptotic results correctly predict the finite-sample performance. The tests successfully control size and are powerful against general types of alternatives. The bootstrap tests have slightly better size.

Section 7 employs the new tests to revisit the published studies that Gregory *et al.* (2004) examined for 'mixed signals', i.e. conflicting cointegration tests. We furthermore update the dataset with publications in the JAE from 2001 to 2010. In one third of all cases individual tests give conflicting results. In these cases our meta tests are particularly useful, providing an unambiguous test decision and therefore a solution to the 'mixed signals' problem. A web appendix (available from our websites) gives additional results.

# 2 Motivation and Setup

## 2.1 Motivation

Applied researchers regularly study whether several nonstationary time series are cointegrated, but are unsure with which test to test the no-cointegration null. We purposely select some examples from the literature (further discussed in Sec. 7) to show that all kinds of mixed signals are possible—some tests rejecting, some tests not rejecting and no test always rejecting:[2]

Clements and Hendry (1995) consider a bivariate system of the inverse velocity of circulation and an opportunity cost of holding money. They find cointegration using the Johansen (1988) procedure (for some detail on the tests see Section 2.2), which we confirm with a $\lambda_{\max}$ $p$-value of 0.0003. However, the Engle and Granger (1987) and Banerjee *et al.* (1998) $p$-values are 0.6843 and 0.0883, producing no and only weak evidence for cointegration. The Boswijk (1994) $p$-value is 0.0001, such that the two rejections and two non-rejections produce a mixed signal.

Cooley and Ogaki (1996) examine the long-run relationship between the logs of real per capita non-durables consumption and wages deflated by non-durables prices. Using the test of Park (1990), they find little evidence against the null hypothesis of cointegration. The tests of Johansen (1988), Banerjee *et al.* (1998) and Boswijk (1994) yield the opposite conclusion, not rejecting the null of no cointegration with $p$-values of 0.0744, 0.5630 and 0.5302. The Engle and Granger (1987) test is consistent with Park (1990), with a $p$-value of 0.0142. We hence again observe mixed signals.

Martens *et al.* (1998) study the cost-of-carry model which, via arbitrage, predicts cointegration between index and index-futures prices. Using Engle and Granger (1987) and Johansen (1988) tests, they find strong evidence for cointegration. For e.g. the May '93 relationship, we also find $p$-values close to zero. However, the Banerjee *et al.* (1998) and Boswijk (1994) $p$-values of 0.1301 and 0.0764 do not produce (strong) evidence for cointegration, again yielding a mixed signal.

Overall, mixed signals can easily arise in applications. Moreover, no uniformly most powerful choice emerges from the studies, motivating the need for a combination procedure of the tests.

## 2.2 Model and Individual Tests

We work with Pesavento's (2004) model:

$$\Delta \boldsymbol{x}_t = \boldsymbol{\tau}_1 + \boldsymbol{v}_{1t} \tag{1a}$$

$$y_t = (\mu_2 - \boldsymbol{\theta}'\boldsymbol{\mu}_1) + (\tau_2 - \boldsymbol{\theta}'\boldsymbol{\tau}_1)t + \boldsymbol{\theta}'\boldsymbol{x}_t + u_t \qquad \text{where} \qquad u_t = \rho u_{t-1} + v_{2t} \tag{1b}$$

Eq. (1a) defines the regressor dynamics, while (1b) describes the (single potential) cointegrating vector. Write $\boldsymbol{z}_t = (\boldsymbol{x}_t',\ y_t)'$. The observed sample is $\boldsymbol{z}_0, \ldots, \boldsymbol{z}_T$. Restrictions on $\boldsymbol{\mu}_1'$, $\mu_2$, $\boldsymbol{\tau}_1$ and $\tau_2$ determine the deterministic components, see Pesavento (2004) for details. These amount to no deterministics, a constant, or a constant plus trend. We refer to these as cases $(i)$, $(ii)$, and $(iii)$. Further, $\boldsymbol{v}_t := (\boldsymbol{v}_{1t}', v_{2t})'$ and let $\boldsymbol{\Omega}$ the long-run covariance matrix of $\boldsymbol{v}_t$. Pesavento (2004) maintains the following assumptions to derive the local power of the tests mentioned above:

*Assumption 1.* $\{\boldsymbol{v}_t\}$ satisfies a Functional CLT, i.e. $T^{-1/2}\sum_{t=1}^{[\cdot T]}\boldsymbol{v}_t \Rightarrow \boldsymbol{\Omega}^{1/2}\boldsymbol{W}(\cdot)$, i.e. $\boldsymbol{z}_t$ is $I(1)$.

*Assumption 2.* There are no cointegrating relationships among the variables in $\boldsymbol{x}_t$.

The vector $\boldsymbol{z}_t$ is cointegrated under Assumptions 1 and 2 if $|\rho| < 1$, such that we can state the null hypothesis of no cointegration as

$\mathcal{H}_0 : \rho = 1$ and Assumptions 1 and 2 are true.

The literature has suggested many tests of $\mathcal{H}_0$, typically against one of the following alternatives, which we list in ascending order of generality ($\mathcal{H}_1^a \Rightarrow \mathcal{H}_1^b \Rightarrow \mathcal{H}_1^c$):

$\mathcal{H}_1^a : |\rho| < 1$ and Assumptions 1 and 2 are true.

$\mathcal{H}_1^b : |\rho| < 1$ and Assumption 2 is true (Assumption 1 may or may not hold).

$\mathcal{H}_1^c : |\rho| < 1$ (Assumptions 1 and 2 may or may not hold).

Practitioners may for instance wish to test against $\mathcal{H}_1^c$, aiming to establish a cointegrating relationship in $\boldsymbol{z}_t$, not necessarily excluding the possibility of more than one cointegrating relationship (e.g. a second one not involving $y_t$). Concretely, we consider the tests of Engle and Granger (1987), Johansen (1988), Boswijk (1994) and Banerjee *et al.* (1998).

The Engle and Granger (1987) test tests $\mathcal{H}_0$ against the alternative of at least one cointegrating relationship. One first computes $\hat{u}_t$, the residual from a regression of $y_t$ on $\boldsymbol{x}_t$ (and appropriate deterministics $\boldsymbol{d}_t$), and then the $t$-statistic $t_\gamma^{\mathrm{ADF}}$ on $\gamma$ in the regression $\Delta\hat{u}_t = \gamma\hat{u}_{t-1} + \sum_{p=1}^{P-1}\nu_p\Delta\hat{u}_{t-p} + \epsilon_t$.

The system-based tests of Johansen (1988) test for $h$ cointegrating relationships. In view of $\mathcal{H}_0$, we consider $h = 0$. One estimates the Vector Error Correction Model (VECM)

$$\Delta\boldsymbol{z}_t = \boldsymbol{\Pi}\boldsymbol{z}_{t-1} + \sum_{p=1}^{P-1}\boldsymbol{\Gamma}_p\Delta\boldsymbol{z}_{t-p} + \boldsymbol{d}_t + \boldsymbol{\varepsilon}_t \tag{2}$$

We employ the $\lambda_{\max}(h) = -T\ln(1-\hat{\pi}_1)$ test statistic. Here, $\hat{\pi}_1$ denotes the largest solution to $|\pi\boldsymbol{S}_{11} - \boldsymbol{S}_{10}\boldsymbol{S}_{00}^{-1}\boldsymbol{S}_{01}| = 0$, where the $\boldsymbol{S}_{ij}$ are moment matrices of reduced rank regression residuals.

Banerjee *et al.* (1998) and Boswijk (1994) develop error correction-based tests. One estimates (by OLS) the equation $\Delta y_t = d_t + \boldsymbol{\pi}_{0x}'\Delta\boldsymbol{x}_t + \varphi_0 y_{t-1} + \boldsymbol{\varphi}_1'\boldsymbol{x}_{t-1} + \sum_{p=1}^{P}(\boldsymbol{\pi}_{px}'\Delta\boldsymbol{x}_{t-p} + \pi_{py}\Delta y_{t-p}) + \epsilon_t$, with $P$ chosen such that $\epsilon_t$ is approximately white noise. Banerjee *et al.*'s test statistic $t_\gamma^{\mathrm{ECR}}$ is the $t$-ratio for $\mathcal{H}_0 : \varphi_0 = 0$, whereas Boswijk's $\hat{F}$ is the Wald statistic for $\mathcal{H}_0 : (\varphi_0,\ \boldsymbol{\varphi}_1')' = \boldsymbol{0}$.

Pesavento (2004) shows that, under (1), the local power of these tests against $\mathcal{H}_1^a$ only depends on the local-to-unity parameter $c := T(\rho - 1)$ and $R^2$, the squared correlation of $\boldsymbol{v}_{1t}$ with $v_{2t}$. Concretely, partition $\boldsymbol{\Omega}$ conformably with $(\boldsymbol{x}_t', y_t)'$,

$$\boldsymbol{\Omega} = \begin{pmatrix} \boldsymbol{\Omega}_{11} & \boldsymbol{\omega}_{12} \\ \boldsymbol{\omega}_{12}' & \omega_{22} \end{pmatrix},$$

and define the squared correlation as $R^2 := \boldsymbol{\omega}_{12}'\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\omega}_{12}/\omega_{22}$.[3] Her Theorems 1, and 3-5 show all limiting functionals to be driven by the same Wiener process $\boldsymbol{W}$, such that her result allows us

to consider the *joint* distribution of the statistics. They further show that the different statistics are non-equivalent functionals of $\boldsymbol{W}$, and differentially affected by the nuisance parameter $R^2$ for $c < 0$. Hence, as formalized by Pesavento (2004) and further discussed in Section 4, different tests are powerful for different $R^2$. This is the basis of the combination procedures presented next.

# 3 Combination Tests

Under $\mathcal{H}_0$, many of the above statistics are only weakly correlated, even asymptotically (Gregory *et al.*, 2004). Further, Pesavento (2004) shows that different tests are most powerful for different $R^2$. Thus, a more robust, and possibly even more powerful, combination test can in principle be achieved. To this end, let $t_i$ be the test statistic of test $i$. Take $\xi_i := t_i$ ($-\xi_i = t_i$) if test $i$ rejects for large (small) values. Also, with $\Xi_i(x) := \Pr_{\mathcal{H}_0}(\xi_i \geqslant x)$, the $p$-value of test $i$ is $p_i := \Xi_i(\xi_i)$.

## 3.1 A Fisher-type test

To reach a joint test decision from the various $\xi_i$, we need a suitable aggregator. One such aggregator is given by Fisher's (1932) famous $\chi^2$ test. Let $\mathcal{I}$ the index set of the $\xi_i$ to be aggregated. We then have the following corollary from Pesavento (2004), whose proof follows directly using the CMT (see also White (2000, Prop. 2.2), for a more detailed argument see Appendix A in the extended online version).

**Corollary 1.** *Consider the test statistic*

$$\tilde{\chi}^2_{\mathcal{I}} := -2 \sum_{i \in \mathcal{I}} \ln(p_i). \tag{3}$$

*As $T \to \infty$, (a) $\tilde{\chi}^2_{\mathcal{I}} \to_d \mathcal{F}_{\mathcal{I}}$ under $\mathcal{H}_0$, with $\mathcal{F}_{\mathcal{I}}$ some random variable. Further, (b) $\tilde{\chi}^2_{\mathcal{I}} \to_{\mathrm{p}} \infty$ under $\mathcal{H}_1$, i.e. $\tilde{\chi}^2_{\mathcal{I}}$ is consistent if at least one of the underlying tests is consistent.*

Part $(a)$ guarantees that the $\tilde{\chi}^2_{\mathcal{I}}$ have well-defined asymptotic null distributions, call them $F_{\mathcal{F}_{\mathcal{I}}}$. These are nuisance-parameter free because of (i) the single $\xi_i$ are nuisance parameter free (cf. e.g. Appendix A) and (ii) the $F_{\mathcal{F}_{\mathcal{I}}}$ take the cross-relation between the $\xi_i$ into account. The $F_{\mathcal{F}_{\mathcal{I}}}$ depend on which and how many tests are combined. Of course we cannot invoke the conventional $\chi^2(2|\mathcal{I}|)$ (with $|\mathcal{I}|$ the cardinality of $\mathcal{I}$) null distribution, as independence of the $\xi_i$ would be necessary.

Moreover, part $(a)$ allows to infer by simulation the *joint* null distribution of the $\xi_i$ and hence the distribution $F_{\mathcal{F}_{\mathcal{I}}}$ of the continuous aggregator (3), exploiting e.g. Pesavento's results.[4] Table 1 reports 5%-critical values $cv^{0.05}_{\mathcal{I}} := F^{-1}_{\mathcal{F}_{\mathcal{I}}}(0.95)$ for combinations relevant to us here (Tables B.2-B.3 in the online appendix report other $\alpha$ and test combinations).[5] Reject if $\tilde{\chi}^2_{\mathcal{I}} > cv^{\alpha}_{\mathcal{I}}$. Since the distributions of the $\xi_i$ depend on $K-1$ as well as the deterministic case $(i)$-$(iii)$, that of the $\tilde{\chi}^2_{\mathcal{I}}$ will do so, too.

The theoretical upper bound for the $cv^{0.05}_{\mathcal{I}}$ for e.g. $|\mathcal{I}| = 2$ is $-2\sum_{i\in\mathcal{I}} \ln(0.05) = 11.983$, which obtains if tests are perfectly correlated (hence, equivalent). Ruling out negative correlation,

Table 1: 5%-critical values $cv_{\mathcal{I}}^{0.05}$ for the $\tilde{\chi}^2_{\mathcal{I}}$ tests

| $K-1$ | case: | $(i)$ | $(ii)$ | $(iii)$ | $(i)$ | $(ii)$ | $(iii)$ | $(i)$ | $(ii)$ | $(iii)$ |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $t_\gamma^{\text{ADF}}$ and $\lambda_{\max}$ | | | $\hat{F}$ and $t_\gamma^{\text{ECR}}$ | | | $\hat{F}$, $\lambda_{\max}$, $t_\gamma^{\text{ADF}}$, $t_\gamma^{\text{ECR}}$ | | |
| 1 | | 11.071 | 11.229 | 11.269 | 11.606 | 11.803 | 11.862 | 21.352 | 21.931 | 22.215 |
| 2 | | 10.838 | 10.895 | 10.858 | 11.556 | 11.716 | 11.795 | 20.776 | 21.106 | 21.342 |
| 3 | | 10.640 | 10.637 | 10.711 | 11.554 | 11.683 | 11.731 | 20.237 | 20.486 | 20.788 |
| 4 | | 10.516 | 10.576 | 10.532 | 11.491 | 11.611 | 11.696 | 19.951 | 20.143 | 20.440 |
| 5 | | 10.406 | 10.419 | 10.448 | 11.478 | 11.621 | 11.639 | 19.747 | 19.888 | 20.170 |

5%-critical values for combination tests based on $\tilde{\chi}^2_{\mathcal{I}}$. $t_\gamma^{\text{ADF}}$ is from Engle and Granger (1987), $\lambda_{\max}$ from Johansen (1988), $\hat{F}$ from Boswijk (1994) and $t_\gamma^{\text{ECR}}$ from Banerjee *et al.* (1998).

a lower bound is 9.487, the 5% $\chi^2(4)$ critical value of Fisher's test under independence. The actual $cv_{\mathcal{I}}^{0.05}$ are close to 11 for $|\mathcal{I}| = 2$, vary little across cases and fall moderately in $K$. Also, $cv_{\mathcal{I}}^{0.05} > 9.487$, reflecting that the $\xi_i$ correlate positively and using $cv_{\mathcal{I}}^{0.05}$ instead of 9.487 is necessary for level-$\alpha$ tests. Moreover, $\tilde{\chi}^2_{\mathcal{I}}$ rejects when all $\xi_i$ reject, as $cv_{\mathcal{I}}^{0.05} < -2\sum_{i\in\mathcal{I}} \ln(0.05)$. The latter implies that $\tilde{\chi}^2_{\mathcal{I}}$ may reject even if no single $\xi_i$ rejects: e.g., if $K = 2$, case $(iii)$ and $p_1 = \ldots = p_4 = 0.0622$, $\tilde{\chi}^2_{\mathcal{I}} = -8 \cdot \ln(0.0622) = 22.215$.

Pesavento's results allow even to obtain the asymptotic distribution under $\mathcal{H}_1^a$, and hence the local power of the $\tilde{\chi}^2_{\mathcal{I}}$ (cf. Sec. 4). In brief, we find $\tilde{\chi}^2_{\mathcal{I}}$ to generally be almost as powerful as the best underlying test. This result is useful as it gives some theoretical guidelines for a quite broad class of models. However, practitioners sometimes want to test against the more general alternatives $\mathcal{H}_1^b$ and $\mathcal{H}_1^c$. Although we cannot derive the local power of the tests under $\mathcal{H}_1^b$ and $\mathcal{H}_1^c$, part $(b)$ of Corollary 1 ensures that $\tilde{\chi}^2_{\mathcal{I}}$ is at least consistent also against such alternatives. The investigation of the tests in such scenarios in Section 6 shows that the ranking of the $\xi_i$ found under $\mathcal{H}_1^a$ does not carry over to alternatives $\mathcal{H}_1^b$ and $\mathcal{H}_1^c$. Yet, $\tilde{\chi}^2_{\mathcal{I}}$ continues to closely track the best single test.

*Remark 1.* We also tried alternative aggregators such as the inverse-normal one $\sum_{i\in\mathcal{I}} \Phi^{-1}(p_i)/\sqrt{|\mathcal{I}|}$, with $\Phi^{-1}$ the standard normal quantile function. It was however slightly inferior to that of the $\tilde{\chi}^2_{\mathcal{I}}$ tests, reported below. The superiority of $\tilde{\chi}^2_{\mathcal{I}}$ is not surprising in that known optimality results under independence (Littell and Folks, 1971) appear to carry over to the dependent case. Intuitively, aggregators such as the sum of the $p$-values are plausibly less powerful: small $p$-values cause $\tilde{\chi}^2_{\mathcal{I}}$ to diverge via ln, and hence high power. This is not the case for the sum aggregator.

### 3.2 Union-of-Rejections tests

Harvey *et al.* (2009) develop 'Union-of-Rejections' ($UR$) tests to combine standard Dickey-Fuller (DF) and GLS-demeaned unit root tests. The $UR$ test rejects when one test rejects, suitably adjusting the critical values to ensure a level-$\alpha$ test. It has robust power as DF (GLS) is more powerful when the series' initial condition is large (small). This situation is analogous to ours, as $R^2$ determines the relative power of the $\xi_i$. We use and extend the $UR$ approach to cointegration testing.

Denote the individual level-$\alpha$ critical value of test $i$ as $cv_i^\alpha$, e.g., $cv_i^{0.05} = |-2.763|$ for $t_\gamma^{\text{ADF}}$, $K = 2$

Table 2: Critical values for the minimum $p$-value test

| $K-1$ | case: | $(i)$ | $(ii)$ | $(iii)$ | $(i)$ | $(ii)$ | $(iii)$ |
|---|---|---|---|---|---|---|---|
| | | $t_\gamma^{\text{ADF}}$ and $\lambda_{\max}$ | | | $\hat{F}$ and $t_\gamma^{\text{ECR}}$ | | |
| 1 | | 0.031 | 0.033 | 0.033 | 0.038 | 0.041 | 0.043 |
| 2 | | 0.030 | 0.030 | 0.030 | 0.037 | 0.038 | 0.040 |
| 3 | | 0.029 | 0.029 | 0.029 | 0.036 | 0.038 | 0.039 |
| 4 | | 0.028 | 0.028 | 0.028 | 0.036 | 0.037 | 0.038 |
| 5 | | 0.028 | 0.028 | 0.028 | 0.035 | 0.036 | 0.037 |

Critical values for the minimum $p$-value test when testing at $\alpha = 0.05$.

and case $(i)$. The 'naive' $UR$ statistic test $UR^{\text{n}}(\xi_1, \xi_2) := \mathbb{I}\{\xi_1 > cv_1^\alpha\} + \mathbb{I}\{\xi_1 \leqslant cv_1^\alpha\}\mathbb{I}\{\xi_2 > cv_2^\alpha\}$, with $\mathbb{I}\{A\}$ the indicator function, rejects if $UR^{\text{n}}(\xi_1, \xi_2) = 1$. $UR^{\text{n}}$ is oversized as it ignores the multiple testing nature of the problem.[6] Harvey *et al.* (2009) therefore suggest to reject if $UR_\psi(\xi_1, \xi_2) = 1$, where $UR_\psi(\xi_1, \xi_2) := \mathbb{I}\{\xi_1 > \psi cv_1^\alpha\} + \mathbb{I}\{\xi_1 \leqslant \psi cv_1^\alpha\}\mathbb{I}\{\xi_2 > \psi cv_2^\alpha\}$ and $\psi$ satisfies $\Pr(\bigcup_{i=1}^2 \xi_i > \psi cv_i^\alpha) = \alpha$. However, there is no need to apply the *same* $\psi$ to both $cv_i^\alpha$. Consider the general $UR$ statistic

$$UR_{\psi_{\mathcal{I}}}(\xi_1, \xi_2) := \mathbb{I}\{\xi_1 > \psi_1 cv_1^\alpha\} + \mathbb{I}\{\xi_1 \leqslant \psi_1 cv_1^\alpha\}\mathbb{I}\{\xi_2 > \psi_2 cv_2^\alpha\} \tag{4}$$

An appealing selection rule for the $\psi_i$ is to ensure the same null rejection probabilities for both $\xi_i$. It takes into account that the $\Xi_i$, and thus the $cv_i^\alpha$, can be defined on different scales:

$$\Pr(\xi_1 > \psi_1 cv_1^\alpha) = \Pr(\xi_2 > \psi_2 cv_2^\alpha) \tag{5}$$

Under (5), the $UR_{\psi_{\mathcal{I}}}$ test is equivalent to the 'non-naive' minimum $p$-value test $\min_{i\in\mathcal{I}} p_i$.[7] The critical values of the non-naive $\min_{i\in\mathcal{I}} p_i$ test, provided in Table 2, yield the level $\alpha' < \alpha$ at which to test to avoid the oversizedness of $UR^{\text{n}}$ (which, in turn, is the same as rejecting if $\min_{i\in\mathcal{I}} p_i < \alpha$). Note $\alpha' \gg \alpha/|\mathcal{I}|$ so that $\min_{i\in\mathcal{I}} p_i$ is more powerful than a Bonferroni-type multiple test.

## 4 Large Sample Results

We now report the large-sample power of the tests discussed in Sections 2 and 3. As for the single tests, the local power functions of $\tilde{\chi}^2_{\mathcal{I}}$ and $UR_{\psi_{\mathcal{I}}}(\xi_1, \xi_2)$ are not available in closed form. They are hence simulated with 25,000 replications of the distributions given in Section 3, for $T = 1,000$. We consider $c \in \{0, -1, \ldots, -30\}$, $R^2 \in \{0, 0.05, \ldots, 0.95\}$ and $K \in \{2, \ldots, 6\}$.

Table 3 reports the local power of several combination and individual tests for case $(iii)$ (cf. online Appendix C for $(i)$ and $(ii)$).[8] Figure 1 plots power against $R^2$, for $c = -15$ and $K - 1 = 1$; $t_\gamma^{\text{ECR}}$ is the best individual test for small $R^2$ (Pesavento, 2004). The power of all tests but $t_\gamma^{\text{ADF}}$ increases in $R^2$. The $\lambda_{\max}$ system test benefits most from an increase in $R^2$, fully exploiting the information contained in the $\boldsymbol{x}_t$. The formal similarity of $\hat{F}$ and $t_\gamma^{\text{ECR}}$ translates into similar power.

The combination tests perform very well, tracking the better test very closely. Their power is sometimes even higher than that of all underlying tests; e.g. for $R^2 = 0.2$, $\tilde{\chi}^2_{\mathcal{I}}(t_\gamma^{\text{ADF}}, \lambda_{\max})$. Figure

Table 3: Local asymptotic power

| $-c$ | 0 | 5 | 10 | 15 | 20 | 0 | 5 | 10 | 15 | 20 |
|---|---|---|---|---|---|---|---|---|---|---|
| | $R^2=0$ | | | | | $R^2=0.25$ | | | | |
| $\tilde{\chi}^2_{\mathcal{I}}(\hat{F},t^{\mathrm{ECR}}_\gamma)$ | 0.050 | 0.073 | 0.148 | 0.290 | 0.487 | 0.048 | 0.081 | 0.191 | 0.405 | 0.668 |
| $\tilde{\chi}^2_{\mathcal{I}}(t^{\mathrm{ADF}}_\gamma,\lambda_{\max})$ | 0.050 | 0.069 | 0.132 | 0.253 | 0.423 | 0.050 | 0.072 | 0.127 | 0.267 | 0.495 |
| $\tilde{\chi}^2_{\mathcal{I}}(4)$ | 0.050 | 0.074 | 0.151 | 0.294 | 0.490 | 0.049 | 0.084 | 0.194 | 0.406 | 0.664 |
| $UR_{\psi_{\mathcal{I}}}(t^{\mathrm{ADF}}_\gamma,\lambda_{\max})$ | 0.049 | 0.070 | 0.142 | 0.279 | 0.471 | 0.051 | 0.069 | 0.121 | 0.247 | 0.456 |
| $UR_{\psi_{\mathcal{I}}}(\hat{F},t^{\mathrm{ECR}}_\gamma)$ | 0.051 | 0.064 | 0.116 | 0.230 | 0.392 | 0.050 | 0.079 | 0.171 | 0.364 | 0.626 |
| $\hat{F}$ | 0.050 | 0.070 | 0.138 | 0.271 | 0.457 | 0.047 | 0.083 | 0.199 | 0.412 | 0.668 |
| $t^{\mathrm{ECR}}_\gamma$ | 0.050 | 0.076 | 0.155 | 0.305 | 0.508 | 0.049 | 0.083 | 0.183 | 0.388 | 0.652 |
| $\lambda_{\max}$ | 0.050 | 0.054 | 0.092 | 0.165 | 0.283 | 0.050 | 0.067 | 0.123 | 0.261 | 0.471 |
| $t^{\mathrm{ADF}}_\gamma$ | 0.050 | 0.074 | 0.150 | 0.290 | 0.486 | 0.050 | 0.070 | 0.115 | 0.222 | 0.398 |
| | $R^2=0.5$ | | | | | $R^2=0.75$ | | | | |
| $\tilde{\chi}^2_{\mathcal{I}}(\hat{F},t^{\mathrm{ECR}}_\gamma)$ | 0.049 | 0.089 | 0.285 | 0.621 | 0.874 | 0.051 | 0.134 | 0.596 | 0.923 | 0.993 |
| $\tilde{\chi}^2_{\mathcal{I}}(t^{\mathrm{ADF}}_\gamma,\lambda_{\max})$ | 0.050 | 0.063 | 0.146 | 0.386 | 0.699 | 0.054 | 0.069 | 0.356 | 0.811 | 0.983 |
| $\tilde{\chi}^2_{\mathcal{I}}(4)$ | 0.049 | 0.080 | 0.231 | 0.552 | 0.840 | 0.053 | 0.107 | 0.524 | 0.906 | 0.993 |
| $UR_{\psi_{\mathcal{I}}}(t^{\mathrm{ADF}}_\gamma,\lambda_{\max})$ | 0.049 | 0.102 | 0.318 | 0.648 | 0.882 | 0.050 | 0.196 | 0.689 | 0.946 | 0.995 |
| $UR_{\psi_{\mathcal{I}}}(\hat{F},t^{\mathrm{ECR}}_\gamma)$ | 0.049 | 0.069 | 0.179 | 0.439 | 0.734 | 0.053 | 0.117 | 0.531 | 0.907 | 0.993 |
| $\hat{F}$ | 0.048 | 0.108 | 0.339 | 0.669 | 0.891 | 0.052 | 0.216 | 0.714 | 0.952 | 0.996 |
| $t^{\mathrm{ECR}}_\gamma$ | 0.048 | 0.079 | 0.228 | 0.537 | 0.823 | 0.051 | 0.077 | 0.385 | 0.801 | 0.970 |
| $\lambda_{\max}$ | 0.048 | 0.078 | 0.221 | 0.511 | 0.794 | 0.051 | 0.153 | 0.607 | 0.937 | 0.996 |
| $t^{\mathrm{ADF}}_\gamma$ | 0.050 | 0.052 | 0.077 | 0.151 | 0.292 | 0.054 | 0.029 | 0.035 | 0.071 | 0.166 |

Case (*iii*). $\tilde{\chi}^2_{\mathcal{I}}(\hat{F},t^{\mathrm{ECR}}_\gamma)$ is (3) based on Boswijk's and Banerjee *et al.*'s tests, and $UR_{\psi_{\mathcal{I}}}(\hat{F},t^{\mathrm{ECR}}_\gamma)$ is the $UR$ test (4). The other combination tests are defined analogously. See also notes to Table 1.

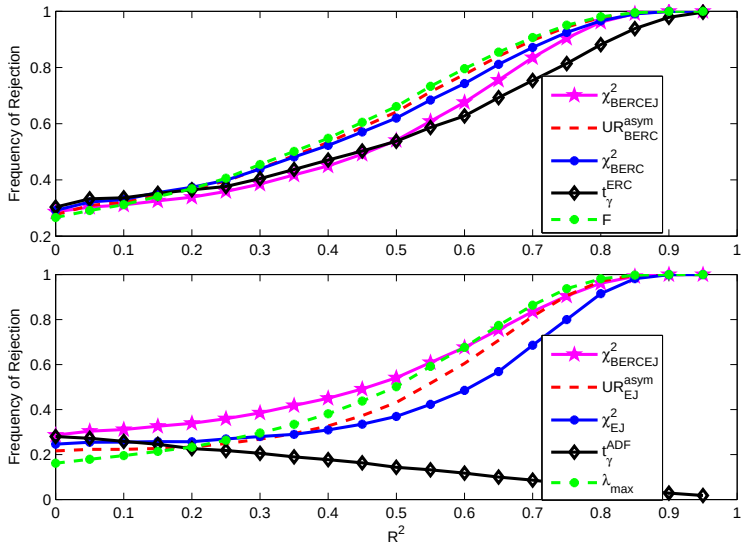
1 shows the power curves of $t^{\mathrm{ADF}}_\gamma$ and $\lambda_{\max}$ to intersect at $R^2\approx 0.2$. Thus, combination tests may outperform the constituent tests when the latter are equally powerful. Intuitively, this is because the $\xi_i$ will then often be individually marginally too small to reject, but, , since they imperfectly correlated, taken together they provide sufficient evidence to reject $\mathcal{H}_0$. The upper panel shows that, unsurprisingly, the power of the combination tests differs less from that of the underlying tests if these perform similarly. Yet, $UR_{\psi_{\mathcal{I}}}(\hat{F},t^{\mathrm{ECR}}_\gamma)$ and $\tilde{\chi}^2_{\mathcal{I}}(\hat{F},t^{\mathrm{ECR}}_\gamma)$ are again closer to the better underlying test (typically $\hat{F}$) whenever there are discernible differences.

Of course, when the difference between the individual tests is large, as in the lower panel of Figure 1 for $R^2\approx 0.6$, the power distance to the best individual test is somewhat larger—but still a lot smaller than that to the worse individual test. Thus, the combination tests cheaply insure against selecting an inferior test, in that one never sacrifices much power, and potentially gains a lot.

Table 3 shows that $\tilde{\chi}^2_{\mathcal{I}}(\hat{F},t^{\mathrm{ECR}}_\gamma,t^{\mathrm{ADF}}_\gamma,\lambda_{\max})=:\tilde{\chi}^2_{\mathcal{I}}(4)$ outperforms $\tilde{\chi}^2_{\mathcal{I}}(t^{\mathrm{ADF}}_\gamma,\lambda_{\max})$, but is (slightly) outperformed by $\tilde{\chi}^2_{\mathcal{I}}(\hat{F},t^{\mathrm{ECR}}_\gamma)$. This is not surprising as $\hat{F}$ and $t^{\mathrm{ECR}}_\gamma$ perform best under (1). Section 6 studies other relevant DGPs and alternatives under which $\lambda_{\max}$ and $t^{\mathrm{ADF}}_\gamma$ outperform $\hat{F}$ and $t^{\mathrm{ECR}}_\gamma$. Consequently $\tilde{\chi}^2_{\mathcal{I}}(4)$ then outperforms $\tilde{\chi}^2_{\mathcal{I}}(\hat{F},t^{\mathrm{ECR}}_\gamma)$. It would thus be wrong to recommend routine use of $\hat{F}$, $t^{\mathrm{ECR}}_\gamma$ or $\tilde{\chi}^2_{\mathcal{I}}(\hat{F},t^{\mathrm{ECR}}_\gamma)$. Overall, the transparent strategy to combine all available tests can be recommended for practice. On the other hand, $t^{\mathrm{ADF}}_\gamma$ and $\lambda_{\max}$ are still the most widely used tests, such that studying their combination likely is relevant for practitioners.

The $\tilde{\chi}^2_{\mathcal{I}}$ are somewhat more powerful than the $UR_{\psi_{\mathcal{I}}}$ when both constituent tests are relatively powerful. The $UR_{\psi_{\mathcal{I}}}$ outperform the $\tilde{\chi}^2_{\mathcal{I}}$ when there is a large power difference between the

Figure 1: Local asymptotic power as a function of $R^2$, $c = -15$




Case (*iii*). $\chi^2_{\text{BERC}}$ is our Fisher test (3) based on Boswijk's and Banerjee *et al.*'s tests. $\chi^2_{\text{EJ}}$ is based on Engle and Granger's and Johansen's tests. $\chi^2_{\text{BERCEJ}}$ combines all four tests. $UR^{\text{asym}}_{\text{BERC}}$ and $UR^{\text{asym}}_{\text{EJ}}$ are the corresponding $UR_{\psi_{\mathcal{I}}}$ tests (4). The individual tests' curves are for comparison.

individual tests. This is intuitive as $UR_{\psi_{\mathcal{I}}}$ looks for only one individual rejecting test, to then effectively ignore the other. On the other hand, $\tilde{\chi}^2_{\mathcal{I}}$ combines evidence from both tests, such that a test with low power can lead $\tilde{\chi}^2_{\mathcal{I}}$ to accept. If both tests are moderately powerful, $\tilde{\chi}^2_{\mathcal{I}}$ will reject.

An interesting issue raised by an anonymous referee is to relate these results to the power envelope traced out by the Neyman-Pearson tests for model (1). This task has however proved analytically challenging, and only results for the no deterministics case (Elliott and Pesavento, 2009, case (*i*)) and the case of known cointegration vectors (Elliott *et al.*, 2005) seem to be feasible. The latter may be relevant in some applications where economic theory predicts specific cointegrating relationships, and allows Elliott *et al.* (2005) to quantify the loss from having to estimate the cointegration vector for several cointegration tests. For case (*iii*) considered here, they find the power envelope to be 0.09, 0.27, 0.58 and 0.86 for $c = 5, 10, 15$ and 20 for $R^2 = 0$ and 0.29, 0.77, 0.97 and 1 for $R^2 = 0.5$, if the cointegration vector is known. Comparing these values to Table 3 we find that the best tests as well as the best-performing meta tests come close to this envelope for small $c$, and, trivially, for $c$ sufficiently large that the best single tests achieves asymptotic power of 1. For intermediate $c$ the need to estimate the cointegration vector in practice for both single and meta tests translates into (expected) larger asymptotic power losses.

*Remark 2.* As discussed, some individual tests are most powerful when $R^2$ is low, and others when $R^2$ is large. One might hence also consider a pretest strategy selecting the best test given an estimate $\hat{R}^2$. However, as several quantities are not consistently estimable in the present framework, such an estimator is likely not feasible (Pesavento, 2007). Moreover, the combination tests are never much less powerful than the best individual test, and generally a lot more powerful

than the worst test. If an estimator was available, it would not, certainly not for $T$ finite, estimate $R^2$ without error. Hence, a pretest would sometimes select the *less* powerful test and thus likely have less power than the strategies advocated here. To illustrate, let $q$ the probability to select the inferior test. Consider e.g. from Table 3 $\lambda_{\max}$, $t_\gamma^{\text{ADF}}$ and $\tilde{\chi}^2_{\mathcal{I}}(t_\gamma^{\text{ADF}}, \lambda_{\max})$ for $R^2 = 0.75$ and $c = -15$. A pretest would need to select the worse test ($t_\gamma^{\text{ADF}}$) in only $q = (0.811 - 0.937)/(0.071 - 0.937) \approx 14.5\%$ of the cases to be inferior to $\tilde{\chi}^2_{\mathcal{I}}(t_\gamma^{\text{ADF}}, \lambda_{\max})$. Unreported calculations show that, for e.g. $\tilde{\chi}^2_{\mathcal{I}}(t_\gamma^{\text{ADF}}, \lambda_{\max})$, $c = -15$ and $K = 2$, $q$ never exceeds 35%, and even $q = 0$ for $R^2 \in [0.15, 0.25] \cup (0.9, 1)$, reflecting that $\tilde{\chi}^2_{\mathcal{I}}$ then is *more* powerful than even a perfect pretest. Moreover, as $q \ll 0.5$, $\tilde{\chi}^2_{\mathcal{I}}$ uniformly outperforms randomly selecting one underlying test.

# 5 Bootstrap Analogs

The previous results rely on asymptotic theory. The combination tests will also share small-sample deficiencies of the underlying tests. Haug (1996) found the small-sample behavior of cointegration tests to be somewhat sensitive to e.g. short-run dynamics. In particular, finite-sample sizes depend on the estimation method for these and other nuisance parameters. The bootstrap has recently been successfully used to improve the small-sample behavior of cointegration tests (Swensen, 2006; Palm *et al.*, 2010). We therefore now introduce bootstrap analogs of the combination tests to provide potentially more reliable small sample inference.

To bootstrap $\tilde{\chi}^2_{\mathcal{I}}$, we require a method to bootstrap cointegration tests. A suitable procedure has recently been proposed by Swensen (2006). In brief, Swensen's procedure resamples residuals from an estimated VECM representation of the DGP to then generate integrated but non-cointegrated time series. We propose the following algorithm to estimate the finite-sample distribution of $\tilde{\chi}^2_{\mathcal{I}}$.

**Algorithm 1.**

1. *Estimate the unrestricted VAR* $\boldsymbol{z}_t = \sum_{p=1}^{P} \boldsymbol{\Phi}_p \boldsymbol{z}_{t-p} + \boldsymbol{d}_t + \boldsymbol{\varepsilon}_t$ *to obtain estimates* $\hat{\boldsymbol{d}}_t, \hat{\boldsymbol{\Phi}}_p$ *and residuals* $\hat{\boldsymbol{\varepsilon}}_t$. *Transform* $\hat{\boldsymbol{\Phi}}_p$, $p = 1, \ldots, P$, *to* $\hat{\boldsymbol{\Gamma}}_p$, $p = 1, \ldots, P-1$, *as in* (2).[9]
2. *Check that the system has no explosive root, i.e.* $\|z\| > 1$, *by solving* $\det\{\hat{\boldsymbol{B}}(z)\} = 0$, *where* $\hat{\boldsymbol{B}}(z) := \boldsymbol{I}_K - \hat{\boldsymbol{\Gamma}}_1 z - \cdots - \hat{\boldsymbol{\Gamma}}_{P-1} z^{P-1}$.[10]
3. *If so, resample* $\{\boldsymbol{\varepsilon}^*_{t,b}\}_{t=P,\ldots,T}^{b=1,\ldots,B}$ *non-parametrically with replacement from* $\{\hat{\boldsymbol{\varepsilon}}_t\}_{t=P,\ldots,T}$.
4. *With* $\{\boldsymbol{\varepsilon}^*_{t,b}\}_{t=P,\ldots,T}^{b=1,\ldots,B}$, *construct* $B$ *series of pseudo observations* $\boldsymbol{z}^*_{t,b}$ *from* $\Delta\boldsymbol{z}^*_{t,b} = \hat{\boldsymbol{d}}_t + \sum_{p=1}^{P-1} \hat{\boldsymbol{\Gamma}}_p \Delta\boldsymbol{z}^*_{t-p,b} + \boldsymbol{\varepsilon}^*_{t,b}$. *For the initial observations, set* $\boldsymbol{z}^*_{t,b} = \boldsymbol{z}_t, t = 0, \ldots, P-1$.[11]
5. *Compute the vector of test statistics* $\boldsymbol{\xi}^*_b := (\xi^*_{1,b}, \ldots, \xi^*_{|\mathcal{I}|,b})'$, *for each* $b = 1, \ldots, B$.
6. *Estimate the cdf of each statistic as* $B^{-1} \sum_{h=1}^{B} \mathbb{I}\{\xi^*_{i,h} \leq x\} =: 1 - \Xi^*_i(x)$ *and calculate p-values* $p^*_{i,b} := \Xi^*_i(\xi^*_{i,b})$. *Calculate the p-values of the* $\xi_i$ *on the original data,* $p^*_i := \Xi^*_i(\xi_i)$.
7. *Obtain the corresponding aggregate* $\tilde{\chi}^2_{\mathcal{I}}$ *test statistic* $\tilde{\chi}^{2,*}_{\mathcal{I},b} = -2 \sum_{i=1}^{|\mathcal{I}|} \ln(p^*_{i,b})$.
8. *Estimate the distribution function* $F_{\mathcal{F}^*_{\mathcal{I}}}$ *of the* $\tilde{\chi}^{2,*}_{\mathcal{I},b}$ *by* $\hat{F}_{\mathcal{F}^*_{\mathcal{I}}}(x) := B^{-1} \sum_{h=1}^{B} \mathbb{I}\{\tilde{\chi}^{2,*}_{\mathcal{I},h} \leq x\}$.

This yields a bootstrap version of $\tilde{\chi}^2_{\mathcal{I}}$, $\tilde{\chi}^{2,*}_{\mathcal{I}} = -2\sum_{i=1}^{|\mathcal{I}|} \ln(p_i^*)$. Reject at level $\alpha$ if $\tilde{\chi}^{2,*}_{\mathcal{I}} > \hat{F}^{-1}_{\mathcal{F}^*_{\mathcal{I}}}(1-\alpha)$. Heuristically, the method can be expected to work as follows. Swensen (2006) proves that his procedure (steps 1-4 in Algorithm 1) yield $\boldsymbol{z}^*_{t,b}$ which have a representation asymptotically equivalent to the true DGP. Moreover, he proves that steps 5 and 6 consistently estimate the null distribution of the $\lambda_{\text{trace}}$ test, hence yielding consistent estimates of $p$-values. Therefore, we can expect the proposition to carry over to the above cointegration tests, as these essentially also rely on the availability of suitable $\boldsymbol{z}^*_{t,b}$. The CMT with $\boldsymbol{\xi} := (\xi_1, \ldots, \xi_{|\mathcal{I}|})'$ as functions of the observations $\boldsymbol{z}_t$, for which an invariance principle holds, ensures a well-defined *joint* distribution of $\boldsymbol{\xi}$. That joint distribution can be consistently estimated with Algorithm 1 under fairly weak regularity conditions (Horowitz, 2001). Section 6 provides numerical support for this argument.

*Remark 3.* Algorithm 1 is only about as computationally demanding as Swensen's (2006). It also requires resampling $B$ pseudo-observations, and no double bootstrapping. The difference to his algorithm is that $|\mathcal{I}|$ instead of one statistic ($\lambda_{\text{trace}}$) need to be calculated for each $b$.

# 6 Monte Carlo Experiments

## 6.1 Setup

We now study the finite-sample properties of the tests in a series of Monte Carlo experiments. We consider four different DGPs:

1. DGP(A): $\Delta x_t = v_{1t}$, $y_t = x_t + u_t$ and $u_t = \rho_T u_{t-1} + v_{2t}$

   closely follows (1). The autoregressive coefficient $\rho_T = 1 + c/T$. $\mathcal{H}_0$ is obtained when $c = 0$. Under the alternative $c = -15$, $\mathcal{H}_1^a$ holds. The $(v_{1t}, v_{2t})'$ are drawn from $\mathcal{N}(\mathbf{0}, \boldsymbol{\Omega})$, where $\boldsymbol{\Omega} = \left(\begin{smallmatrix} 1 & R \\ R & 1 \end{smallmatrix}\right)$. We take $R^2 = 0.25$ (online Appendix E reports results for other $R^2$ and $c$).

2. DGP(B): $\Delta \boldsymbol{z}_t = \boldsymbol{\Pi}_T \boldsymbol{z}_{t-1} + \boldsymbol{\Gamma} \Delta \boldsymbol{z}_{t-1} + \boldsymbol{u}_t$ where $\boldsymbol{\Gamma} = 0.2\boldsymbol{I}_2$ and $\boldsymbol{u}_t = (u_{1t}, u_{2t})' \overset{iid}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{I}_2)$

   still fulfills $\mathcal{H}_1^a$ for $c < 0$, but introduces short-run dynamics. These are nuisance-parameters that do not affect large-sample power, but may impact the tests' finite-sample performance. For (B) $\mathcal{H}_0$ is obtained when $\boldsymbol{\Pi}_T = \mathbf{0}$. We parameterize $\mathcal{H}_1$ by $\boldsymbol{\Pi}_T = \frac{c}{T}(0 \quad 1)'(1 \quad -1)$.[12]

3. DGP(C): $\Delta \boldsymbol{z}_t = \frac{2}{3}\begin{pmatrix} \rho_T - 1 \\ \frac{1}{2}(\rho_T - 1) \end{pmatrix}(1 \quad 1)\boldsymbol{z}_{t-1} + \tilde{\boldsymbol{v}}_t$, $\tilde{\boldsymbol{v}}_t \overset{iid}{\sim} \mathcal{N}\left(\mathbf{0}, \frac{1}{9}\left(\begin{smallmatrix} 8 & 0 \\ 0 & 5 \end{smallmatrix}\right)\right)$

   violates $\mathcal{H}_1^a$: DGP(C) does not impose that $x_t$ has an exact unit root under the local alternative and thus only satisfies $\mathcal{H}_1^b$. Therefore, Pesavento's results and thus the local power curves from Section 4 do not hold under (C). Notwithstanding, DGP(C), first considered by Engle and Granger (1987) (the above representation is given by Elliott *et al.* (2005)), is a plausible cointegration model: $|\rho_T| < 1$ implies the existence of a stationary equilibrium error $a_{2t}$ (note that e.g. Johansen (1995) allows for non-$I(1)$ variables in cointegrating relationships). Resulting from this somewhat unusual property under the alternative, DGP(C) has the advantage

of not imposing weak exogeneity, as DGP(A) does (given absence of short run dynamics, see Elliott *et al.*, 2005, p. 36).

4. DGP(D): $\boldsymbol{z}_t = \begin{pmatrix} x_{2t} \\ x_{1t} \\ y_t \end{pmatrix}$ and $\Delta \boldsymbol{z}_t = \frac{c}{T} \begin{pmatrix} 0 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & -1 & 1 \end{pmatrix} \boldsymbol{z}_{t-1} + 0.2\boldsymbol{I}_3 \Delta \boldsymbol{z}_{t-1} + \boldsymbol{u}_t$, $\boldsymbol{u}_t \overset{iid}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_3)$

   DGP(D) can be considered a trivariate extension of (B) where the researcher erroneously adds to the regressions underlying $t_\gamma^{\mathrm{ADF}}$, $t_\gamma^{\mathrm{ECR}}$ and $\hat{F}$ a variable $x_{2t}$ that turns out to be redundant, i.e. has a zero coefficient in the cointegrating vector. This variable, however, is cointegrated with $x_{1t}$, hence violates Assumption 2 and invalidates $\mathcal{H}_1^a$. Nonetheless, since $x_{1t}$ cointegrates with $y_t$, $\mathcal{H}_1^c$ still holds.

Recall that departures from $\mathcal{H}_1^a$ do not affect the validity of our combination procedure. First, the critical values of $UR_{\psi_\mathcal{I}}$ and $\tilde{\chi}^2_\mathcal{I}$ obviously derive from the properties of the system under $\mathcal{H}_0$, where $\rho_T = 1$ and the above considerations do not apply. Second, the consistency of the tests only requires at least one diverging test statistic (cf. Corollary 1(b)). Divergence of individual test statistics for $|\rho_T| < 1$ is shown by e.g. Phillips and Ouliaris (1990, Thm. 5.1).[13]

The above DGPs are widely used in Monte Carlo studies. See e.g. Pesavento (2004, 2007) for (A), Swensen (2006) for (B), or Engle and Granger (1987), Haug (1996) and Gregory *et al.* (2004) for (C).[14] The DGPs are local, such that power ought to be roughly constant in $T$. We use 5,000 replications. We choose $T \in \{50, 100, 150, 200\}$, corresponding to typical sample sizes encountered in applied work. To mitigate the effect of initial conditions under $\mathcal{H}_1$, we simulate for $T + 30$ periods and discard the first 30 observations. The bootstrap tests use $B = 10,000$ resamples.

We mainly consider $\tilde{\chi}^2_\mathcal{I}(4)$ and compare it to the single $t_\gamma^{\mathrm{ADF}}$, $\lambda_{\max}$, $t_\gamma^{\mathrm{ECR}}$ and $\hat{F}$ tests,[15] as well as to bootstrap versions $t_\gamma^{\mathrm{ADF},*}$, $\lambda^*_{\max}$, $t_\gamma^{\mathrm{ECR},*}$ and $\hat{F}^*$ (these are by-products of Algorithm 1). We also compute the 'naive' $UR^{\mathrm{n}}$ test which reveals the size distortion incurred by selecting the most rejective from a set of tests.

The tests require choosing a lag length $\hat{P}$. We initially impose the correct lag order (i.e. $P = 0$ in (A) and (C) and $P = 1$ in (B) and (D)) in order to focus on size distortions resulting from multiple testing and return to the practically more relevant case of unknown $P$ below. For $t_\gamma^{\mathrm{ADF}}$, we select $P = 1$ under (B) too, as this yields a sufficiently accurate approximation for $\boldsymbol{\Gamma} = 0.2\boldsymbol{I}_2$. All results are for case $(iii)$.

## 6.2 Results

Table 4 reports size at $\alpha = 0.05$.[16] As expected, the 'naive' tests are oversized. Their sizes exceed that of the individual tests by up to 4 percentage points. Both the $UR_{\psi_\mathcal{I}}$ and, to a lesser extent, the $\tilde{\chi}^2_\mathcal{I}$ are slightly upward size distorted for small $T$, due to distortions of the underlying tests. However, this distortion quickly vanishes for larger $T$. The single bootstrap, and hence combination, tests are slightly better sized, as the $\hat{F}_{\mathcal{F}^*_\mathcal{I}}$ more accurately approximate the finite-sample distributions than the asymptotic distributions $F_{\mathcal{F}_\mathcal{I}}$.

Table 4: Small-sample size

| | | $\lambda_{\max}$ and $t_\gamma^{\text{ADF}}$ | | | | | $\hat{F}$ and $t_\gamma^{\text{ECR}}$ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| I. Bootstrap | | | | | | | | | | | | |
| DGP | $T$ | $\lambda_{\max}^*$ | $t_\gamma^{\text{ADF},*}$ | naive$^*$ | $\tilde{\chi}_{\mathcal{I}}^{2,*}$ | $UR_{\psi_{\mathcal{I}}}^*$ | $\hat{F}^*$ | $t_\gamma^{\text{ECR},*}$ | naive$^*$ | $\tilde{\chi}_{\mathcal{I}}^{2,*}$ | $UR_{\psi_{\mathcal{I}}}^*$ | $\tilde{\chi}_{\mathcal{I}}^{2,*}(4)$ |
| (A) | 50 | 0.050 | 0.052 | 0.087 | 0.049 | 0.048 | 0.049 | 0.048 | 0.058 | 0.049 | 0.049 | 0.050 |
| | 100 | 0.053 | 0.054 | 0.090 | 0.052 | 0.054 | 0.055 | 0.053 | 0.063 | 0.054 | 0.055 | 0.057 |
| | 150 | 0.048 | 0.053 | 0.085 | 0.055 | 0.053 | 0.054 | 0.057 | 0.064 | 0.056 | 0.055 | 0.054 |
| | 200 | 0.051 | 0.052 | 0.086 | 0.053 | 0.050 | 0.045 | 0.046 | 0.053 | 0.045 | 0.044 | 0.048 |
| (B) | 50 | 0.054 | 0.055 | 0.087 | 0.051 | 0.053 | 0.053 | 0.057 | 0.066 | 0.055 | 0.055 | 0.055 |
| | 100 | 0.049 | 0.051 | 0.077 | 0.049 | 0.049 | 0.048 | 0.053 | 0.059 | 0.051 | 0.051 | 0.052 |
| | 150 | 0.050 | 0.049 | 0.078 | 0.049 | 0.050 | 0.049 | 0.049 | 0.057 | 0.049 | 0.050 | 0.048 |
| | 200 | 0.048 | 0.049 | 0.074 | 0.047 | 0.047 | 0.046 | 0.048 | 0.055 | 0.046 | 0.046 | 0.047 |
| (C) | 50 | 0.049 | 0.048 | 0.083 | 0.048 | 0.046 | 0.050 | 0.050 | 0.060 | 0.050 | 0.052 | 0.048 |
| | 100 | 0.054 | 0.054 | 0.089 | 0.055 | 0.053 | 0.054 | 0.052 | 0.060 | 0.051 | 0.054 | 0.055 |
| | 150 | 0.047 | 0.051 | 0.082 | 0.048 | 0.049 | 0.051 | 0.048 | 0.057 | 0.049 | 0.050 | 0.048 |
| | 200 | 0.050 | 0.050 | 0.084 | 0.052 | 0.049 | 0.049 | 0.049 | 0.059 | 0.050 | 0.051 | 0.049 |
| (D) | 50 | 0.052 | 0.054 | 0.091 | 0.055 | 0.053 | 0.056 | 0.060 | 0.074 | 0.058 | 0.057 | 0.061 |
| | 100 | 0.050 | 0.053 | 0.086 | 0.052 | 0.050 | 0.054 | 0.053 | 0.065 | 0.056 | 0.054 | 0.054 |
| | 150 | 0.050 | 0.049 | 0.083 | 0.050 | 0.050 | 0.053 | 0.054 | 0.066 | 0.057 | 0.055 | 0.055 |
| | 200 | 0.049 | 0.046 | 0.079 | 0.049 | 0.049 | 0.049 | 0.049 | 0.063 | 0.051 | 0.051 | 0.049 |
| II. Asymptotic | | | | | | | | | | | | |
| DGP | $T$ | $\lambda_{\max}$ | $t_\gamma^{\text{ADF}}$ | naive | $\tilde{\chi}_{\mathcal{I}}^2$ | $UR_{\psi_{\mathcal{I}}}$ | $\hat{F}$ | $t_\gamma^{\text{ECR}}$ | naive | $\tilde{\chi}_{\mathcal{I}}^2$ | $UR_{\psi_{\mathcal{I}}}$ | $\tilde{\chi}_{\mathcal{I}}^2(4)$ |
| (A) | 50 | 0.054 | 0.079 | 0.112 | 0.057 | 0.081 | 0.081 | 0.076 | 0.090 | 0.080 | 0.082 | 0.072 |
| | 100 | 0.050 | 0.060 | 0.094 | 0.042 | 0.061 | 0.065 | 0.063 | 0.075 | 0.062 | 0.064 | 0.055 |
| | 150 | 0.054 | 0.060 | 0.094 | 0.048 | 0.063 | 0.055 | 0.053 | 0.061 | 0.054 | 0.053 | 0.050 |
| | 200 | 0.051 | 0.055 | 0.089 | 0.042 | 0.058 | 0.058 | 0.058 | 0.068 | 0.058 | 0.059 | 0.052 |
| (B) | 50 | 0.065 | 0.074 | 0.110 | 0.067 | 0.078 | 0.077 | 0.072 | 0.087 | 0.075 | 0.078 | 0.071 |
| | 100 | 0.056 | 0.063 | 0.093 | 0.061 | 0.066 | 0.060 | 0.061 | 0.070 | 0.059 | 0.060 | 0.060 |
| | 150 | 0.057 | 0.061 | 0.091 | 0.057 | 0.060 | 0.058 | 0.058 | 0.066 | 0.057 | 0.058 | 0.057 |
| | 200 | 0.051 | 0.053 | 0.081 | 0.050 | 0.052 | 0.050 | 0.050 | 0.057 | 0.051 | 0.051 | 0.049 |
| (C) | 50 | 0.054 | 0.078 | 0.111 | 0.059 | 0.080 | 0.084 | 0.077 | 0.093 | 0.078 | 0.082 | 0.073 |
| | 100 | 0.050 | 0.058 | 0.091 | 0.043 | 0.061 | 0.061 | 0.060 | 0.071 | 0.060 | 0.061 | 0.052 |
| | 150 | 0.054 | 0.059 | 0.095 | 0.050 | 0.066 | 0.058 | 0.059 | 0.068 | 0.060 | 0.060 | 0.054 |
| | 200 | 0.051 | 0.055 | 0.089 | 0.040 | 0.058 | 0.055 | 0.054 | 0.063 | 0.054 | 0.054 | 0.047 |
| (D) | 50 | 0.100 | 0.074 | 0.147 | 0.082 | 0.097 | 0.089 | 0.072 | 0.098 | 0.078 | 0.084 | 0.083 |
| | 100 | 0.077 | 0.063 | 0.113 | 0.068 | 0.074 | 0.074 | 0.064 | 0.085 | 0.068 | 0.070 | 0.070 |
| | 150 | 0.065 | 0.064 | 0.107 | 0.067 | 0.069 | 0.062 | 0.060 | 0.074 | 0.061 | 0.061 | 0.067 |
| | 200 | 0.059 | 0.055 | 0.093 | 0.058 | 0.057 | 0.056 | 0.052 | 0.066 | 0.055 | 0.055 | 0.055 |

Rejection rates at nominal level of 5%. 5,000 replications and 10,000 bootstrap replications. $t_\gamma^{\text{ADF}}$ and $\lambda_{\max}$ refer to Engle and Granger (1987) and Johansen (1988) tests. $\hat{F}$ and $t_\gamma^{\text{ECR}}$ are from Boswijk (1994) and Banerjee *et al.* (1998). naive rejects when $t_\gamma^{\text{ADF},*}$ or $\lambda_{\max}^*$ or both reject. $UR_{\psi_{\mathcal{I}}}$ is the test defined by (4) and (5). $\tilde{\chi}_{\mathcal{I}}^2$ is the Fisher test (3). Starred tests are bootstrap counterparts. $UR^*$ and $\tilde{\chi}_{\mathcal{I}}^{2,*}$ are described in Algorithm 1. $\tilde{\chi}_{\mathcal{I}}^2(4)$ combines all four tests.

Table 5 reports size-adjusted small sample power of the asymptotic tests.[17] As one might have expected, for DGP(A), the results from Section 4 predict the finite-sample results well, in that $t_\gamma^{\text{ADF}}$, $\lambda_{\max}$ and $\tilde{\chi}_{\mathcal{I}}^2(t_\gamma^{\text{ADF}}, \lambda_{\max})$ again have similar power for this $R^2$. Further, $\hat{F}$ and $t_\gamma^{\text{ECR}}$ are the most powerful single tests. The rightmost column of Table 4 shows that $\tilde{\chi}_{\mathcal{I}}^2(4)$ outperforms $\tilde{\chi}_{\mathcal{I}}^2(t_\gamma^{\text{ADF}}, \lambda_{\max})$ rather markedly. Noticeably, $\tilde{\chi}_{\mathcal{I}}^2(4)$ is only slightly outperformed by $\hat{F}$, $t_\gamma^{\text{ECR}}$ and

Table 5: Small-sample power

| | | $\lambda_{\max}$ and $t_\gamma^{\text{ADF}}$ | | | | $\hat{F}$ and $t_\gamma^{\text{ECR}}$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| DGP | $T$ | $\lambda_{\max}$ | $t_\gamma^{\text{ADF}}$ | $\tilde{\chi}^2_{\mathcal{I}}$ | $UR_{\psi_{\mathcal{I}}}$ | $\hat{F}$ | $t_\gamma^{\text{ECR}}$ | $\tilde{\chi}^2_{\mathcal{I}}$ | $UR_{\psi_{\mathcal{I}}}$ | $\tilde{\chi}^2_{\mathcal{I}}(4)$ |
| (A) | 50 | 0.274 | 0.257 | 0.326 | 0.269 | 0.433 | 0.413 | 0.433 | 0.428 | 0.421 |
| | 100 | 0.271 | 0.257 | 0.341 | 0.254 | 0.422 | 0.398 | 0.411 | 0.416 | 0.398 |
| | 150 | 0.251 | 0.234 | 0.300 | 0.246 | 0.455 | 0.425 | 0.446 | 0.447 | 0.413 |
| | 200 | 0.275 | 0.240 | 0.319 | 0.264 | 0.429 | 0.382 | 0.407 | 0.413 | 0.396 |
| (B) | 50 | 0.380 | 0.309 | 0.402 | 0.383 | 0.270 | 0.237 | 0.253 | 0.257 | 0.330 |
| | 100 | 0.554 | 0.347 | 0.517 | 0.508 | 0.378 | 0.326 | 0.356 | 0.368 | 0.436 |
| | 150 | 0.596 | 0.374 | 0.574 | 0.567 | 0.437 | 0.368 | 0.408 | 0.421 | 0.516 |
| | 200 | 0.636 | 0.377 | 0.619 | 0.607 | 0.455 | 0.386 | 0.418 | 0.434 | 0.560 |
| (C) | 50 | 0.174 | 0.275 | 0.272 | 0.217 | 0.207 | 0.240 | 0.229 | 0.210 | 0.257 |
| | 100 | 0.168 | 0.285 | 0.281 | 0.223 | 0.213 | 0.240 | 0.231 | 0.217 | 0.261 |
| | 150 | 0.162 | 0.258 | 0.247 | 0.201 | 0.214 | 0.238 | 0.231 | 0.220 | 0.254 |
| | 200 | 0.176 | 0.258 | 0.276 | 0.218 | 0.216 | 0.239 | 0.233 | 0.227 | 0.277 |
| (D) | 50 | 0.377 | 0.385 | 0.488 | 0.431 | 0.190 | 0.150 | 0.169 | 0.186 | 0.306 |
| | 100 | 0.564 | 0.451 | 0.619 | 0.576 | 0.246 | 0.212 | 0.228 | 0.236 | 0.416 |
| | 150 | 0.644 | 0.465 | 0.670 | 0.624 | 0.322 | 0.257 | 0.289 | 0.305 | 0.513 |
| | 200 | 0.688 | 0.509 | 0.724 | 0.690 | 0.339 | 0.276 | 0.303 | 0.315 | 0.574 |

See notes to Table 4. $R^2 = 0.25$ (for DGP(A)) and $c = -15$.

$\tilde{\chi}^2_{\mathcal{I}}(\hat{F}, t_\gamma^{\text{ECR}})$ even under (A).

For DGP(B) we find $\lambda_{\max}$ to be the most powerful individual test, followed by the meta tests involving $\lambda_{\max}$. Noteworthy again, the decrease in power by moving from the most powerful to the meta test is small, while the power gain from the least powerful test ($t_\gamma^{\text{ECR}}$) is large. E.g., for $T = 200$, $\tilde{\chi}^2_{\mathcal{I}}(4)$ loses 7.5 percentage points to $\lambda_{\max}$ but gains over 17 points to $t_\gamma^{\text{ECR}}$.

For DGP(C), $t_\gamma^{\text{ADF}}$ is the most powerful individual test. Again, $\tilde{\chi}^2_{\mathcal{I}}(4)$ is a close second, and even the best for $T = 200$. Hence, when testing against $\mathcal{H}_1^b$, the ranking of local power curves from Section 4 no longer holds. Consequently, it would be premature to recommend routine application of either $\hat{F}$ or $t_\gamma^{\text{ECR}}$ provided the researcher is interested in testing against alternatives such as $\mathcal{H}_1^b$.

The results for DGP(D) should be compared to those for (B). Again, $\lambda_{\max}$ is the most powerful individual test and even gains power relative to (B). This is not surprising as there are now two cointegration relationships in the system, and $\lambda_{\max}$ rejects if it detects at least one of them. Likewise, $t_\gamma^{\text{ADF}}$ gains power. On the other hand, $\hat{F}$ and $t_\gamma^{\text{ECR}}$ lose power. Apparently, violation of Assumption 2 is detrimental for the power of the error-correction based tests. Nonetheless, $\tilde{\chi}^2_{\mathcal{I}}(4)$ outperforms all individual tests except $\lambda_{\max}$, and is much closer to $\lambda_{\max}$ than to either $\hat{F}$ or $t_\gamma^{\text{ECR}}$.

For each DGP, there is always a meta test combining two tests that slightly outperforms $\tilde{\chi}^2_{\mathcal{I}}(4)$. However, there is always another DGP for which $\tilde{\chi}^2_{\mathcal{I}}(4)$ clearly outperforms that meta test. Overall, the other meta tests hence sometimes have higher power, but are less robust than $\tilde{\chi}^2_{\mathcal{I}}(4)$.

As stated above, the lag length $P$ is not known in practice, and it would be a severe limitation of our procedures if it one needed to know $P$ for them to work. We therefore now provide some results when $P$ is chosen with the BIC. Table 6 shows that the above qualitative findings remain

Table 6: Rejection rates under lag-length selection

| | | $\lambda_{\max}$ and $t_\gamma^{\text{ADF}}$ | | | | | $\hat{F}$ and $t_\gamma^{\text{ECR}}$ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| (a) Size | | | | | | | | | | | | |
| Bootstrap | $T$ | $\lambda_{\max}^*$ | $t_\gamma^{\text{ADF},*}$ | naive$^*$ | $\tilde{\chi}_{\mathcal{I}}^{2,*}$ | $UR_{\psi_{\mathcal{I}}}^*$ | $\hat{F}^*$ | $t_\gamma^{\text{ECR},*}$ | naive$^*$ | $\tilde{\chi}_{\mathcal{I}}^{2,*}$ | $UR_{\psi_{\mathcal{I}}}^*$ | $\tilde{\chi}_{\mathcal{I}}^{2,*}(4)$ |
| | 50 | 0.101 | 0.070 | 0.140 | 0.083 | 0.090 | 0.108 | 0.110 | 0.129 | 0.110 | 0.111 | 0.105 |
| | 100 | 0.076 | 0.063 | 0.108 | 0.070 | 0.073 | 0.074 | 0.073 | 0.086 | 0.075 | 0.072 | 0.077 |
| | 150 | 0.061 | 0.052 | 0.088 | 0.057 | 0.057 | 0.058 | 0.057 | 0.069 | 0.059 | 0.060 | 0.058 |
| | 200 | 0.059 | 0.055 | 0.089 | 0.055 | 0.055 | 0.054 | 0.057 | 0.065 | 0.056 | 0.055 | 0.055 |
| Asymptotic | $T$ | $\lambda_{\max}$ | $t_\gamma^{\text{ADF}}$ | naive | $\tilde{\chi}_{\mathcal{I}}^2$ | $UR_{\psi_{\mathcal{I}}}$ | $\hat{F}$ | $t_\gamma^{\text{ECR}}$ | naive | $\tilde{\chi}_{\mathcal{I}}^2$ | $UR_{\psi_{\mathcal{I}}}$ | $\tilde{\chi}_{\mathcal{I}}^2(4)$ |
| | 50 | 0.146 | 0.085 | 0.184 | 0.107 | 0.137 | 0.108 | 0.094 | 0.119 | 0.098 | 0.107 | 0.101 |
| | 100 | 0.092 | 0.073 | 0.128 | 0.080 | 0.094 | 0.076 | 0.073 | 0.085 | 0.072 | 0.076 | 0.076 |
| | 150 | 0.080 | 0.065 | 0.109 | 0.071 | 0.078 | 0.068 | 0.064 | 0.076 | 0.064 | 0.066 | 0.067 |
| | 200 | 0.061 | 0.055 | 0.089 | 0.056 | 0.059 | 0.054 | 0.055 | 0.061 | 0.055 | 0.055 | 0.056 |
| (b) Power | | | | | | | | | | | | |
| Bootstrap | $T$ | $\lambda_{\max}^*$ | $t_\gamma^{\text{ADF},*}$ | naive$^*$ | $\tilde{\chi}_{\mathcal{I}}^{2,*}$ | $UR_{\psi_{\mathcal{I}}}^*$ | $\hat{F}^*$ | $t_\gamma^{\text{ECR},*}$ | naive$^*$ | $\tilde{\chi}_{\mathcal{I}}^{2,*}$ | $UR_{\psi_{\mathcal{I}}}^*$ | $\tilde{\chi}_{\mathcal{I}}^{2,*}(4)$ |
| | 50 | 0.193 | 0.206 | 0.461 | 0.245 | 0.223 | 0.147 | 0.126 | 0.313 | 0.134 | 0.135 | 0.163 |
| | 100 | 0.404 | 0.311 | 0.594 | 0.408 | 0.402 | 0.283 | 0.258 | 0.397 | 0.268 | 0.280 | 0.341 |
| | 150 | 0.539 | 0.349 | 0.651 | 0.523 | 0.516 | 0.379 | 0.323 | 0.433 | 0.352 | 0.355 | 0.454 |
| | 200 | 0.592 | 0.353 | 0.692 | 0.576 | 0.564 | 0.427 | 0.353 | 0.464 | 0.390 | 0.408 | 0.504 |
| Asymptotic | $T$ | $\lambda_{\max}$ | $t_\gamma^{\text{ADF}}$ | naive | $\tilde{\chi}_{\mathcal{I}}^2$ | $UR_{\psi_{\mathcal{I}}}$ | $\hat{F}$ | $t_\gamma^{\text{ECR}}$ | naive | $\tilde{\chi}_{\mathcal{I}}^2$ | $UR_{\psi_{\mathcal{I}}}$ | $\tilde{\chi}_{\mathcal{I}}^2(4)$ |
| | 50 | 0.200 | 0.233 | 0.547 | 0.269 | 0.232 | 0.200 | 0.171 | 0.334 | 0.188 | 0.190 | 0.218 |
| | 100 | 0.418 | 0.292 | 0.655 | 0.434 | 0.416 | 0.318 | 0.266 | 0.430 | 0.289 | 0.306 | 0.366 |
| | 150 | 0.509 | 0.339 | 0.696 | 0.503 | 0.492 | 0.397 | 0.343 | 0.477 | 0.370 | 0.383 | 0.461 |
| | 200 | 0.595 | 0.357 | 0.699 | 0.591 | 0.578 | 0.439 | 0.360 | 0.470 | 0.399 | 0.410 | 0.511 |

This DGP studies DGP(B) where the number of lags is chosen with the BIC. To avoid trivial duplication with $UR_{\psi_{\mathcal{I}}}$, the power of the naive test is not size-adjusted. See also notes to Table 4.

intact. Unsurprisingly, size is somewhat worse for small $T$. Size-adjusted power is similar to what we found above. As expected, the bootstrap tests exhibit slightly better size behavior overall.[18]

In summary, we find our meta tests, in particular $\tilde{\chi}_{\mathcal{I}}^2(4)$, to be attractive because they are robust and offer cheap insurance when the researcher is interested in testing against more general alternatives such as $\mathcal{H}_1^b$ and $\mathcal{H}_1^c$, not knowing whether Assumptions 1 and 2 hold for data under study.[19] Yet, such knowledge about the DGP will rarely be available in practice (cf. Remark 2). Indeed, it seems implausible that one wishes to conduct inference about a key feature of the time series—cointegration—whilst having accurate knowledge about e.g. the processes generating $\boldsymbol{x}_t$.

# 7 Mixed Signals Revisited

## 7.1 Setup

We revisit the studies Gregory *et al.* (2004) investigated for 'mixed signals', i.e. conflicting cointegration tests. They analyze 34 studies published in the Journal of Applied Econometrics from 1994 to March/April 2001. We perform an analogous exercise for the issues from May/June 2001 to papers scheduled for publication as of August 2010.[20] In total, we construct 286 data sets. Of

these, 127 are from after April 2001. Thus, cointegration continues to receive unabated attention. When necessary, we perform some preliminary data transformations such as removal of seasonal patterns. We have many more tests than studies because, e.g., we can calculate several time-series tests from panel data studies. The data sets show large differences in sample size, ranging from 24 to 7693. The number of variables $K$ ranges from 2 to 11.

We document the extent to which mixed signals arise in applications and how our meta tests can heal this problem. We do not suggest that the authors of the studies have been strategic in choosing which test to report. The original studies employ different specifications. To make results comparable, we follow Gregory *et al.* (2004) in imposing a unifying but standard methodology. If a test requires a dependent variable $y_t$, we follow the choice of the original paper if possible. If there is no obvious $y_t$, we choose it based on the highest coefficient of determination of first-stage regressions. We also need to allow for variation in lag lengths $\hat{P}$ across data sets. We determine $\hat{P}$ using the BIC as described e.g. in Lütkepohl (2005, Sec. 4.3.2). We search over $1 \leq \hat{P} \leq \min\left(8\left(\frac{T}{100}\right)^{1/5}, \frac{T-2}{2(K+2)}\right)$, and impose the same $\hat{P}$ for all tests. Our qualitative conclusions would be the same for other selection methods for $\hat{P}$. All tests include a constant and a trend.

### 7.2 Results

We compare the results of individually using $\lambda_{\max}$, $t_\gamma^{\mathrm{ADF}}$, $t_\gamma^{\mathrm{ECR}}$ and $\hat{F}$ with $\tilde{\chi}^2_{\mathcal{I}}(4)$. First reconsider the noteworthy patterns of rejections from Section 2.1. For Clements and Hendry (1995), $\lambda_{\max}$ and $\hat{F}$ rejected, showing that the related $t_\gamma^{\mathrm{ECR}}$ and $\hat{F}$ may not agree for the *same* samples. Only $t_\gamma^{\mathrm{ADF}}$ rejected in Cooley and Ogaki (1996), showing that $t_\gamma^{\mathrm{ADF}}$, often thought to be less powerful, rejects while the system- and error-correction based tests do not. In Martens *et al.* (1998), $\lambda_{\max}$ and $t_\gamma^{\mathrm{ADF}}$ rejected, showing that $t_\gamma^{\mathrm{ECR}}$ and $\hat{F}$ may not reject although $\lambda_{\max}$ does. Overall, they show that mixed signals do not stem from a single test always or never rejecting.

How does $\tilde{\chi}^2_{\mathcal{I}}(4)$ resolve these mixed signals? For example one, $\tilde{\chi}^2_{\mathcal{I}}(4) = 39.870$ exceeds the 5% critical value of 22.215, hence agreeing with $\lambda_{\max}$ and $\hat{F}$. On the other hand, $\tilde{\chi}^2_{\mathcal{I}}(4) = 16.126$ for example two, i.e. $\tilde{\chi}^2_{\mathcal{I}}(4)$ joins the non-rejecting $\lambda_{\max}$, $t_\gamma^{\mathrm{ECR}}$ and $\hat{F}$. Apparently, the $p$-value of $t_\gamma^{\mathrm{ADF}}$ is not small enough to have $\tilde{\chi}^2_{\mathcal{I}}(4)$ reject. The very small $p$-values in example three produce a large, and thus rejecting, meta statistic. Hence, $\tilde{\chi}^2_{\mathcal{I}}(4)$ aggregates the single tests such that, depending on the relative strengths of rejection and acceptance, either aggregate result can obtain.

More generally, we check whether all individual tests from the Gregory *et al.* (2004) data and the updated set agree or not in their decision, see left panel of Table 7. If there are conflicts we check which result the test used in the original paper had suggested (more precisely what would have been the outcome of our version with e.g. the chosen lag-length criterion), see the right panel of Table 7.[21] We then compare the results to that of $\tilde{\chi}^2_{\mathcal{I}}(4)$.

Table 7 thus reports the frequencies for all possible outcomes.[22] When all tests do or do not reject, the meta test does so too. However, agreeing tests make up only 65% (= (56 + 131)/286) of all

Table 7: Test results in applied studies and the $\tilde{\chi}^2_{\mathcal{I}}$ test

| number of cases in which... | ...individual test results... | | | | ...in case of conflicting results: 'preferred' test† | | | |
|---|---|---|---|---|---|---|---|---|
| | agree | | conflict | | | | | |
| | $r$ | $\neg r$ | | $\sum$ | | $r$ | $\neg r$ | $\sum$ |
| $\tilde{\chi}^2_{\mathcal{I}}(4) : r$ | 56 | 0 | 46 | 102 | $\tilde{\chi}^2_{\mathcal{I}}(4) : r$ | 25 | 12 | 37 |
| $\tilde{\chi}^2_{\mathcal{I}}(4) : \neg r$ | 0 | 131 | 53 | 184 | $\tilde{\chi}^2_{\mathcal{I}}(4) : \neg r$ | 26 | 14 | 40 |
| $\sum$ | 56 | 131 | 99 | 286 | $\sum$ | 51 | 26 | 77 |

$r$: test rejects; $\neg r$: test does not reject (at $\alpha = 0.05$). †: Test type on which conclusions in the original study were based (see fn. 21). Absolute frequencies of cointegration-test results. Individual tests are Engle and Granger (1987), Boswijk (1994), Banerjee *et al.* (1998) and Johansen (1988) tests. The $\tilde{\chi}^2_{\mathcal{I}}(4)$ abbreviates $\tilde{\chi}^2_{\mathcal{I}}(\lambda_{\max}, t_\gamma^{\text{ADF}}, t_\gamma^{\text{ECR}}, \hat{F})$ and combines these tests as described in Sec. 3.

data sets. For the remaining 35% of conflicting results $\tilde{\chi}^2_{\mathcal{I}}(4)$ is most useful, yielding a definite conclusion. In 54% (= 53/99) of the cases $\tilde{\chi}^2_{\mathcal{I}}(4)$ accepts. In the remaining cases $\tilde{\chi}^2_{\mathcal{I}}(4)$ rejects.

Moreover, rejecting whenever at least one (but not all) of the tests rejected would have lead to a substantial overstatement of cointegration (99 vs. 46 cases). Similarly, the conservative strategy of only rejecting when all tests reject would have understated the pervasiveness of cointegration. Also, the tests that have been 'preferred' in the studies are more rejective than our meta test (51 vs. 37 rejections in 77 tests). Hence, evidence for cointegration would have been less pronounced if the studies could have relied on a suitable meta test.[23] Finally, whether or not the preferred test rejected is not informative on whether or not $\tilde{\chi}^2_{\mathcal{I}}$ rejects conditional on observing 'mixed signals'. This is reflected by similar conditional probabilities: $53/99 \simeq 26/51 \simeq 14/26 \approx 1/2$. Thus, we cannot infer from a published result what the $\tilde{\chi}^2_{\mathcal{I}}$ test would indicate, conditional on a further individual test leading to a conflicting test result.

## 8 Conclusion

This paper proposes meta tests that combine information from individual cointegration tests. The tests take into account the multiple testing nature of running several individual tests and hence control size. The meta tests find and employ the distribution of aggregators of the underlying tests (e.g., Fisher's), by appropriately modifying the critical values of the underlying tests, as well as by bootstrap methods. By contrast, running more than one test and drawing inferences from the most rejective test yields an oversized test. Asymptotic and Monte Carlo results establish attractive power properties. An application to a large and up-to-date set of studies confirms our tests' practical value, yielding an unambiguous test decision in cases of conflicting individual tests.

The setup we put forward is fairly general and hence can be adopted to other testing problems for which several (imperfectly correlated) tests exist. Essentially, only the distribution of a suitable aggregator or an appropriate bootstrap method are needed. Examples include testing for unit roots or heteroscedasticity, for which the sieve and wild bootstrap would be suitable.

A major practical advantage of our proposed tests is that they relieve the applied researcher from the discretionary and often arbitrary choice between cointegration tests to reach a decision.

# Notes

[1]STATA and MATLAB code implementing the procedures suggested in this paper is available at www.rug.nl/staff/c.h.hanck/research.

[2]We do *not* suggest that the authors of the studies have been in any way strategic in their choice of which cointegration test to report. In fact, since we impose (see Section 7 for details) a common selection procedure regarding e.g. trend and lag length selection, our results could possibly differ from what the authors would have found. Also, cointegration testing may or may not have been a key concern in any of the applied work studied here.

[3]Kremers *et al.*'s (1992) 'common factor restriction' is an example for $R^2 = 0$.

[4]Clearly, it would be nice to express the limiting random variable of $\tilde{\chi}^2_{\mathcal{I}}$ as an explicit functional of $\boldsymbol{W}$. We think this is difficult analytically, as it is complicated and possible in special cases only even for sums of standard and independent random variables. Here, the $\xi_i$ are nonstandard and dependent in a complicated way.

[5]We obtain these from 100,000 draws from the $F_{\mathcal{F}_{\mathcal{I}}}$, approximating $\boldsymbol{W}$ with suitably normalized Gaussian random walks of length $T = 1,000$.

[6]The null rejection probability of test $i$ is $\Pr(\xi_i > cv_i^{\alpha}) = \alpha$. The size of $UR^{\text{n}}(\xi_1, \xi_2)$ therefore is $\Pr(\bigcup_{i=1}^{2} \xi_i > cv_i^{\alpha}) = \sum_{i=1}^{2} \Pr(\xi_i > cv_i^{\alpha}) - \Pr(\bigcap_{i=1}^{2} \xi_i > cv_i^{\alpha}) = 2\alpha - \Pr(\bigcap_{i=1}^{2} \xi_i > cv_i^{\alpha}) \geqslant \alpha$, as $\Pr(\bigcap_{i=1}^{2} \xi_i > cv_i^{\alpha}) \leqslant \Pr(\xi_i > cv_i^{\alpha}) = \alpha$.

[7]Appendix A gives a formal argument. One can further show that (5) minimizes the instances where both $\xi_i$ reject under $\mathcal{H}_0$, i.e. the tests are made as 'uncorrelated' as possible. A detailed argument is available, see also B.4.

[8]We simulate critical values for $R^2 = 0$. Thus, the deviations from $\alpha$ for $R^2 \neq 0$ are due to simulation variability.

[9]See e.g. Lütkepohl (2005, p. 247) for the procedure. One could also estimate a VAR for $\Delta \boldsymbol{z}_t$, imposing $\mathcal{H}_0$ (cf. Swensen, 2006). However, as Paparoditis and Politis (2003) show for unit-root tests, this may lead to lower power.

[10]See Swensen (2006, Remark 1) and Johansen (1995, p. 71) for a discussion of this condition. Note that under $h = 0$, $\hat{\boldsymbol{\alpha}}\hat{\boldsymbol{\beta}}' = \boldsymbol{0}$ in Swensen's notation, such that we have $\hat{A}(z) = (1 - z)\hat{\boldsymbol{B}}(z)$, with the l.h.s. in Swensen's notation again. Thus his condition (iii) is equivalent to step 2 of our algorithm.

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

[12]Elliott *et al.* (2005) show that variants of DGP(A) and (B) are closely related, yet they differ in how short-run dynamics enter the DGPs. DGP(B) can be written as

$$[(\boldsymbol{I} - \boldsymbol{\Gamma} L)(1 - L) - (\rho_T - 1)\boldsymbol{\Pi}_T L]\boldsymbol{z}_t = \boldsymbol{u}_t \tag{6}$$

whereas an equivalent way of writing (1) for the corresponding case of a VAR(2) is

$$[(\boldsymbol{I} - \boldsymbol{\Phi} L)(1 - L) - (\boldsymbol{I} - \boldsymbol{\Phi} L)(\rho_T - 1)\boldsymbol{\Pi}_T L]\boldsymbol{z}_t = \boldsymbol{u}_t \tag{7}$$

(see Pesavento's eq. (2.1)). As $\boldsymbol{I} - \boldsymbol{\Phi} L$ also affects the error-correction term in (7) one cannot find a $\boldsymbol{\Phi}$ such that (6) and (7) imply the same dynamics in our parametrization. This also implies that it is no longer directly possible to infer the $R^2$s for DGP(B).

[13]They e.g. show that $t_{\gamma}^{\text{ADF}} = \mathcal{O}_p(T^{1/2})$ under cointegration in a triangular system like (1). This readily extends to (C), as divergence follows from stationarity of the residuals, which is clearly also given if the series are not $I(1)$.

[14]Appendix E demonstrates that all qualitative findings remain intact when generating (B) and (C) with an unrestricted $\boldsymbol{\Omega}$ as in (A). Moreover, we show that non-diagonality of neither $\boldsymbol{\Pi}$ nor $\boldsymbol{\Gamma}$ affects the conclusions.

[15]For $t_\gamma^{\text{ADF}}$ we use response surface critical values. We also studied Phillips and Ouliaris (1990) and $\lambda_{\text{trace}}$ tests. Since these are very strongly correlated with $t_\gamma^{\text{ADF}}$ and $\lambda_{\max}$ resp. (Gregory *et al.*, 2004), adding these to $\tilde{\chi}^2_{\mathcal{I}}$ or $UR_{\psi_{\mathcal{I}}}$ barely affects the latters' performance.

[16]Furthermore, we ran all simulations at the 1% and 10% level. We also get similar results with a version of (C) with AR(1) error terms instead of white noise $\boldsymbol{u}_t$.

[17]Power results of the bootstrap tests were very similar. The largest power difference over all DGPs, tests and $T$ was 4.4 percentage points, and the mean absolute difference just over one point. See online Table E.7 for details.

[18]Online Table E.6 offers additional results for DGP(D).

[19]DGP(B) and Table E.4 show that the meta tests also provide insurance against nuisance parameters (short-run dynamics and $R^2$).

[20]We searched for 'cointegration' and 'cointegrated' on the Wiley webpage. Of the 34 hits, we excluded 5, e.g. an editorial for a special issue, Monte Carlo papers or those using data already considered by Gregory *et al.* (2004). The modified 2001-2010 data are available upon request. The raw 1994-2001 data are available at http://qed.econ.queensu.ca/jae/2004-v19.1/gregory-haug-lomuto/.

[21]For this purpose, we categorize the studies according to whether they use a residual- (i.e. those by Engle and Granger, 1987, or Phillips and Ouliaris, 1990) or system-based Johansen (1988) test. That is, we identify all Johansen tests with $\lambda_{\max}$ and all residual-based tests with $t_\gamma^{\text{ADF}}$. Given the highly positive correlation within classes of tests (Gregory *et al.*, 2004), this approximation is accurate. In 22 ($99 - 77$) cases of conflicting test results, the original studies do not report a cointegration test, being concerned with e.g. estimating cointegration vectors.

[22]Online Appendix F reports results for $\tilde{\chi}^2_{\mathcal{I}}(\lambda_{\max}, t_\gamma^{\text{ADF}})$; results for other (bootstrap) combination tests are available.

[23]That the preferred test is more rejective than $\tilde{\chi}^2_{\mathcal{I}}$ here does not contradict the favorable power properties of $\tilde{\chi}^2_{\mathcal{I}}$ found in Section 6, as $\tilde{\chi}^2_{\mathcal{I}}$ can, and should, of course only be shown to be powerful in a class of level-$\alpha$ tests. Whether the way researchers identify their 'preferred' test leads to a level-$\alpha$ test or suffers from data-mining is impossible to say without knowledge of the decision process.

# References


Banerjee, A., Dolado, J. J., and Mestre, R. (1998) Error-correction mechanism tests for cointegration in a single-equation framework. *Journal of Time Series Analysis* **19**(3), 267–283.

Boswijk, H. P. (1994) Testing for an unstable root in conditional and unconditional error correction models. *Journal of Econometrics* **63**, 37–60.

Clements, M. P. and Hendry, D. F. (1995) Forecasting in cointegrated systems. *Journal of Applied Econometrics* **10**(2), 127–146.

Cooley, T. F. and Ogaki, M. (1996) A time series analysis of real wages, consumption, and asset returns. *Journal of Applied Econometrics* **11**(2), 119–134.

Elliott, G., Jansson, M., and Pesavento, E. (2005) Optimal power for testing potential cointegrating vectors with known parameters for nonstationarity. *Journal of Business & Economic Statistics* **23**(1), 34–48.

Elliott, G. and Pesavento, E. (2009) Testing the null of no cointegration when covariates are known to have a unit root. *Econometric Theory* **25**, 1829–1850.

Engle, R. F. and Granger, C. W. (1987) Co-integration and error correction: Representation, estimation, and testing. *Econometrica* **55**(2), 251–76.

Fisher, R. (1932) *Statistical Methods for Research Workers*. London: Oliver and Boyd.

Gregory, A. W., Haug, A. A., and Lomuto, N. (2004) Mixed signals among tests for cointegration. *Journal of Applied Econometrics* **19**(1), 89–98.

Harvey, D. I., Leybourne, S. J., and Taylor, A. M. R. (2009) Unit root testing in practice: Dealing with uncertainty over the trend and initial condition. *Econometric Theory* **25**, 587–636.

Haug, A. A. (1996) Tests for cointegration: A Monte Carlo comparison. *Journal of Econometrics* **71**, 89–115.

Horowitz, J. L. (2001) The bootstrap. Heckman, J. J. and Leamer, E. E., eds., *Handbook of Econometrics*, vol. 5, chap. 52, 3159–3228, Amsterdam: Elsevier.

Johansen, S. (1988) Statistical analysis of cointegration vectors. *Journal of Economic Dynamics and Control* **12**(2–3), 231–254.

Johansen, S. (1995) *Likelihood-Based Inference in Cointegrated Vector Autoregressive Models*. Oxford University Press.

Kremers, J. J., Ericsson, N. R., and Dolado, J. J. (1992) The power of cointegration tests. *Oxford Bulletin of Economics and Statistics* **54**(3), 325–348.

Littell, R. C. and Folks, J. L. (1971) Asymptotic optimality of Fisher's method of combining independent tests. *Journal of the American Statistical Association* **66**(336), 802–806.

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

Martens, M., Kofman, P., and Vorst, T. C. F. (1998) A threshold error-correction model for intraday futures and index returns. *Journal of Applied Econometrics* **13**(3), 245–263.

Palm, F. C., Smeekes, S., and Urbain, J.-P. (2010) A sieve bootstrap test for cointegration in a conditional error correction model. *Econometric Theory* **26**, 647–681.

Paparoditis, E. and Politis, D. N. (2003) Residual-based block bootstrap for unit root testing. *Econometrica* **71**(3), 813–855.

Park, J. Y. (1990) Testing for unit roots and cointegration by variable addition. *Advances in Econornetrics* **8**, 107–133.

Pesavento, E. (2004) Analytical evaluation of the power of tests for the absence of cointegration. *Journal of Econometrics* **122**, 349–384.

Pesavento, E. (2007) Residuals-based tests for the null of no-cointegration: An analytical comparison. *Journal of Time Series Analysis* **28**(1), 111–137.

Phillips, P. C. B. and Ouliaris, S. (1990) Asymptotic properties of residual based tests for cointegration. *Econometrica* **58**(1), 165–193.

Swensen, A. R. (2006) Bootstrap algorithms for testing and determining the cointegration rank in VAR models. *Econometrica* **74**(6), 1699–1714.

White, H. (2000) A reality check for data snooping. *Econometrica* **68**(5), 1097–1126.