Bootstrapping the Likelihood Ratio Cointegration Test in Error Correction Models with Unknown Lag Order

Christian Kascha[a], Carsten Trenkler[b,*]

[a] *Norges Bank, Research Department, Bankplassen 2, 0107 Oslo, Norway*
[b] *University of Mannheim, Department of Economics, Chair of Empirical Economics, 68131 Mannheim, Germany*

Abstract

The finite-sample size and power properties of bootstrapped likelihood ratio systems cointegration tests are investigated via Monte Carlo simulations when the true lag order of the data generating process is unknown. Recursive bootstrap schemes are employed which differ in the way the lag order is chosen. The order is estimated by minimizing different information criteria and by combining the corresponding order estimates. In comparison to the standard asymptotic likelihood ratio test based on an estimated lag order, it is found that bootstrapping can lead to improvements in small samples even when the true lag order is unknown while the power loss is moderate.

Keywords: Cointegration Tests, Bootstrapping, Information Criteria

1. Introduction

In this paper, likelihood ratio cointegration tests with asymptotical and bootstrap critical values are compared in terms of their size and power in the case when the true lag order in the vector error correction model (VECM) is not known a priori. These tests have been compared only in situations in which the true lag order is known, as the theory on bootstrapping systems cointegration tests was developed only recently by Swensen (2006) and extended by Trenkler (2009) and Cavaliere, Rahbek, and Taylor (2010b,c). Monte Carlo experiments are conducted using three different data generating processes (DGPs) and recursive bootstrap procedures which differ in the way the lag order is chosen. The order is estimated by minimizing different information criteria and by combining the corresponding order estimates. In comparison to the asymptotic likelihood ratio test, it is found that bootstrapping can lead to improvements in small samples even in the unknown lag order case.

Most tests for cointegration are formulated in the well-known VECM framework for a n-dimensional time series $y_t = (y_{1,t}, \dots, y_{n,t})'$ observed for $t = 1, \dots, T$,

$$\Delta y_t \quad = \quad \mu_0 + \mu_1 t + \Pi y_{t-1} + \sum_{j=1}^{k} \Gamma_j \Delta y_{t-j} + \epsilon_t, \quad t = k+2, \dots, T, \tag{1}$$

where μ_0 and μ_1 are $(n \times 1)$ parameter vectors, Π and $\Gamma_1, \dots, \Gamma_k$ are $(n \times n)$ parameter matrices. The initial observations $y_1, \dots, y_{k+1}$ are taken as given. To assure that y_t is an $I(1)$ process we make the following assumptions. The roots of $\det((1-z)I_n - \Pi z - \Gamma_1(1-z)z - \cdots \Gamma_k(1-z)z^{k-1})$, where $\Pi = \alpha\beta'$, are either outside the unit circle or equal to one. Moreover, α and β are $(n \times r)$ matrices with full column rank r and the matrix $\alpha'_\perp \Gamma \beta_\perp$ has full rank, where $\Gamma = I_n - \sum_{j=1}^{k} \Gamma_j$ and where $\alpha_\perp$ and $\beta_\perp$ are the orthogonal complements to α and β.

[*] Corresponding address: University of Mannheim, Department of Economics, Chair of Empirical Economics, L7, 3-5, 68131 Mannheim, Germany, Tel.: +49 621 181 1852, Fax: +49 621 181 1931.

Email addresses: christian.kascha@norges-bank.no (Christian Kascha), trenkler@uni-mannheim.de (Carsten Trenkler)

For simplicity, it is assumed that $\epsilon_t \sim$ i.i.d.$(0, \Sigma)$ and $E(\epsilon_t^4) < \infty$. When the true VECM lag order k is used, the asymptotic validity of the asymptotic likelihood ratio test procedure and of the recursive *i.i.d.* bootstrap considered later on can be shown under weaker assumptions (see e.g. Cavaliere et al., 2010b). Note that $p = k + 1$ is the order of the underlying vector autoregressive (VAR) process for y_t. When the matrix Π has rank $r > 0$ the series are said to cointegrate. The most popular test for cointegration is the likelihood ratio (LR) test for the null hypothesis $\mathcal{H}_0 : r = r_0$ versus $\mathcal{H}_1 : r > r_0$ proposed by Johansen (1988, 1991).

The asymptotic distribution of the LR test was first derived by Johansen (1988) under the assumption that the true lag length is used. Even in this favourable case, many authors such as Toda (1995), Ho and Sorensen (1996) or Gonzalo and Pitarakis (1999) have shown that the size of the LR test in small samples can substantially differ from its nominal value when asymptotic critical values are used. The problem has been addressed in two different ways. The first approach tries to correct or modify the test statistic such that its finite sample distribution is closer to the one obtained from asymptotic theory. Examples are the corrections proposed by Reinsel and Ahn (1992); Reimers (1992) and, in particular, Johansen (2002). The second approach uses bootstrap methods to obtain critical values of the finite sample distribution of the test statistic (see e.g. Swensen, 2006; Ahlgren and Antell, 2008).

In practice, however, the lag order k is unknown and has to be estimated prior to testing for cointegration. This is usually done by applying information criteria with respect to unrestricted VAR models fitted to y_t. In addition, it might be that the true lag order in (1) is infinite because the data are generated by a vector autoregressive moving average (VARMA) process. Saikkonen and Luukkonen (1997) showed that the use of Johansen's LR test is justified in the VARMA case in that the test has the same limiting distribution as in the finite-order case under some restrictions on the process and the lag order. See also Lütkepohl and Saikkonen (1999) and Bauer and Wagner (2005). The estimation of the lag order aggravates the above-mentioned problems. Indeed, there are many simulation studies pointing to intolerable size distortions when the wrong lag order is chosen (see e.g. Boswijk and Franses, 1992; Cheung and Lai, 1993; Yap and Reinsel, 1995). Similar pessimistic results when the lag order is estimated have been obtained by Lütkepohl and Saikkonen (1999).

Swensen (2006) and Trenkler (2009) found encouraging results for some recursive bootstrap systems cointegration tests. In their recursive bootstraps, an i.i.d. sample from the residuals is drawn and the bootstrap data are generated recursively according to the chosen time series model. The situation of an unknown lag order has not been investigated in these papers. Only van Giersbergen (1996) examined whether a stationary bootstrap can help to reduce distortions due to lag order misspecification. However, the stationary bootstrap is of limited applicability here because it relies on critical auxiliary parameters which are difficult to estimate in practice. The results on the asymptotic LR test suggest, however, that it is crucial to compare bootstrap cointegration tests with the asymptotic LR test in a setup in which the lag order is unknown.

Therefore, this paper examines via Monte Carlo simulations how lag selection within different strategies affects the small sample properties of recursive bootstrap cointegration tests. Also, various ways to handle the additional uncertainty stemming from the lag order estimation are investigated following proposals in the literature on bootstrap unit root tests (e.g. Richard, 2009) and on bootstrapping within a VAR framework with unknown lag order (e.g. Kilian, 1998). Such a setup, however, presents some major theoretical difficulties as the lag order becomes random as well, see Leeb and Pötscher (2005) and Pötscher and Leeb (2009). Their work demonstrates that the usual inference ignoring that a model has been randomly selected, the so-called post-model-selection setup, becomes invalid. As the bootstrap aims at inferring the small sample distribution of the LR test statistic, this also means that the standard pointwise consistency proofs are not sufficient in a setup where the lag order is unknown. Thus, the simulation results in this paper have to be viewed in light of these findings. However, the fact that the bootstrap turns out to perform well in some circumstances indicates that it might be worth studying the approach analytically.

The rest of the paper is organized as follows. The test procedures are described in Section 2. Section 3 contains a description of the design of the Monte Carlo simulations and a discussion of the results. Section 4 concludes.

2. Test Procedures

In this section, tests for cointegration in the VECM framework in (1) are described. The focus is on the VECM with a restricted trend term and r cointegrating relations such that one can write $\mu_1 = \alpha\rho$ with α as above and ρ being a scalar. Model (1) can then be written as

$$\Delta y_t = \mu_0 + \alpha(\beta' y_{t-1} + \rho t) + \sum_{j=1}^{k} \Gamma_j \Delta y_{t-j} + \epsilon_t. \tag{2}$$

Given the lag order k, denote by R_{0t} and R_{1t} the residuals obtained from regressing Δy_t and $(y_{t-1}, t)'$ on $(1, \Delta y_{t-1}, \ldots, \Delta y_{t-k})'$, respectively. Define further $S_{ij} = T^{-1} \sum_{t=k+2}^{T} R_{it} R'_{jt}$ and denote by $\hat{\lambda}_1 > \ldots > \hat{\lambda}_n > \hat{\lambda}_{n+1} = 0$ the ordered eigenvalues of $|\lambda S_{11} - S_{10} S_{00}^{-1} S_{01}| = 0$. For simplicity, the dependence on the lag order is omitted for all these quantities but it is retained for the test statistics. Johansen's LR trace test statistic for $\mathcal{H}_0 : r = r_0$ versus $\mathcal{H}_1 : r > r_0$ is

$$LR_{r_0}(k) = -(T - k - 1) \sum_{i=r_0+1}^{n} \log(1 - \hat{\lambda}_i). \tag{3}$$

This is the most commonly used test statistic and its (non-standard) asymptotic distribution can be found in Johansen (1996). Here, asymptotical critical values computed by Doornik (1998) are used. As pointed out in the introduction, Johansen (2002) proposed a small sample correction following Bartlett (1937). Swensen (2006), however, showed that bootstrap approximations work better than the correction suggested by Johansen (2002) in a number of situations. Simulations undertaken for this paper have confirmed this. Therefore, this study does not report results on the Bartlett-corrected LR test. The use of the true lag order will be explicitly indicated by the notation $LR_{r_0}(k_{true})$.

In order to determine the cointegrating rank, the LR test is applied sequentially. To be precise, one starts by testing the null hypothesis $\mathcal{H}_0 : r = 0$. If it is rejected, then $\mathcal{H}_0 : r = 1$ is tested and so forth. The sequence stops if the null hypothesis is not rejected the first time and the corresponding rank under the null hypothesis is the estimate for r. For further details see Johansen (1996).

In practice, the lag order in (1) is of course unknown. A researcher might therefore use some information criterion, $IC(k)$, and choose $\hat{k}_{IC} = \operatorname{argmin}_k IC(k)$, where the minimization is over $k = 0, \ldots, k_T$ and k_T is a given upper bound on the possible lag orders. Paulsen (1984) shows that the standard order selection criteria are consistent for multivariate autoregressive processes with unit roots. From the results in Bauer and Wagner (2005) and Lütkepohl and Saikkonen (1999) it follows that the LR test with a lag order selected by an information criterion has asymptotically the same distribution as $LR_{r_0}(k_{true})$ under appropriate conditions. In particular, one has to assume that $k_T = o((T/\log T)^{1/2})$. Therefore, $k_{50} = 3$, $k_{100} = 4$, and $k_{200} = 6$ are used, which corresponds to taking the largest integers such that $k_T \leq (T/\log T)^{1/2}$. Since $k_T = o((T/\log T)^{1/2})$ only imposes an upper bound on the divergence rate of k_T regarding T, one could also apply larger values for k_T than the ones chosen. However, increasing k_T can lead to excessive size distortions when the sample size is small (e.g. $T = 50$). Corresponding simulation evidence is reported in Subsection 3.3.

The information criteria take the general form

$$IC(k) = \ln |\hat{\Sigma}(k)| + C_T \frac{kn^2}{N}, \tag{4}$$

where $N = T - k_T - 1$, $\hat{\Sigma}(k) = \sum_{t=k_T+2}^{T} \hat{\epsilon}_t \hat{\epsilon}'_t / N$ is an estimate of the error term covariance matrix Σ and the $\hat{\epsilon}_t$ are obtained by estimating an unrestricted VAR model of order $k+1$, on $y_{k_T+1-k}, \ldots, y_T$. This study considers Akaike's (1973) information criterion (AIC) with $C_T = 2$, the Hannan-Quinn (HQ) criterion (Hannan and Quinn, 1979) with $C_T = \ln \ln N$ and Schwarz's (1978) Bayesian information criterion (SC), $C_T = \ln N$. In addition, the authors employ the modified Akaike information criterion ($MAIC$) of Qu and Perron (2007) that imposes r_0 at the lag specification stage in order to obtain a better estimate of the Kullback-Leibler divergence in small samples. It is given by

$$MAIC(k, r_0) \quad = \quad \ln |\hat{\Sigma}(k, r_0)| + 2\frac{\widetilde{LR}_{r_0}(k) + kn^2}{N}, \tag{5}$$

where $\widetilde{LR}_{r_0}(k) = -N \sum_{i=r_0+1}^{n} \ln(1 - \hat{\lambda}_i)$ is just the usual LR test statistic but obtained by maximizing the likelihood of the parameters based on $y_{k_T+2}, \ldots, y_T$ in the VECM. Thus, the $MAIC$ is almost identical to the AIC but includes the extra term $\widetilde{LR}_{r_0}(k)$. Since the application of the $MAIC$ often leads to inferior outcomes relative to the other information criteria, no detailed results regarding $MAIC$ are presented. The $MAIC$ will only be considered as a part of strategies that combine various information criteria. Nevertheless, it should be noted that the use of $MAIC$ can avoid overrejections from which the cointegration tests suffer in some cases.

When the LR test in (3) is used with a lag order selected by one of the information criteria it is denoted by $LR_{r_0}(\hat{k}_{AIC})$, $LR_{r_0}(\hat{k}_{HQ})$ and so forth. The notation $LR_{r_0}(\hat{k}_{IC})$ is used when the tests are referred to in general.

The baseline recursive bootstrap procedure for a nominal level γ works as follows.

1. Compute $\hat{k}_{IC}$ and estimate model (2) under $\mathcal{H}_0 : r = r_0$ to obtain estimates $\hat{\mu}_0, \hat{\rho}, \hat{\alpha}, \hat{\beta}$ and $\hat{\Gamma}_i$, $i = 1, \ldots, \hat{k}_{IC}$, and the residuals $\hat{\epsilon}_{\hat{k}_{IC}+2}, \ldots, \hat{\epsilon}_T$.
2. Compute B bootstrap replications of y_t^*, $t = \hat{k}_{IC} + 2, \ldots, T$, recursively by

$$\Delta y_t^* \quad = \quad \hat{\mu}_0 + \hat{\alpha}(\hat{\beta}' y_{t-1}^* + \hat{\rho} t) + \sum_{j=1}^{\hat{k}_{IC}} \hat{\Gamma}_j \Delta y_{t-j} + \epsilon_t^*,$$

   where the ϵ_t^* are drawn with replacement from the residuals $\hat{\epsilon}_{\hat{k}_{IC}+2}, \ldots, \hat{\epsilon}_T$. The starting values of the recursion, $y_1^*, \ldots, y_{\hat{k}_{IC}+1}^*$ are set equal to $y_1, \ldots, y_{\hat{k}_{IC}+1}$.
3. For each replication $b = 1, \ldots, B$, given $\hat{k}_{IC}$, estimate model (2) under r_0 and compute the LR test as in (3). Denote the bootstrap statistics by $LR_{r_0}^*(\hat{k}_{IC})_b$.
4. Estimate the p-value of the test statistic as

$$p^*(LR_{r_0}(\hat{k}_{IC})) \quad = \quad \frac{1}{B} \sum_{b=1}^{B} \mathbf{1}(LR_{r_0}^*(\hat{k}_{IC})_b > LR_{r_0}(\hat{k}_{IC})), \tag{6}$$

   where $\mathbf{1}(\cdot)$ denotes the indicator function. If $p^*(LR_{r_0}(\hat{k}_{IC})) < \gamma$ reject $\mathcal{H}_0 : r = r_0$.

The bootstrap test versions with the corresponding criterion are denoted by $BOOT_{r_0}(\hat{k}_{AIC})$, $BOOT_{r_0}(\hat{k}_{HQ})$ et cetera. In order to estimate the cointegrating rank the bootstrap procedure is applied sequentially as the asymptotic LR test, i.e. steps 1 to 4 are repeated with $r_0 + 1$ if $\mathcal{H}_0 : r = r_0$ is not rejected in step 4.

Two remarks on the bootstrap procedure are in order. First, the baseline bootstrap procedure is an adoption of the recursive scheme of Cavaliere, Rahbek, and Taylor (2010a) suggested for the case of a known lag order. The latter bootstrap scheme assures that the bootstrap data are always asymptotically $I(1)$ with cointegrating rank r_0 even if $r_0 < r$, as may be the case in a sequential application of the bootstrap. In contrast, Swensen (2006) suggested estimating a VAR model without imposing the cointegrating rank null hypothesis to obtain residuals and estimates of μ_0 and Γ_i, $i = 1, \ldots, k$. Hence, a combination of estimates from two different models would be applied. However, this combination can cause inferior small-sample properties in the case of nonzero deterministic terms as pointed out by Trenkler (2009). Moreover, the bootstrap approach of Swensen (2006) only results in bootstrap data that are asymptotically $I(1)$ with cointegrating rank r_0 if $r = r_0$ or if a certain eigenvalue condition holds, see Swensen (2009, Assumption 2). For these reasons, the bootstrap scheme of Swensen (2006) has not been analyzed in detail.

Second, Swensen (2006, 2009) proved the asymptotic validity of the bootstrap in the case of a known lag order. There are two crucial requirements. First, the parameter estimators are consistent. Second, the appropriately scaled partial sums of bootstrap and sample error terms converge to the same Brownian motion. However, it is not trivial to analyze the asymptotic properties of the bootstrap cointegration test in the case of a pre-estimated lag length. In the following, we highlight some important problems that emerge in such a post-model-selection setup.

While consistency of the parameter estimators is retained when using a consistent (e.g. SC or HQ) or conservative model selection criterion (e.g. AIC) (see Pötscher, 1991, Lemmata 1 and 2), the derivation of appropriate limiting distributions in a post-model-selection setup is much more intricate. A consequence of Pötscher (1991, Lemma 1) is that the asymptotic distribution of the post-model-selection parameter estimators is still normal in the case of consistent model selection. However, this result only holds pointwise, but not uniformly with respect to the parameter space. More importantly, it is impossible to estimate the finite-sample distribution of the post-model-selection estimators uniformly consistent. This applies both to consistent and conservative model selection (compare Leeb and Pötscher, 2005, and the references therein).

Bootstrap approaches are also effected by the latter 'impossibility' result (see Leeb and Pötscher, 2005). While the bootstrap may be pointwise consistent if a consistent model selection procedure has been applied, it is not consistent in case of conservative model selection (compare e.g. Leeb and Pötscher, 2005; Kulperger and Ahmed, 1992; Knight, 1999). Although methods exist to restore consistency of the bootstrap in some setups (see e.g. Samworth, 2003), the consistency result again only holds pointwise, but not uniformly. Note, however, that a consistent estimation of the finite-sample distribution of the post-model-selection estimators does not need to be preferable in finite samples. Pötscher and Leeb (2009) point out that consistent estimation procedures may perform rather poorly in worst-case scenarios. In fact, the use of conservative model selection procedures and, thereby, of inconsistent estimators of distributions, can mitigate the problems that emerge due to the non-uniform convergence (compare also Pötscher, 1991, Section 4).

The foregoing implies that the bootstrap cointegration tests may perform arbitrarily badly for subsets of the parameter space, in particular for parameter subsets that make lag order selection a difficult problem. Hence, simulation results have to be carefully interpreted. While our results show that bootstrap cointegration tests with prior estimation of the lag order can perform poorly in some setups, they do not have particularly worse finite-sample properties than the test versions based on the true lag order. Although the simulation setups may have missed critical parameter setups, the identification of some well working bootstrap tests indicates that it may be possible and worthwhile to study their asymptotic properties analytically. Such an analytical approach, however, is left for future research. Note, in this respect, that results provided by Chang, Park, and Song (2006) and Palm, Smeekes, and Urbain (2010) for a sieve bootstrap setup could be a promising starting point for analyzing the asymptotic properties of an appropriate sieve bootstrap cointegration rank test.

The authors also considered the fast double bootstrap (FDB) of Davidson and MacKinnon (2007) as a potential refinement of the bootstrap procedure. The results for the FDB were, however, very similar to the results of the standard bootstrap and are therefore not discussed. Note that Ahlgren and Antell (2008) reported slight improvements when applying the FDB in case of a known lag order.

Several modifications of the baseline bootstrap procedure in terms of lag order determination are analyzed. Let k^D denote the lag length used in the model on which the LR test is based, let k^{BT} denote the lag length used in the model on which the bootstrap data generation is based and, finally, let k^b denote the lag length that is used for computing the test statistics in the bootstrap replications $b = 1, \ldots, B$. Richard (2009) suggested to implement the bootstrap disentangling these lag choices and, in particular, to determine k^{BT} by imposing the rank of the null hypothesis. To incorporate his proposals, several modifications of the baseline bootstrap regarding the triple (k^D, k^{BT}, k^b) have been considered. It turned out that only one modification, denoted by $BOOT_{r_0}(\hat{k}_{IC}, \hat{k}_{IC,r_0}, \hat{k}_{IC})$, worked satisfactorily. The procedures determines k^D via an information criterion in an unrestricted VAR while k^{BT} is determined via an information criterion imposing $\mathcal{H}_0$ and $k^b = k^D$ is set for for all b. For the determination of the lag length under $\mathcal{H}_0$, the information criteria are computed analogously. Only the standard information criteria AIC, HQ, and SC are applied in connection with the modifications of the baseline bootstrap. This approach will also be referred to as modified recursive bootstrap. We note that our results regarding this and other modified bootstrap procedures, confirm the findings of Richard (2009) in terms of the endogenization of the lag order choice.

Since it turned out that none of the information criteria is satisfactory in all simulation setups, several ways of combining them have been studied. First, the lag length is estimated as the minimum, maximum or average of orders suggested by the different information criteria AIC, HQ, SC, and $MAIC$. Using the minimum leads to quite oversized tests in some situations. By contrast, the maximum lag length leads to far too conservative tests and a considerable loss of power in many setups. Therefore, only results on the average are reported and the corresponding tests are denoted by $LR_{r_0}(\hat{k}_{AVE})$ and $BOOT_{r_0}(\hat{k}_{AVE})$.

An alternative approach of combining is to take the union or intersection of tests based on different information criteria. For the union, a null hypothesis is rejected when at least one of the tests rejects. This strategy often leads to excessive size distortions and is therefore not recommended. In the case of the tests' intersection, a null hypothesis is rejected when all tests reject. For finite-order VAR processes, taking the intersection of asymptotic and bootstrap tests often leads to results similar to those when using $LR_{r_0}(\hat{k}_{AVE})$ and $BOOT_{r_0}(\hat{k}_{AVE})$, respectively. In the case of VARMA processes, however, the approach is less beneficial than $LR_{r_0}(\hat{k}_{AVE})$ and $BOOT_{r_0}(\hat{k}_{AVE})$ since taking the intersection leads to tests which are much more size-distorted in some situations while too conservative in others. Given these outcomes, the results for the intersection and union of tests are not reported. Note, however, that a 5% level, as in the main simulations, has been applied before combining the individual tests. Adjusting the individual test level may lead to improved performance. This could be done e.g. from a multiple testing perspective (Romano and Wolf, 2005).

3. Monte Carlo Simulations

3.1. Simulation Design

Three different DGPs for sample sizes $T = 50,\ 100$, and 200 are simulated. These sample sizes are typical for macroeconomic applications. The computations are performed using programs written in GAUSS V8 for Windows. The RNDNS function with a fixed seed has been used to generate standard normally distributed random numbers. The number of replications is $R = 5000$. For determining the quantiles of the empirical bootstrap distributions, $B = 1000$ replications are computed.

The DGPs have mainly been chosen because they were used in the literature to exemplify the size distortions of the LR test. The first DGP was suggested by Toda (1994, 1995)

$$y_t = \begin{bmatrix} a_1 & 0 \\ 0 & 1 \end{bmatrix} y_{t-1} + \varepsilon_t, \;\; \varepsilon_t \sim \text{i.i.d. } N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \theta \\ \theta & 1 \end{bmatrix} \right), \tag{7}$$

where the initial value y_0 is set to zero. The parameter a_1 determines the cointegrating rank. If $|a_1| < 1$, $r = 1$ and θ describes the instantaneous correlation between the stationary and nonstationary components. In the simulations, $\theta = 0.8$ is used. In contrast, if $a_1 = 1$ the cointegrating rank is zero and θ is set to zero since, in this case, the test results do not depend on this parameter (Toda, 1994, 1995). The series' initial values are set to zero. Other bivariate VAR(1) processes of interest can be obtained from (7) by linear transformations which leave the LR tests invariant (compare Toda, 1994, 1995).

Since the process in (7) is a rather simple one, a more complex, data-based DGP from an empirical study of King, Plosser, Stock, and Watson (1991) (KPSW) is also used. King et al. (1991) analyze a small macroeconomic model for the U.S. which consists of the logarithms of per-capita private real GNP, per-capita real consumption, and per-capita gross private domestic fixed investment. A subset-VECM with one lag and two restricted cointegrating relationships is estimated on quarterly data for the period 1949:1-1988:4. Subset restrictions have been imposed using a *Top-Down* strategy based on the AIC as implemented in JMulTi (Lütkepohl and Krätzig, 2004, Chapter 3). The process is

$$\Delta y_t = \begin{bmatrix} -0.038 \\ -0.186 \\ 0.032 \end{bmatrix} + \begin{bmatrix} 0 & -0.026 \\ 0.217 & -0.150 \\ 0.126 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \end{bmatrix} y_{t-1} + \begin{bmatrix} 0 & 0 & 0.154 \\ 0 & 0.282 & 0.660 \\ 0.272 & 0.162 & 0 \end{bmatrix} \Delta y_{t-1} + \epsilon_t, \tag{8}$$

Table 1: Rejection Frequencies of Tests for Bivariate Toda DGP (7).

	Panel A:			Panel B:			Panel C:		
	$a_1 = 1$ $(r = 0)$, $\mathcal{H}_0 : r = 0$			$a_1 = 0.9$ $(r = 1)$, $\mathcal{H}_0 : r = 1$			$a_1 = 0.9$ $(r = 1)$, $\mathcal{H}_0 : r = 0$		
	$T = 50$	$T = 100$	$T = 200$	$T = 50$	$T = 100$	$T = 200$	$T = 50$	$T = 100$	$T = 200$
$BOOT_{r_0}(k_{true})$	0.0478	0.0502	0.0504	0.0148	0.0336	0.0618	0.1192	0.3888	0.9248
$BOOT_{r_0}(\hat{k}_{AIC})$	0.0834	0.0668	0.0566	0.0198	0.0370	0.0618	0.1666	0.4048	0.9172
$BOOT_{r_0}(\hat{k}_{HQ})$	0.0656	0.0496	0.0492	0.0168	0.0322	0.0602	0.1404	0.3876	0.9222
$BOOT_{r_0}(\hat{k}_{SC})$	0.0534	0.0486	0.0502	0.0154	0.0334	0.0624	0.1250	0.3876	0.9254
$BOOT_{r_0}(\hat{k}_{AVE})$	0.0482	0.0436	0.0470	0.0124	0.0320	0.0572	0.1098	0.3298	0.8786
$BOOT_{r_0}(\hat{k}_{AIC}, \hat{k}_{AIC,r_0}, \hat{k}_{AIC})$	0.0668	0.0558	0.0494	0.0194	0.0334	0.0606	0.1166	0.3178	0.8342
$BOOT_{r_0}(\hat{k}_{HQ}, \hat{k}_{HQ,r_0}, \hat{k}_{HQ})$	0.0580	0.0474	0.0506	0.0166	0.0328	0.0590	0.1214	0.3712	0.9192
$BOOT_{r_0}(\hat{k}_{SC}, \hat{k}_{SC,r_0}, \hat{k}_{SC})$	0.0518	0.0476	0.0490	0.0166	0.0332	0.0624	0.1200	0.3832	0.9236
$LR_{r_0}(k_{true})$	0.0584	0.0528	0.0532	0.0152	0.0344	0.0682	0.1342	0.4024	0.9294
$LR_{r_0}(\hat{k}_{AIC})$	0.1206	0.0800	0.0622	0.0222	0.0408	0.0698	0.2094	0.4304	0.9248
$LR_{r_0}(\hat{k}_{HQ})$	0.0848	0.0564	0.0546	0.0182	0.0352	0.0690	0.1634	0.4086	0.9288
$LR_{r_0}(\hat{k}_{SC})$	0.0658	0.0532	0.0534	0.0162	0.0344	0.0684	0.1426	0.4034	0.9294
$LR_{r_0}(\hat{k}_{AVE})$	0.0694	0.0522	0.0520	0.0146	0.0336	0.0658	0.1410	0.3574	0.8914

Note: The table shows rejection frequencies for replications of the Monte Carlo simulation. The number of simulations is 5000. The true cointegrating rank is r. The nominal significance level is 0.05. In the table, *LR* is Johansen's likelihood ratio test and *BOOT* denotes the bootstrap versions of the LR test. See section 2 for explanation.

where $\epsilon_t \sim$ i.i.d. $N(0, \Sigma)$ and

$$\Sigma = 10^{-4} \begin{bmatrix} 0.588 & 0.821 & 0.465 \\ & 4.870 & 1.688 \\ & & 1.376 \end{bmatrix}.$$

The series' initial values correspond to the empirical data. The same process was used by Trenkler (2009) in a related study. It turned out that both asymptotic and bootstrap cointegration tests displayed poor finite-sample properties for this process. Hence, the DGP (8) may be regarded as a demanding reference for the test procedures.

The third DGP is a mixed VARMA process which was also used by Yap and Reinsel (1995) and Lütkepohl and Saikkonen (1999). This process allows one to obtain results for infinite-order VAR processes. It is given by

$$\Delta y_t = P^{-1} \left(\begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix} - I_3 \right) P y_{t-1} + \epsilon_t - P_\theta \begin{bmatrix} \lambda_\theta & 0 & 0 \\ 0 & 0.297 & 0 \\ 0 & 0 & -0.202 \end{bmatrix} P_\theta^{-1} \epsilon_{t-1}, \tag{9}$$

where $\epsilon_t \sim$ i.i.d. $N(0, \Sigma)$ and

$$P = \begin{bmatrix} -0.29 & -0.47 & -0.57 \\ -0.01 & -0.85 & 1.00 \\ -0.75 & 1.39 & -0.55 \end{bmatrix}, \Sigma = \begin{bmatrix} 0.47 & 0.20 & 0.18 \\ & 0.32 & 0.27 \\ & & 0.30 \end{bmatrix}, P_\theta = \begin{bmatrix} -0.816 & -0.657 & -0.822 \\ -0.624 & -0.785 & 0.566 \\ -0.488 & 0.475 & 0.174 \end{bmatrix},$$

where we use $y_0 = \epsilon_0 = 0$. The values of the λ_i, $i = 1, 2, 3$, determine the cointegration properties of the series. That is, the number of λ_i with $|\lambda_i| < 1$ is the cointegrating rank of the system. The precise values are given in the tables later on. Note that the size of λ_θ affects how well the VARMA can be approximated by a VAR. A low value for λ_θ in modulus implies that all eigenvalues of the moving-average matrix are small and a finite-order VAR should be able to capture the true dynamics well since the other two eigenvalues are small as well. If λ_θ is large in modulus the moving-average part has one large eigenvalue and a VAR with a larger lag order is needed to approximate the DGP. In the simulations, λ_θ assumes values -0.5, 0, and 0.5.

3.2. Main Simulations

The results for the size and power of the different test procedures are given in Tables 1-4. Results are reported for a nominal size of 0.05.

For the simple bivariate Toda-DGP, Table 1 gives a comparison of the empirical size (Panel A and B) and power of the different test procedures (Panel C). The cointegrating rank is $r = 0$ in Panel A and $r = 1$ in Panels B and C. For the asymptotic and baseline bootstrap tests, one can see that using AIC, HQ, or SC to determine the lag order prior to employing either asymptotic tests or the bootstrap leads to higher empirical size values compared to applying the corresponding test with the true lag order if $T = 50$. Regarding the larger sample sizes, one only observes an upward size effect for AIC, while for HQ and SC the empirical sizes are rather similar to the ones obtained when $k_{true} = 0$ is used. This results from the fact that the correct lag order is suggested in about 98% or more of the replications by HQ and SC if $T = 100$ and $T = 200$. By contrast, the fraction of correct suggestions is only between 0.80 and 0.85 for AIC. Since $k_{true} = 0$, a too high lag order is chosen in the remaining replications. Thus, an overestimation of the lag length leads to larger size values and not to smaller ones for the standard criteria. Interestingly, the size-increasing effect of estimating the lag order is much stronger for the asymptotic tests. In comparison to these tests, application of the bootstrap reduces the size, but much less so when the corresponding asymptotic test with a particular lag selection criterion is only slightly oversized. Thus, the bootstrap correction does not mechanically reduce sizes but is sensitive to the size distortion of the corresponding asymptotic tests. Accordingly, choosing the bootstrap can be very useful to avoid or reduce excessive size distortions, which one observes for the standard criteria in a number of cases, compare Panel A and B of Table 1.

Regarding the power, the baseline bootstrap tests using AIC, HQ, or SC have a rather similar power as the bootstrap test with true order, compare panel C. More importantly, the power is only slightly smaller than those of the asymptotic tests with AIC, HQ, or SC. Hence, there is no relevant price to pay for bootstrapping in the case of an unknown lag order for the Toda-DGP (7).

The modified bootstrap procedure, $BOOT_{r_0}(\hat{k}_{IC}, \hat{k}_{IC,r_0}, \hat{k}_{IC})$, and the test based on averaging the lag order estimates, $BOOT_{r_0}(\hat{k}_{AVE})$, have sizes that are comparable in magnitude to $BOOT_{r_0}(\hat{k}_{IC})$, for $IC = AIC, HQ, SC$. The only notable difference is that the excessive size distortion of $BOOT_{r_0}(\hat{k}_{AIC})$ for the case $a_1 = 1$ is avoided, see Table 1, Panel A. The relative performance of $BOOT_{r_0}(\hat{k}_{AVE})$ regarding $LR_{r_0}(\hat{k}_{AVE})$ is similar to that of the baseline bootstrap tests with AIC, HQ, or SC.

From Panel C, one sees that the modified bootstrap tests and the procedures based on an average lag order can have clearly lower power than the standard asymptotic and bootstrap tests. This applies particularly when the sample size is small. Note that the results of the modified bootstraps depend much more on the chosen standard information criterion than is the case for the baseline bootstraps. If the test results for the information criteria differ strongly, then AIC produces the smallest and SC the largest power in connection with the modified bootstrap.

Table 2 provides results for the more complex KPSW-DGP (8) with a true cointegrating rank of two. Thus, Panel A ($\mathcal{H}_0 : r_0 = 2$) gives results on the tests' size while Panels B and C ($\mathcal{H}_0 : r_0 = 1$ and $\mathcal{H}_0 : r_0 = 0$) give results on the power of the tests. Since the VECM has one lag, underestimation of k can occur in this case. In fact, the information criteria, in particular HQ and SC, underestimate k quite often for $T = 50$. This may explain why one observes higher rejection frequencies for the baseline bootstrap tests with estimated lag order compared to $BOOT_{r_0}(k_{true})$ if $T = 50$ or $T = 100$. The low empirical size values seen in Panel A seem to be the result of distortions due to parameter estimation as even $LR_{r_0}(k_{true})$ is undersized. This may also have an negative effect on the tests' power. The modified bootstrap tests and $BOOT_{r_0}(\hat{k}_{AVE})$ perform similar to the bootstrap tests with standard information criteria. In the case of the KPSW-DGP, the overall effect of the bootstrap on the size of the tests is most often advantageous for $T = 100$ and $T = 200$ but small in any case.

Comparing the power of corresponding bootstrap and asymptotic tests, one sees that a price in terms of power loss has to be paid when bootstrapping, particularly in the case of $\mathcal{H}_0 : r = 0$ if the sample size is not large enough ($T = 50$, $T = 100$). However, the relative power loss of the baseline bootstrap tests is somewhat lower than in the case of using a true lag order, in particular for SC with $T = 50$. As a result, the introduction of lag order uncertainty tends to favour the baseline bootstrap in relative terms. Nevertheless, all procedures perform rather poorly in the current setup.

Lastly, two versions of the VARMA-DGP (9) are considered in Tables 3 and 4. First, Table 3 shows results for the size of the tests when $\lambda_1 = \lambda_2 = \lambda_3 = 1$ such that the true cointegrating rank is $r = 0$ and $\mathcal{H}_0 : r_0 = 0$ is tested. Since there is no true finite lag order, only results for tests that estimate the order are presented. The tests using asymptotic critical values conditional on an estimated lag length are quite oversized for different sample sizes and different values of λ_θ. It appears that the introduction of a moving-average component with small eigenvalues in modulus can already lead to severe size distortions. The tests are less size-distorted for $\lambda_\theta = -0.5$ and $\lambda_\theta = 0$ which imply mostly non-negative eigenvalues of the moving-average part. It seems unlikely that differences in the lag order estimates with respect to λ_θ play a crucial role in terms of the size distortion since no clear pattern could be detected.

The application of the bootstrap clearly reduces the empirical sizes. Nevertheless, the bootstrap tests can still be quite oversized even though the excessive size distortions are reduced compared with the asymptotic test counterparts. However, in the case of $BOOT_{r_0}(\hat{k}_{AIC}, \hat{k}_{AIC,r_0}, \hat{k}_{AIC})$ and, to a lesser extent, for $BOOT_{r_0}(\hat{k}_{HQ}, \hat{k}_{HQ,r_0}, \hat{k}_{HQ})$ and $BOOT_{r_0}(\hat{k}_{AVE})$, the size distortions may be acceptable.

Table 4 shows results regarding the size of the tests for $\lambda_1 = 1$, $\lambda_2 = 0.8$ and $\lambda_3 = 0.7$. Hence, the true cointegrating rank is two. All tests are very conservative for $T = 50$ and their empirical size increases with the sample size such that they are usually oversized for $T = 200$, especially in the case of $\lambda_\theta = 0.5$. There is a clear tendency for all information criteria to suggest larger models, the larger the sample size is. This may have caused the increasing size values.

The effect of applying the bootstrap procedure on the size depends on the value of λ_θ and the sample size.

Table 2: Rejection Frequencies of Tests for KPSW DGP (8) with $r = 2$.

	Panel A: $\mathcal{H}_0 : r = 2$			Panel B: $\mathcal{H}_0 : r = 1$			Panel C: $\mathcal{H}_0 : r = 0$		
	$T = 50$	$T = 100$	$T = 200$	$T = 50$	$T = 100$	$T = 200$	$T = 50$	$T = 100$	$T = 200$
$BOOT_{r_0}(k_{true})$	0.0036	0.0114	0.0298	0.0166	0.0756	0.3264	0.1192	0.4686	0.9914
$BOOT_{r_0}(\hat{k}_{AIC})$	0.0042	0.0122	0.0316	0.0230	0.0798	0.3286	0.1778	0.4788	0.9896
$BOOT_{r_0}(\hat{k}_{HQ})$	0.0046	0.0122	0.0286	0.0250	0.0756	0.3268	0.1890	0.4708	0.9914
$BOOT_{r_0}(\hat{k}_{SC})$	0.0040	0.0104	0.0302	0.0232	0.0728	0.3278	0.2250	0.4964	0.9918
$BOOT_{r_0}(\hat{k}_{AVE})$	0.0034	0.0114	0.0306	0.0172	0.0670	0.3056	0.1280	0.4036	0.9790
$BOOT_{r_0}(\hat{k}_{AIC}, \hat{k}_{AIC,r_0}, \hat{k}_{AIC})$	0.0066	0.0120	0.0304	0.0256	0.0676	0.3154	0.1124	0.3848	0.9512
$BOOT_{r_0}(\hat{k}_{HQ}, \hat{k}_{HQ,r_0}, \hat{k}_{HQ})$	0.0058	0.0112	0.0286	0.0258	0.0758	0.3276	0.1374	0.4632	0.9908
$BOOT_{r_0}(\hat{k}_{SC}, \hat{k}_{SC,r_0}, \hat{k}_{SC})$	0.0044	0.0116	0.0294	0.0258	0.0686	0.3302	0.1808	0.4756	0.9914
$LR_{r_0}(k_{true})$	0.0038	0.0104	0.0300	0.0382	0.0976	0.3598	0.2916	0.5990	0.9956
$LR_{r_0}(\hat{k}_{AIC})$	0.0084	0.0116	0.0300	0.0692	0.1040	0.3598	0.4102	0.6120	0.9946
$LR_{r_0}(\hat{k}_{HQ})$	0.0070	0.0104	0.0300	0.0542	0.0982	0.3598	0.3552	0.6038	0.9956
$LR_{r_0}(\hat{k}_{SC})$	0.0040	0.0110	0.0300	0.0394	0.0930	0.3596	0.3366	0.6146	0.9958
$LR_{r_0}(\hat{k}_{AVE})$	0.0048	0.0106	0.0292	0.0412	0.0886	0.3378	0.3012	0.5520	0.9896

Note: The table shows rejection frequencies for replications of the Monte Carlo simulation. The number of simulations is 5000. The true cointegrating rank is $r = 2$. The nominal significance level is 0.05. In the table, *LR* is Johansen's likelihood ratio test and *BOOT* denotes the bootstrap versions of the LR test. See section 2 for explanation.

Table 3: Rejection Frequencies for VARMA DGP (9) with $\lambda_1 = \lambda_2 = \lambda_3 = 1$ $(r = 0)$.

	Panel A:			Panel B:			Panel C:		
	$\lambda_\theta = -0.5,\ \mathcal{H}_0 : r = 0$			$\lambda_\theta = 0,\ \mathcal{H}_0 : r = 0$			$\lambda_\theta = 0.5,\ \mathcal{H}_0 : r = 0$		
	$T = 50$	$T = 100$	$T = 200$	$T = 50$	$T = 100$	$T = 200$	$T = 50$	$T = 100$	$T = 200$
$BOOT_{r_0}(\hat{k}_{AIC})$	0.1654	0.0930	0.0862	0.2052	0.1608	0.0786	0.4126	0.2470	0.1924
$BOOT_{r_0}(\hat{k}_{HQ})$	0.2404	0.1088	0.0908	0.2178	0.2474	0.1258	0.5238	0.4652	0.2272
$BOOT_{r_0}(\hat{k}_{SC})$	0.3118	0.2228	0.0886	0.2158	0.2920	0.2752	0.5710	0.7080	0.3302
$BOOT_{r_0}(\hat{k}_{AVE})$	0.1096	0.0694	0.0708	0.1302	0.1152	0.0732	0.2414	0.1862	0.1574
$BOOT_{r_0}(\hat{k}_{AIC}, \hat{k}_{AIC,r_0}, \hat{k}_{AIC})$	0.0698	0.0646	0.0650	0.0916	0.0714	0.0668	0.0920	0.1186	0.1256
$BOOT_{r_0}(\hat{k}_{HQ}, \hat{k}_{HQ,r_0}, \hat{k}_{HQ})$	0.0846	0.0746	0.0846	0.1372	0.1212	0.0900	0.1990	0.1970	0.2072
$BOOT_{r_0}(\hat{k}_{SC}, \hat{k}_{SC,r_0}, \hat{k}_{SC})$	0.1776	0.0960	0.0864	0.1992	0.2454	0.1684	0.4412	0.3704	0.2646
$LR_{r_0}(\hat{k}_{AIC})$	0.3614	0.1592	0.1184	0.3318	0.2012	0.0990	0.5476	0.3066	0.2242
$LR_{r_0}(\hat{k}_{HQ})$	0.3634	0.1636	0.1168	0.2816	0.2814	0.1402	0.6084	0.5004	0.2616
$LR_{r_0}(\hat{k}_{SC})$	0.3982	0.2706	0.1160	0.2642	0.3176	0.2882	0.6344	0.7306	0.3564
$LR_{r_0}(\hat{k}_{AVE})$	0.2308	0.1182	0.0960	0.2082	0.1508	0.0892	0.3626	0.2446	0.1860

Note: The table shows rejection frequencies for replications of the Monte Carlo simulation. The number of simulations is 5000. The true cointegrating rank is $r = 0$. The nominal significance level is 0.05. In the table, *LR* is Johansen's likelihood ratio test and *BOOT* denotes the bootstrap versions of the LR test. See section 2 for explanation.

However, the bootstrap approach generally corrects the size of the asymptotic tests towards the nominal size no matter whether the $LR_{r_0}(\hat{k}_{IC})$ tests are under- or oversized, respectively. Hence, for the current DGP, the size of the bootstrap tests may also systematically increase compared to the corresponding asymptotic tests. This is in contrast to the Toda- and KPSW-DGPs, where the empirical sizes of the bootstrap tests fall in general.

It turns out that $BOOT_{r_0}(\hat{k}_{AIC}, \hat{k}_{AIC,r_0}, \hat{k}_{AIC})$, $BOOT_{r_0}(\hat{k}_{AVE})$, and $LR_{r_0}(\hat{k}_{AVE})$ are preferable in terms of avoiding too many and severe underrejections while keeping the excessive size distortions in an acceptable range for the relevant setups. Due to the observed strong size distortions for the VARMA processes, the tests' small sample power is not evaluated.

3.3. Additional Simulations

Additional simulations have been undertaken to address various issues.

First, the distribution of the rank estimates was investigated for the different test procedures when used in a sequential manner for the Toda and KPSW-DGPs. For both processes, the outcomes for the distributions corresponded quite closely to the results for the tests' size and power. Also, the ratio of correct rank estimates was quite low for both DGPs and all tests. Detailed results can be obtained from the authors.

Second, the effect of k_T on the tests based on the standard information criteria AIC, HQ, and SC are discussed. Table 5 shows how the choice of k_{50} influences the LR and baseline bootstrap tests' rejection frequencies in case of the Toda-DGP with $r = 0$ $(a_1 = 1)$. The empirical size of $LR_{r_0}(\hat{k}_{AIC})$ and $LR_{r_0}(\hat{k}_{HQ})$ clearly increases for large values of k_{50}. Accordingly, the tests clearly overreject. One does not observe such a feature for $LR_{r_0}(\hat{k}_{SC})$. Bootstrapping in case of AIC and HQ reduces the size distortions but they are still visible. Even more extreme results can be found for the KPSW- and VARMA-DGPs. In those setups SC is also affected. The results indicate that practitioners should avoid using large maximum lag orders if only small samples are available.

Finally, the empirical sizes of the baseline bootstrap are investigated in the case of large sample sizes of $T = 1000$ when the lag order is pre-estimated by an information criterion for the three DGPs. For SC and HQ, the corresponding asymptotic and the bootstrap tests display the same rejection frequencies as when using the true order k in the case of the finite-order VAR processes. With respect to AIC one observes small deviations of the rejection frequencies from the cases with the true order, for both the asymptotic and bootstrap tests. Thus, the findings do not indicate that pre-estimating the lag-order leads to changing outcomes in large samples for the VAR processes considered, at least with respect to the consistent information criteria SC and HQ. In addition, the setup of a VARMA-DGP does not give rise to further concerns.

4. Conclusion

In this paper, the properties of bootstrap LR cointegration tests are investigated in terms of size and power in the vector error correction model when the lag order is not known a priori and has to be estimated by some information criterion. The paper contains Monte Carlo experiments for three different data generating processes that allow one to compare various ways of implementing a recursive bootstrap procedure with the corresponding tests based on asymptotical critical values.

The results of the simulations can be summarized as follows. When the lag order is not known a priori, bootstrapping remains advantageous in that it can bring empirical sizes closer to the nominal ones both when the asymptotic tests are over- or undersized. Hence, the introduction of lag order uncertainty does not impair the relative performance of the bootstrap. In fact, it even tends to favour the bootstrap in some cases.

However, none of the considered procedures works very well for both finite-order VARs and VARMA-DGPs. In particular, many test procedures suffer from excessive size distortions in case of the VARMA processes. In fact, there is a general trade-off between controlling the size for VARMA processes and a reasonable power in the case of finite-order VARs. Overall, two bootstrap procedures, $BOOT_{r_0}(\hat{k}_{AIC}, \hat{k}_{AIC,r_0}, \hat{k}_{AIC})$

Table 4: Rejection Frequencies of Tests for VARMA DGP (9) with $\lambda_1 = 1, \lambda_2 = 0.8, \lambda_3 = 0.7$ ($r = 2$).

	Panel A: $\lambda_\theta = -0.5,\ \mathcal{H}_0 : r = 2$			Panel B: $\lambda_\theta = 0,\ \mathcal{H}_0 : r = 2$			Panel C: $\lambda_\theta = 0.5,\ \mathcal{H}_0 : r = 2$		
	$T = 50$	$T = 100$	$T = 200$	$T = 50$	$T = 100$	$T = 200$	$T = 50$	$T = 100$	$T = 200$
$BOOT_{r_0}(\hat{k}_{AIC})$	0.0062	0.0364	0.0684	0.0082	0.0328	0.0490	0.0410	0.1602	0.1998
$BOOT_{r_0}(\hat{k}_{HQ})$	0.0056	0.0408	0.0872	0.0084	0.0356	0.0498	0.0552	0.3076	0.2616
$BOOT_{r_0}(\hat{k}_{SC})$	0.0052	0.0326	0.0868	0.0086	0.0354	0.0504	0.0574	0.4240	0.4226
$BOOT_{r_0}(\hat{k}_{AVE})$	0.0052	0.0332	0.0656	0.0062	0.0304	0.0466	0.0270	0.1344	0.2016
$BOOT_{r_0}(\hat{k}_{AIC}, \hat{k}_{AIC,r_0}, \hat{k}_{AIC})$	0.0090	0.0290	0.0590	0.0088	0.0292	0.0494	0.0256	0.0986	0.1560
$BOOT_{r_0}(\hat{k}_{HQ}, \hat{k}_{HQ,r_0}, \hat{k}_{HQ})$	0.0078	0.0386	0.0852	0.0074	0.0346	0.0490	0.0420	0.2206	0.2396
$BOOT_{r_0}(\hat{k}_{SC}, \hat{k}_{SC,r_0}, \hat{k}_{SC})$	0.0064	0.0328	0.0880	0.0104	0.0360	0.0510	0.0568	0.3768	0.3358
$LR_{r_0}(\hat{k}_{AIC})$	0.0080	0.0308	0.0700	0.0080	0.0308	0.0508	0.0364	0.1518	0.2022
$LR_{r_0}(\hat{k}_{HQ})$	0.0060	0.0344	0.0876	0.0070	0.0330	0.0522	0.0456	0.2992	0.2634
$LR_{r_0}(\hat{k}_{SC})$	0.0050	0.0294	0.0882	0.0068	0.0338	0.0524	0.0486	0.4132	0.4242
$LR_{r_0}(\hat{k}_{AVE})$	0.0044	0.0290	0.0664	0.0048	0.0278	0.0478	0.0208	0.1230	0.2060

Note: The table shows rejection frequencies for replications of the Monte Carlo simulation. The number of simulations is 5000. The true cointegrating rank is $r = 2$. The nominal significance level is 0.05. In the table, LR is Johansen's likelihood ratio test and $BOOT$ denotes the bootstrap versions of the LR test. See section 2 for explanation.

Table 5: Rejection Frequencies of Tests for Bivariate Toda DGP (7) with $r = 0$, $\mathcal{H}_0 : r = 0$ and $T = 50$.

	$k_{50} = 4$	$k_{50} = 8$	$k_{50} = 12$
$BOOT_{r_0}(\hat{k}_{AIC})$	0.0834	0.1236	0.1208
$BOOT_{r_0}(\hat{k}_{HQ})$	0.0656	0.0734	0.1134
$BOOT_{r_0}(\hat{k}_{SC})$	0.0534	0.0536	0.0560
$LR_{r_0}(\hat{k}_{AIC})$	0.1206	0.2150	0.5984
$LR_{r_0}(\hat{k}_{HQ})$	0.0848	0.0982	0.2278
$LR_{r_0}(\hat{k}_{SC})$	0.0658	0.0660	0.0700

Note: The table shows rejection frequencies for replications of the Monte Carlo simulation. The number of simulations is 5000. The true cointegrating rank is $r = 0$. The nominal significance level is 0.05. In the table, LR is Johansen's likelihood ratio test and $BOOT$ denotes the bootstrap versions of the LR test. See section 2 for explanation.

and $BOOT_{r_0}(\hat{k}_{AVE})$, are the best choices in balancing this trade-off. They avoid extreme upward size distortions without losing too much power. The choice between the two approaches boils down to deciding on either a better size control for VARMA processes ($BOOT_{r_0}(\hat{k}_{AIC}, \hat{k}_{AIC,r_0}, \hat{k}_{AIC})$) or higher power in the case of finite-order VARs ($BOOT_{r_0}(\hat{k}_{AVE})$). This recommendation also means that neither the baseline nor the modified bootstrap work better in general. The relative performance can sometimes strongly depend on what information criterion or combination of criteria is applied.

In light of the results of Leeb and Pötscher (2005) on inference in a post-model-selection setup, the simulation results indicate that it might be worth investigating the performance of the bootstrap in the unknown lag order setup analytically. Also, an interesting topic for future research would be to test for cointegration in a VARMA framework that allows for lag order uncertainty such that a finite-order VAR model is contained as a special case, for example, along the lines of Bauer and Wagner (2009). Such a procedure would ideally perform well both in the VAR as well as in the more general VARMA case.

Aknowledgements

We thank Anders Rygh Swensen, participants of the ESEM 2009 in Barcelona and of the Econometrics Seminar at the University of Kiel. We are grateful to an Associate Editor and two anonymous referees for very helpful comments. The views expressed in this paper are our own and do not necessarily reflect the views of Norges Bank.

Ahlgren, N., Antell, J., 2008. Bootstrap and fast double bootstrap tests of cointegration rank with financial time series. Computational Statistics & Data Analysis 52 (10), 4754–4767.

Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. In: Petrov, B. N., Csaki, F. (Eds.), 2nd International Symposium on Information Theory. Akademia Kiado, Budapest.

Bartlett, M. S., 1937. Properties of sufficiency and statistical tests. Proceedings of the Royal Society of London, Series A 160 (901), 268–282.

Bauer, D., Wagner, M., 2005. Autoregressive approximations of multiple frequency I(1) processes. Economics Series 174, Institute for Advanced Studies.

Bauer, D., Wagner, M., 2009. Using subspace algorithm cointegration analysis: Simulation performance and application to the term structure. Computational Statistics & Data Analysis 53 (6), 1954–1973.

Boswijk, P., Franses, P. H., 1992. Dynamic specification and cointegration. Oxford Bulletin of Economics and Statistics 54 (3), 369–81.

Cavaliere, G., Rahbek, A., Taylor, A., Feb. 2010a. Bootstrap sequential determination of the co-integration rank in var models. CREATES Research Papers 2010-07, School of Economics and Management, University of Aarhus.

Cavaliere, G., Rahbek, A., Taylor, A., 2010b. Co-integration rank testing under conditional heteroskedasticity. Econometric Theory, forthcoming.

Cavaliere, G., Rahbek, A., Taylor, A., 2010c. Testing for co-integration in vector autoregressions with non-stationary volatility. Journal of Econometrics, forthcoming.

Chang, Y., Park, J. Y., Song, K., 2006. Bootstrapping cointegrating regressions. Journal of Econometrics 133 (2), 703–739.

Cheung, Y.-W., Lai, K. S., 1993. Finite-sample sizes of Johansen's likelihood ration tests for conintegration. Oxford Bulletin of Economics and Statistics 55 (3), 313–28.

Davidson, R., MacKinnon, J. G., 2007. Improving the reliability of bootstrap tests with the fast double bootstrap. Computational Statistics and Data Analysis 51 (7), 3259–3281.

Doornik, J. A., 1998. Approximations to the asymptotic distributions of cointegration tests. Journal of Economic Surveys 12 (5), 573–93.

Gonzalo, J., Pitarakis, J. Y., 1999. Dimensionality effect in cointegration analysis. In: Engle, R. F., White, H. (Eds.), Cointegration, Causality and Forecasting: Festschrift in Honour of Clive Granger. Oxford University Press, Oxford, pp. 212–229.

Hannan, E. J., Quinn, B. G., 1979. The determination of the order of an autoregression. Journal of the Royal Statistical Society. Series B (Methodological) 41 (2), 190–195.

Ho, M. S., Sorensen, B. E., 1996. Finding cointegration rank in high dimensional systems using the Johansen test: An illustration using data based monte carlo simulations. The Review of Economics and Statistics 78 (4), 726–32.

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

Johansen, S., 1991. Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models. Econometrica 59 (6), 1551–1580.

Johansen, S., 1996. Likelihood-based Inference in Cointegrated Vector Autoregressive Models. Oxford University Press, Oxford.

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., 1998. Accounting for lag order uncertainty in autoregressions: the endogenous lag order bootstrap algorithm. Journal of Time Series Analysis 19 (5), 531–548.

King, R. G., Plosser, C. I., Stock, J. H., Watson, M. W., 1991. Stochastic trends and economic fluctuations. American Economic Review 81 (4), 819–840.

Knight, K., 1999. Epi-convergence in distribution and stochastic equi-semicontinuity. Working paper, Department of Statistics, University of Toronto.

Kulperger, R. J., Ahmed, S. E., 1992. A bootstrap theorem for a preliminary test estimator. Communications in Statistics - Theory and Methods 21 (7), 2071–2082.

Leeb, H., Pötscher, B. M., 2005. Model selection and inference: Facts and fictions. Econometric Theory 21 (1), 21–59.

Lütkepohl, H., Krätzig, M., 2004. Applied Time Series Econometrics. Cambridge University Press.

Lütkepohl, H., Saikkonen, P., 1999. Order selection in testing for the cointegrating rank of a var process. In: Engle, R. F., White, H. (Eds.), Cointegration, Causality, and Forecasting. A Festschrift in Honour of Clive W.J. Granger. Oxford University Press, Oxford.

Palm, F. C., Smeekes, S., Urbain, J.-P., 2010. A sieve bootstrap test for cointegration in a conditional error correction model. Econometric Theory, forthcoming.

Paulsen, J., 1984. Order determination of multivariate autoregressive time series with unit roots. Journal of Time Series Analysis 5 (2), 115–127.

Pötscher, B. M., 1991. Effects of model selection on inference. Econometric Theory 7 (2), 163–185.

Pötscher, B. M., Leeb, H., 2009. On the distribution of penalized maximum likelihood estimators: The LASSO, SCAD, and thresholding. Journal of Multivariate Analysis 100 (9), 2065–2082.

Qu, Z., Perron, P., 2007. A modified information criterion for cointegration tests based on a var approximation. Econometric Theory 23 (4), 638–685.

Reimers, H., Dec. 1992. Comparisons of tests for multivariate cointegration. Statistical Papers 33 (1), 335–359.

Reinsel, G. C., Ahn, S. K., 1992. Vector AR models with unit roots and reduced rank structure: Estimation, likelihood ratio test, and forecasting. Journal of Time Series Analysis 13 (4), 353–375.

Richard, P., 2009. Modified fast double sieve bootstraps for adf tests. Computational Statistics & Data Analysis 53 (12), 4490–4499.

Romano, J. P., Wolf, M., 2005. Stepwise multiple testing as formalized data snooping. Econometrica 73 (4), 1237–1282.

Saikkonen, P., Luukkonen, R., 1997. Testing cointegration in infinite order vector autoregressive processes. Journal of Econometrics 81 (1), 93–126.

Samworth, R., December 2003. A note on methods of restoring consistency to the bootstrap. Biometrika 90 (4), 985–990.

Schwarz, G., 1978. Estimating the dimension of a model. The Annals of Statistics 6 (2), 461–464.

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

Swensen, A. R., 2009. Corrigendum to "Bootstrap algorithms for testing and determining the cointegration rank in VAR models". Econometrica 77 (5), 1703–1704.

Toda, H. Y., 1994. Finite sample properties of likelihood ratio tests for cointegrating ranks when linear trends are present. The Review of Economics and Statistics 76 (1), 66–79.

Toda, H. Y., 1995. Finite sample performance of likelihood ratio tests for cointegrating ranks in vector autoregressions. Econometric Theory 11 (5), 1015–1032.

Trenkler, C., 2009. Bootstrapping systems cointegration tests with a prior adjustment for deterministic terms. Econometric Theory 25 (1), 243–269.

van Giersbergen, N. P. A., 1996. Bootstrapping the trace statistic in var models: Monte carlo results and applications. Oxford Bulletin of Economics and Statistics 58 (2), 391–408.

Yap, S. F., Reinsel, G. C., 1995. Estimation and testing for unit roots in a partially nonstationary vector autoregressive moving average model. Journal of the American Statistical Association 90 (429), 253–267.