IMK
Institut für Makroökonomie
und Konjunkturforschung

WORKING PAPER

Das IMK ist ein Institut
der Hans-Böckler-Stiftung

No. 199 • August 2019 • Hans-Böckler-Stiftung

ON (BOOTSTRAPPED) COINTEGRATION TESTS IN PARTIAL SYSTEMS

Sven Schreiber[1]

ABSTRACT

As applied cointegration analysis faces the challenge that (a) potentially relevant variables are unobservable and (b) it is uncertain which covariates are relevant, partial systems are often used and potential (stationary) covariates are ignored. Recently it has been argued that a nominally significant cointegration outcome using the bootstrapped rank test (Cavaliere, Rahbek, and Taylor, 2012) in a bivariate setting might be due to test size distortions when a larger data-generating process (DGP) with covariates is assumed. This study reviews the issue systematically and generally finds noticeable but only mild size distortions, even when the specified DGP includes a large borderline-stationary root. The previously found drastic test size problems in an application of a long-run Phillips curve (inflation and unemployment in the euro area) appear to hinge on the particular construction of a time series for the output gap as a covariate. We conclude that the problems of the bootstrapped rank test are not severe and that it is still to be recommended for applied research.

[1] Macroeconomic Policy Institute Duesseldorf (IMK at Hans Boeckler Foundation), and Free University Berlin.
Email: mail@sven.schreiber.name.

On (bootstrapped) cointegration tests in partial systems*

by Sven Schreiber†

August 2019

Abstract

As applied cointegration analysis faces the challenge that (a) potentially relevant variables are unobservable and (b) it is uncertain which covariates are relevant, partial systems are often used and potential (stationary) covariates are ignored. Recently it has been argued that a nominally significant cointegration outcome using the bootstrapped rank test (Cavaliere, Rahbek, and Taylor, 2012) in a bivariate setting might be due to test size distortions when a larger data-generating process (DGP) with covariates is assumed. This study reviews the issue systematically and generally finds noticeable but only mild size distortions, even when the specified DGP includes a large borderline-stationary root. The previously found drastic test size problems in an application of a long-run Phillips curve (inflation and unemployment in the euro area) appear to hinge on the particular construction of a time series for the output gap as a covariate. We conclude that the problems of the bootstrapped rank test are not severe and that it is still to be recommended for applied research.

JEL codes: C32 (multiple time series), C15 (statistical simulation methods), E31 (inflation)

Keywords: bootstrap, cointegration rank test, empirical size

1 Introduction

The cointegration rank test conducted in a multivariate system ("Johansen procedure") is a widespread and popular tool for applied time series analysis. It has long been known that

*I am grateful for valuable feedback from Robert Taylor, Anders Rahbek, Giuseppe Cavaliere, and Jack Lucchetti. They are not responsible for any errors that may remain.

†Macroeconomic Policy Institute Duesseldorf (IMK at Hans Boeckler Foundation), and Free University Berlin. Email: mail@sven.schreiber.name.

asymptotic inference with that test suffers from substantial size distortions in small samples typical of macroeconomic datasets. Johansen himself developed a finite-sample Bartlett correction for the trace test statistic (Johansen, 2002), and later on bootstrap techniques were proposed (Cavaliere, Rahbek, and Taylor, 2012, 2015). This could be considered as the state of the art.

Recently, however, by conducting an extensive array of simulations Benati (2015) arrived at the interesting result that even the bootstrapped version of the rank test could still be subject to considerable size distortions.[1] In one of the many simulations in his paper he essentially analyzed the performance of the bootstrapped rank test in a partial system, i.e. in a situation where the VAR used for the test is lower-dimensional than the DGP, even when only stationary covariates are omitted, not variables in the cointegration relationships themselves. Let inflation be denoted as π_t and unemployment as u_t, while the short- and long-term interest rates s_t and l_t are transformed a priori to the stationary term spread $(l-s)_t$ together with the differenced short rate Δs_t and the output gap y_t: Then the analysis concerns $x_{2,t} = (\pi_t, u_t)'$ with $N = 2$ versus $x_{5,t} = (\pi_t, u_t, l_t - s_t, \Delta s_t, y_t)'$ with $N = 5$. For the bivariate system he reports in his Table 2 a p-value of 0.049 for the bootstrapped test of a cointegrating rank $r = 0$ versus $r = 1$. This finding would usually suggest to reject non-cointegration of euro-area inflation and unemployment at the 5% level of significance. By simulation under the null hypothesis he then found a considerable size distortion of the bootstrapped test based on $x_{2,t}$ when the DGP was assumed to contain $x_{5,t}$ and dismissed the nominal findings of cointegration as a "statistical fluke".

Because the reliability of the cointegration test is crucial for many applied research areas, simulations using the actual data are also supplemented here with some simulations of artificial data.[2] Our main finding is that generally the bootstrapped rank test does not

[1] Benati's paper was not meant as an econometrics methods study but investigated the existence of long-run Phillips curve relationships in various economies (synthetical euro area, UK, USA, Canada, and Australia). In this context the term "long-run Phillips curve" refers to a connection between π, the growth rate of the price level (not wage inflation), and u, the level of the unemployment rate; see section 4.2 for plots of the euro area data. We focus here on the results for the euro area and follow the choice of Benati's synthetical sample that actually predates the introduction of the euro (quarterly data 1970-1998).

[2] The original application also considered cointegration ranks $r > 1$ including interest rate levels, and checked CPI inflation as a variant. The datasets are not strictly identical, but we obtain qualitatively the same results, see the appendix (A). For the bootstrap procedures we use the johansensmall.gfn function package (version $\geq$2.6) by Sven Schreiber and Andreas Noack Jensen for the open-source gretl program and freely available online from within gretl. Similar code for Matlab is for example available on De Angelis' homepage https://sites.google.com/view/luca-de-angelis/research.

over-reject to any alarming extent. This is true for example in simulations of the full 5-dimensional system with $x_{5,t}$ when the output gap y_t is measured as a a standard HP-filter cycle of real output. In the literature the bootstrapped rank test was found to have somewhat inflated test sizes when there is a large (stationary) root in the null model (Cavaliere, Rahbek, and Taylor, 2015), but this effect appears to be limited in the given partial system setting. We can qualitatively replicate the over-rejection of Benati (2015) only with a particular output gap measure that was formerly distributed with the ECB's area-wide model dataset (AWM), some properties of which we will discuss below. Hence overall we conclude that the problems of the bootstrapped rank test are not severe and that it is still to be recommended for applied research.

2 Theoretical considerations

Before turning to the simulations and replications, we briefly revisit the relevant theoretical background for cointegration in potentially partial systems.

First of all, note that the meaning of a "partial" system is different from the one used in Harbo, Johansen, Nielsen, and Rahbek (1998) and related works. There the considered systems are specified conditional on contemporaneous values of some of the $I(1)$ variables that are part of the cointegrating relations. In contrast, we use the term "partial" to refer to a model that completely disregards some stationary variables of the underlying full system. If the full system vector $x_{N,t}$ is N-dimensional and suitably ordered, we define a partial system as modelling the subvector $x_{M,t} = Fx_{N,t}$, where $F = [I_M : 0]$, $M < N$. Sometimes the process representing $x_{M,t}$ is called a subprocess or marginal process; this subprocess is assumed to contain all $I(1)$ components of $x_{N,t}$, such that for the remainder process it holds that $[0 : I_{N-M}]x_{N,t} \sim I(0)$.

The standard starting point that we will adopt is that the data of the full system $x_{N,t}$ are generated by a finite-order VAR. It is well known that in general the subprocess $x_{M,t}$ will then not possess a finite-order VAR representation but instead some VARMA form, which in turn entails an infinite-order VAR model. Before addressing any bootstrap techniques, an important question thus concerns the cointegration analysis of inifite-order VARs.

In this context, one important insight which can be attributed to Saikkonen and Luukko-

nen (1997) and Lütkepohl and Saikkonen (1999) is that the application of the standard Johansen rank test in $VAR(\infty)$ systems is asymptotically valid. Of course, for the asymptotics to work the chosen lag order must not grow too fast relative to the sample size, but this restriction is either irrelevant for practical applications in given samples or is easy to implement in an automated fashion.

Therefore, given that (1) the partial system $x_{M,t}$ has a $VAR(\infty)$ representation, that (2) the cointegration rank test using a finite lag order is asymptotically still justified, and that (3) the mentioned bootstrap approaches to the rank test are also known to be asymptotically justified, by implication the bootstrapped rank test could in principle be expected to be valid for partial systems, too.

However, approximating a $VAR(\infty)$ with a $VAR(p)$ obviously leaves some autocorrelation in the residuals "by construction". This is not the situation for which the *iid*-residual bootstrap is designed and hence it is not obvious whether it continues to be valid. In such a situation, the residual-based block bootstrap might be promising; see Jentsch, Politis, and Paparoditis (2015), who deal with the VECM coefficients for a given cointegration rank, however. Also, as mentioned by Kilian and Lütkepohl (2017, p.348), "no formal results ... about the validity of conducting inference about structural impulse responses in cointegrated VAR models based on the residual-based block bootstrap" exist. While our topic here is not structural impulse responses, a similar gap seems to apply to rank testing, especially in the $VAR(\infty)$ context of a partial system.

Until the statistical theory is completely settled, we must turn primarily to simulation studies. Also, it appears essential to obtain a good approximation to the $VAR(\infty)$ in the first place, such that the difference becomes negligible. Intuitively, if the residuals of a $VAR(p)$ fitted to the partial system are close to being white noise, then there is hope that a standard *iid*-residual bootstrap will work as usual. Building on this insight, we will therefore choose the VAR lag order for the partial systems endogenously based on diagnostic autocorrelation testing as part of the simulation algorithm.

3 Bootstrap test specifications

Throughout this note we focus on the popular case of an unrestricted constant, which was formally justified in Cavaliere, Rahbek, and Taylor (2015). For lag length selection in the test VARs we deliberately choose not to use information criteria. The reason is that the non-autocorrelation of residuals is essential for the validity of the standard *iid*-residual based bootstrap, and some of the lag order suggestions by information criteria led to substantial remaining residual autocorrelation. Thus we specify lag orders based on passing a diagnostic autocorrelation test instead.

We focus on the case where the permanent effects on inflation of many shocks are unrestricted (allowed but not forced to be permanent) because it leaves the reduced-form coefficients of the VAR unchanged, allowing the standard application of the Johansen rank test.

The original simulation study used a five-dimensional DGP including inflation and unemployment that imposed absence of cointegration, and then applied the bootstrapped rank test of the null hypothesis $r = 0$ vs. $r \geq 1$ to the bivariate sub-system of simulated inflation and unemployment (in levels) in each simulation draw. Table 3 in Benati (2015) shows that the bootstrap procedure rejected the null hypothesis of no cointegration at a nominal 5% significance in 18.3% of the simulation draws. Thus he concluded that the bootstrap test grossly exceeded its nominal significance level, and that therefore the original test rejection with a p-value of just under 5% might be "a fluke".

The original study's suggested simulation design is absolutely reasonable. However, this test approach is not the only possible one, at least two different test variants come to mind when further variables are suspected to be relevant for the system dynamics. To systematically address these issues, we enumerate the following three possibilities of cointegration testing with stationary co-variates in small samples:

1. (Bivariate, Benati's method) The null model is given by an unrestricted autoregression for the vector $x'_{0,t} = (\Delta u_t, \Delta\pi_t, y_t, \Delta s_t, l_t - s_t)$, where y_t is the output gap, and $l_t - s_t$ is the term spread between longer-term and short-term interest rates. To ensure a common lag length in levels, the $K-$th lag coefficients for the differences of

unemployment and inflation are set to zero for the simulation DGP:

$$x_{0,t} = c + \sum_{i=1}^{K-1} A_i x_{0,t-i} + (0_{5,2}|\tilde{A}_K) x_{0,t-K} + \varepsilon_t,$$

where $\tilde{A}_K$ is an unrestricted 5×3 matrix for the K-th coefficients of the three stationary co-variates. Use this model to generate pseudo data, then run the Cavaliere, Rahbek, and Taylor (2015) bootstrapped cointegration test with an unrestricted constant on each simulated draw of the bivariate data $x_{2,t}^{*'} = (u_t^*, \pi_t^*)$ with a lag order K.[3]

2. (Swensen, unmodelled covariates method) Another bootstrap possibility in the presence of stationary covariates is given by Swensen (2011). The null model is again set up and simulated as in 1, and the bootstrap test is also applied to the bivariate vector $x_{2,t}^{*'} = (u_t^*, \pi_t^*)$. However, the test system is augmented with lags of the co-variates $x_{3,t}^{*'} = (y_t^*, \Delta s_t^*, (l_t - s_t)^*)$, i.e. $x_{3,t-1}^* ... x_{3,t-K}^*$ are added as unrestricted regressors.[4]

3. (Full system method) If the researcher suspects that there are some important covariates which are known to be $I(0)$, it seems natural to simply include them in the test system. Thus the null model and the bootstrap framework is again given as in method 1, but here the vector to be tested is $x_{5,t}^{*'} = (u_t^*, \pi_t^*, y_t^*, \Delta s_t^*, (l_t - s_t)^*)$, and since the co-variates add three stationary directions to the system already under the null, the relevant hypothesis to test cointegration between unemployment and inflation is $r = 3$ vs. $r = 4$ (again with K lags).

4 Simulation results

In order to have full control and to avoid any unknown properties of actual data we start with the following artificial setup, where the role of unemployment and inflation is taken by

[3] It is not obvious from Benati's description how exactly he handles the lag structure in his simulation, i.e. whether or not he chooses a different lag length for the bivariate subsystem. We determine the lag length in each rank test based on autocorrelation diagnostics.

[4] We do not include contemporaneous values of the covariates as this would obviously violate the necessary assumption of uncorrelatedness. These pseudo covariates are re-generated in each simulation run, but are then held fixed for the inner bootstrap. This corresponds to the test variant described in remark 6 in Swensen (2011). His remark 3 also applies in our implementation, as we use the restricted non-cointegrated model in the bootstrap algorithm.

v_t and w_t.

4.1 Size simulations with artificial data

Consider the vector $x_3 = (v, w, z)'$ where the first two components (v_t, w_t) are $I(1)$ while the last one (z_t) is a stationary co-variate. Due to the presence of z_t the formal cointegration rank (dimension of the stationary directions) of the full system is one, even though the $I(1)$ variables are not cointegrated. The VECM representation is given by $\Delta x_{3,t} = \alpha\beta' x_{3,t-1} + \Gamma_1 \Delta x_{3,t-1} + c + \varepsilon_t$ with a diagonal covariance matrix and the trivial cointegration vector $\beta = (0, 0, 1)'$. The loading coefficients are $\alpha = (0.1, 0.3, a_z)'$, the unrestricted constant term is arbitrarily[5] set to $c = (0.9, -0.5, 0.3)'$ and the short-run dynamics are specified as:

$$\Gamma_1 = \begin{bmatrix} 0.4 & 0.3 & 0.1 \\ 0 & 0.5 & 0.1 \\ 0 & 0 & 0 \end{bmatrix}.$$

The covariate here is specified as an exogenous AR(1) process. Because of the insight from Cavaliere, Rahbek, and Taylor (2015) that a large stationary roots in the system can affect the empirical size of the boostrapped rank test, we analyze the cases $a_z = -0.5$ (small root) and $a_z = -0.08$ (large root). As usual, the corresponding levels form VAR with two unit roots is $x_{3,t} = B_1 x_{3,t-1} + B_2 x_{3,t-2} + \varepsilon_t$, where $B_1 = \alpha\beta' + I_3 + \Gamma_1$ and $B_2 = -\Gamma_1$. With $a_z = -0.5$ the roots of the system are: 1, 1, 0.5, 0.5, 0.4, 0, while with $a_z = -0.08$ they are: 1, 1, 0.92, 0.5, 0.4, 0. In the latter case obviously the largest stationary root is quite close to the unit circle and implies considerable persistence.

Running the test size simulations with the bootstrapped test variants described in Section 3, and using these two DGP variants, we obtain the results in table 1. First of all, despite the small sample length of $T = 100$ the test size distortions are relatively mild. In the full-system approach we even do not observe any impact of the larger stationary root on the rejection frequency. In the bivariate partial-system setup (first row in the table) there is an increase from an effective size of 7.1% to a size of 8.3% in the presence of additional high persistence, i.e. by roughly one percentage point.

[5] Since the rank test with an unrestricted constant term is not similar and depends on the presence of the drift term, it cannot be omitted.

Table 1: Bootstrapped test size simulation, artificial DGPs

(simulated rejection frequencies under H0, resampling as-if-iid)	small root (0.5)	large root (0.92)
Bivariate, $r_0 = 0$	0.070	0.083
Swensen 2 + 1 covar., $r_0 = 0$	0.071	0.065
Full 3-dim, $r_0 = 1$	0.052	0.050

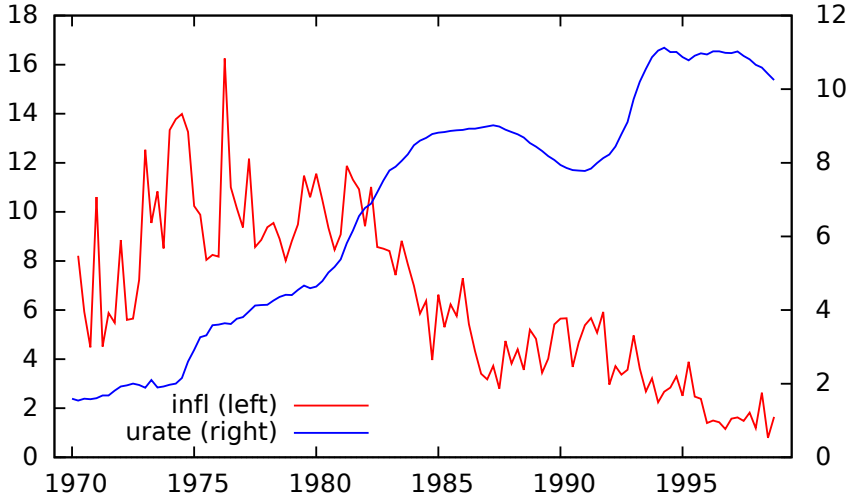
Notes: Nominal 0.05 significance level; 5000 replications; sample size $T = 100$.

Nevertheless, while these results are far from the previously reported distortions with apparent test sizes $> 15\%$ (at nominal 5%), given a borderline rejecting test result in actual data (for a chosen nominal significance level) it may of course make a difference for the decision whether the effective level of the test is α or 1.5α.

4.2 Simulated empirical size

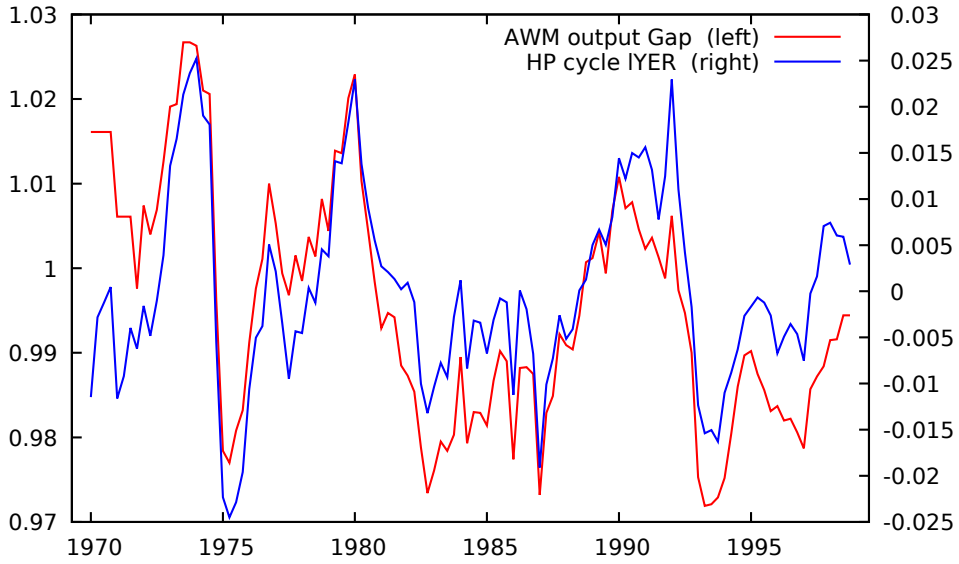
We now turn to the actual data analysis. The underlying system in these subsections 4.2 through 4.4 is a 5-dimensional VAR using the cycle component of a standard Hodrick-Prescott (HP) filter applied to real GDP as the relevant measure of the output gap y_t (see Figure 1). The two $I(1)$ series are reported in Figure 2, and the interest rate data as further stationary covariates in Figure 3.

Figure 2: Inflation and unemployment rates



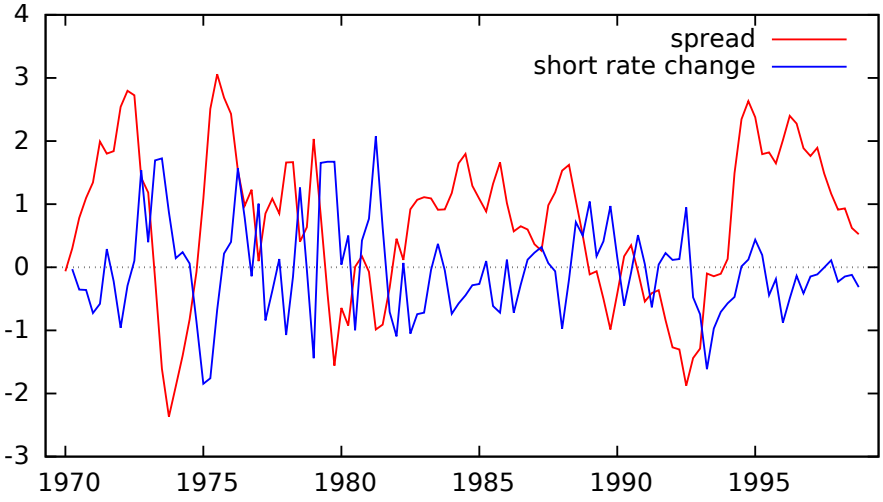
Notes: Data from the ECB's AWM, $400 \times \Delta\log(YED)$ and $100 \times URX$.

Figure 1: Output gaps



Notes: HP cycle is the result of a standard Hodrick-Prescott filter on log real GDP. (AWM is the output gap series from an earlier vintage of the ECB's area-wide model database, displayed for comparison. See also appendix A.)

Figure 3: Interest rates



Notes: Data from the ECB's AWM, $LTN - STN$ and ΔSTN.

We simulate the effective size (rejection probability under the null) of the bootstrapped cointegration test in the three different test strategies. Following Benati's approach we take the parameters of a non-cointegrated 5-dimensional VAR fitted to the data as the posited

Table 2: Test size simulations

(simulated rejection frequencies under H0)	as-if-iid	wild
Bivariate, $r_0 = 0$	0.069	0.083
Swensen 2 + 3 covariates, $r_0 = 0$	0.079	0.077
Full 5-dim, $r_0 = 3$	0.033	0.040

Notes: Simulation of the size of the bootstrapped rank test. Nominal 5%; 2000 simulation replications; the bootstrap test in each simulation draw uses 1000 replications. The time series length is $T = 109$.

DGP; to this end the two $I(1)$ variables are differenced and the stationary variables are left as is. We use 4 lags to obtain the parameters under the null, as this satisfies both the AC and ARCH residual tests.[6] For fitting the model to the simulated data in each draw we do not impose the original lag length but the algorithm chooses the lag order endogenously based on diagnostic residual testing. As explained above, this is important to obtain empirical residuals as close to white noise as possible.

Table 2 reports the size simulation results. For the rightmost column "wild", the rank test is based on a wild bootstrap scheme from the cited literature to account for potential heteroskedasticity. The takeaway from that simulation is that again there are only mild size distortions. The empirical sizes of the bivariate partial-system test and of Swensen's approach are roughly equal, and the full-system approach is mildly conservative which implies that its size is only about half of the sizes of the other aproaches (for a nominal 0.05 level).

4.3 Test results

Although the main motivation for this study is the behavior of the bootstrap rank test in partial systems in general, it is also interesting to replicate the test outcome from the original application. Given that we do not have the strictly identical dataset and vintages we do not expect identical results anyway, but the primary difference concerns the lag length specification: Because of the importance of non-correlated residuals we base our lag choice on diagnostic tests instead of information criteria.

[6] Having approximately white noise innovations is preferable because we use resampling for the simulation. If we drew the simulation innovations from a parametric model instead the lag length would of course be less important. In any case there are no qualitative differences whether one bootstrap variant or the other is used.

Table 3: Bootstrapped cointegration rank tests (inflation / unemployment)

(bootstrapped p-values)	iid	wild
Bivariate	0.011	0.027
Swensen 2 + 3 covar., $r_0 = 0$	0.182	0.213
Full 5-dim, $r_0 = 3$	0.159	0.185

Notes: 4999 replications; lags are chosen based on diagnostic tests: bivariate – 7 lags, Swensen's approach – 7 lags, full system – 4 lags. The respective sample size T is 113 minus the lag order.

Our test results on the actual data are reported in Table 3; the bivariate results in the first row are qualitatively similar to Benati (2015), namely that the null of no cointegration is nominally rejected at the 5% level but not at the 1% level.[7] The p-values are actually a little lower than in the original study, such that even after considering the noticeable size distortions from Table 2 the result would seem significant at the 5% level. However, the test conclusion is still not very clearcut.

While the residuals are free from autocorrelation in the bivariate specification with seven lags, there are always remaining ARCH effects, so the wild bootstrap variant (right column) may be preferred for the bivariate case. The other approaches were not considered in the original application.

Swensen's approach, where the bivariate system is augmented with the stationary covariates, is also subject to ARCH-type residuals, again suggesting the use of the wild bootstrap. Here the bootstrapped p-value is far above conventional critical levels (0.213), suggesting non-rejection of no cointegration. Finally, the full-system setup with four lags is well behaved, so the *iid* bootstrap is the method of choice, but it shares with Swensen's setup the non-rejection result (p-value 0.159).

4.4 Power assessment

The test results in Table 3 represent a dilemma. Given that in Table 2 we found that the size distortions of the bootstrapped rank test variants are not dramatic, we do not prefer one approach in Table 3 over any other based on the size assessment – that is, if we share the prior belief about the relevant covariates; otherwise the bivariate test would be preferred.

[7] As a memo item note that the standard bivariate rank test without a bootstrap and using asymptotic critical values here has a p-value ten times lower at 0.001.

Table 4: Test power simulations

(simulated rejection frequencies)	iid	wild
Bivariate, $r_0 = 0$	0.810	0.798
Swensen 2 + 3 covariates, $r_0 = 0$	0.139	0.128
Full 5-dim, $r_0 = 3$	0.224	0.235

Notes: Simulation of the power of the bootstrapped rank test for the fixed alternative given by the cointegrated system (cointegration between unemployment and inflation plus the three stationary covariates) estimated from actual data. Nominal 5%; 2000 simulation replications; the bootstrap test in each simulation draw uses 1000 replications. The time series length is $T = 109$.

But obviously the test outcomes are very different, so a test decision is difficult.

Therefore we turn to an assessment of the empirical power of the three test approaches. To this end we run a similar simulation as before in Section 4.2, but using as the DGP a system under the alternative hypothesis, with cointegration: the parameters are taken from the estimated error correction system (VECM) of the actual data under an assumed rank of 4, including the cointegration coefficients β. Three of the four columns of β are trivial unit vectors picking the stationary covariates, which technically increases the cointegration rank. The only "actual" cointegration relationship is still the one between unemployment and inflation. Then we simulate artificial data many times with resampled innovation processes, and each time we run the bootstrapped cointegration rank test on the artificial data.

The results of that simulation exercise are reported in Table 4. There is a surprisingly large gap between the power of around 80% in the bivariate case and the power below 25% or even 15% in the full-system and Swensen approaches. This means that the latter two approaches would quite rarely result in rejection of the null hypothesis even if it were false. Against this background it appears that the bivariate setup is the most reliable, combining only mild size distortions with large power advantages. Given the present covariates, the most natural test conclusion would therefore seem to be that euro area unemployment and inflation are cointegrated at a significance level of 5%, but not at the 1% level.

5 Conclusions

The issue of how cointegration rank tests behave when they are applied in partial systems is important, because in practice (a) either potentially relevant variables are unobservable, or (b) it is fundamentally uncertain which covariates might be relevant. This study has partly confirmed the worrying insight that rejection results in partial systems may sometimes be misleading. However, the good news is that the amount of the size distortion appears far smaller than previously suggested in the literature.

The conjecture (inspired from Cavaliere, Rahbek, and Taylor 2015) that the size distortion may be due to additional large (stationary) roots in the DGP in the background was only partly reflected in simulations with artificial data, and the effect did not appear large. For the original application of a euro-area long-run Phillips curve we were only able to replicate dramatic size distortions by simulations when the special AWM gap variable from Figure 1 was used as a covariate. (Various vintages of that series were formerly published as part of the area-wide model dataset of the ECB, see also the appendix). That time series possesses a mean in the second subsample which is lower by about 72% of the series' standard deviation; thus it may not really be stationary, which is unusual for such a gap concept. We also suspect that this output gap measure was constructed depending on the in-sample development of inflation, and that this causes the decline in the medium to long run. Hence it induces a larger root in the fitted model that was then used as the DGP in the simulations. Nevertheless, the quantitatively dramatic consequences of basing the simulations on this particular co-variate remains surprising.

In contrast, the test size distortions are very limited with a standard HP filter gap in the background, even though its univariate autoregressive root is also quite large (0.85). Therefore, the econometric evidence for cointegration in this sample between unemployment and inflation remains intact, unless one is completely convinced a priori that the true output gap was given by the AWM measure. We also showed that using full-system methods instead does not pay off, suffering from a severe lack of power.

Finally, it should be acknowledged that this study has addressed a very specific methodological aspect of Benati (2015), which also includes an impressive amount of other empirical and theoretical work. It is not the purpose of this note to question the broad conclusions

of his work, summarized as "uncertainty ... is ... substantial" (p. 27). We fully agree. Nevertheless, we regard it as important to clarify for applied economists that conducting cointegration tests in small samples with a bootstrap remains a justified practice and that its results cannot be easily discarded as "statistical flukes".

References

BENATI, L. (2015): "The long-run Phillips curve: A structural VAR investigation," *Journal of Monetary Economics*, 76, 15–28.

CAVALIERE, G., A. RAHBEK, AND A. M. R. TAYLOR (2012): "Bootstrap Determination of the Co-integration Rank in Vector Autoregressive Models," *Econometrica*, 80(4), 1721–1740.

CAVALIERE, G., A. RAHBEK, AND A. M. R. TAYLOR (2015): "Bootstrap Determination of the Co-integration Rank in VAR Models with Unrestricted Deterministic Components," *Journal of Time Series Analysis*, 36, 272–289.

HARBO, I., S. JOHANSEN, B. NIELSEN, AND A. RAHBEK (1998): "Asymptotic Inference on Cointegrating Rank in Partial Systems," *Journal of Business & Economic Statistics*, 16(4), 388–399.

JENTSCH, C., D. N. POLITIS, AND E. PAPARODITIS (2015): "Block Bootstrap Theory for Multivariate Integrated and Cointegrated Processes," *Journal of Time Series Analysis*, 36, 416–441.

JOHANSEN, S. (2002): "A Small Sample Correction for the Test of Cointegrating Rank in the Vector Autoregressive Model," *Econometrica*, 70(5), 1929–1961.

KILIAN, L., AND H. LÜTKEPOHL (2017): *Structural Vector Autoregressive Analysis*, Themes in Modern Econometrics. Cambridge University Press.

LÜTKEPOHL, H., AND P. SAIKKONEN (1999): "Order selection in testing for the cointegrating rank of a VAR process," in *Cointegration, Causality, and Forecasting – A Festschrift in Honour of Clive W.J. Granger*, ed. by R. F. Engle, and H. White, pp. 168–199. Oxford University Press.

SAIKKONEN, P., AND R. LUUKKONEN (1997): "Testing cointegration in infinite order vector autoregressive processes," *Journal of Econometrics*, 81, 93–129.

SWENSEN, A. R. (2011): "A bootstrap algorithm for testing cointegration rank in VAR models in the presence of stationary variables," *Journal of Econometrics*, 165, 152–162.

Table 5: Test size simulations under 5-dim DGP with YGA

(simulated rejection frequencies under H0)	resampling as-if-iid	wild
Bivariate, $r_0 = 0$	0.349	0.327
Swensen 2 + 3 covar., $r_0 = 0$	0.067	0.086
Full 5-dim, $r_0 = 3$	0.023	0.023

Notes: nominal level 0.05; 2000 replications, 7 lags in DGP, sample 1970Q2-1998Q4 (including initial values).

A Supplementary results with the AWM gap

The euro-area output gap measure in Benati (2015) is not a standard HP-filtered cycle but was based on a certain vintage "from the ECB's database" (quote from the online appendix to Benati, 2015). The precise calculation method of that series is unknown.

As a proxy we used the output gap series that we obtained from an earlier vintage of the ECB's area-wide model (AWM) database. In Figure 1 this proxy and the HP gap were compared. At business-cycle frequencies the two series are highly correlated, as should be expected. However, while the HP cycle measure fluctuates around a constant mean (by construction), the AWM gap is more persistent in the longer run, starting with a sequence of higher-than-average values and finishing the sample with many lower-than-average values. Its AR(1) root is 0.90, opposed to the slightly lower root of the HP cycle of 0.85. Given the limited effects of a large stationary root (see Section 4.1) we do not expect this property alone to have a large impact.

In the test size simulations analogous to Section 4.2, using this described AWM gap instead then requires 7 lags under the null to obtain innovations close to white noise. We observe in Table 5 that again the full-system approach is somewhat conservative, Swensen's approach is mildly oversized, but that now the bivariate partial-system test approach is dramatically oversized with an empirical size over 30% for a nominal 5%. This is even more drastic than Benati's original finding (based on a different lag length and probably slightly different data). Together with the actual test outcomes in Section 4.3 this represents a qualitatively successful replication of the original results.

It could be seen in Figure 1 that the initial values of the earlier AWM output gap are artificially extended and perhaps not very intuitive. As a robustness analysis we therefore

Figure 4: Shorter AWM output gap (YGA) range

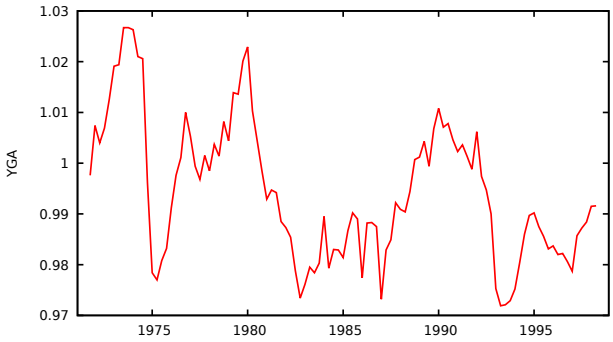


Table 6: Robustness: Test size simulations under 5-dim DGP with shorter YGA

(simulated rejection frequencies under H0)	resampling as-if-iid
Bivariate, $r_0 = 0$	0.230
Swensen 2 + 3 covar., $r_0 = 0$	0.089
Full 5-dim, $r_0 = 3$	0.035

Notes: nominal level 0.05; 2500 replications, 4 lags in DGP, sample 1971Q4-1998Q2 (including initial values).

took a more recent vintage of the AWM database where the output gap variable (YGA) only ranges from 1971Q4 to 1998Q2, see Figure 4. (Note that in more recent vintages of the AWM dataset the constructed YGA variable does not appear anymore.) In this shorter sample without the starting episode 4 lags are sufficient, and the corresponding simulation results are given in Table 6. It can be seen that the results correspond qualitatively to the ones in Table 5. The bivariate partial-system test results of the actual data are of course unaffected by any variations of the covariates in the background simulations and are therefore not repeated.

For completeness we also report the bootstrapped actual test results with the older AWM gap in Table 7. (Again, the bivariate test by definition does not depend on the output gap variable and was already shown in Table 3.) For the Swensen approach there are always remaining ARCH effects, thus the wild bootstrap results may be preferred, with a p-value of 0.011 suggesting rejection of no cointegration at the 5% significance level. Given the only mild size distortions of the Swensen approach this appears to be a valid result. The full-system approach here implies well-behaved residuals, so the preferred variant is the iid bootstrap, yielding a p-value of 0.366, not providing evidence in favor of cointegration.

Table 7: Test results with actual data (AWM gap)

(bootstrapped p-values)	iid	wild
Swensen 2 + 3 covar., $r_0 = 0$	0.007	0.011
Full 5-dim, $r_0 = 3$	0.366	0.335

Notes: 2000 replications; lag choices: Swensen – 5 lags, Full-system – 7 lags.

This may be accurate or could also be due to the lack of power demonstrated before.

Impressum

Publisher: Hans-Böckler-Stiftung, Hans-Böckler-Str. 39, 40476 Düsseldorf, Germany
Phone: +49-211-7778-331, IMK@boeckler.de, http://www.imk-boeckler.de

IMK Working Paper is an online publication series available at:
http://www.boeckler.de/imk_5016.htm

ISSN: 1861-2199