entropy

MDPI

Article

Asymptotic Properties of Estimators for Seasonally Cointegrated State Space Models Obtained Using the CVA Subspace Method

Dietmar Bauer * and Rainer Buschmeier

Department of Business Administration and Economics, Bielefeld University, Universitaetsstrasse 25, 33615 Bielefeld, Germany; RBuschmeier@uni-bielefeld.de

* Correspondence: Dietmar.Bauer@uni-bielefeld.de

Abstract: This paper investigates the asymptotic properties of estimators obtained from the so called CVA (canonical variate analysis) subspace algorithm proposed by Larimore (1983) in the case when the data is generated using a minimal state space system containing unit roots at the seasonal frequencies such that the yearly difference is a stationary vector autoregressive moving average (VARMA) process. The empirically most important special cases of such data generating processes are the I(1) case as well as the case of seasonally integrated quarterly or monthly data. However, increasingly also datasets with a higher sampling rate such as hourly, daily or weekly observations are available, for example for electricity consumption. In these cases the vector error correction representation (VECM) of the vector autoregressive (VAR) model is not very helpful as it demands the parameterization of one matrix per seasonal unit root. Even for weekly series this amounts to 52 matrices using yearly periodicity, for hourly data this is prohibitive. For such processes estimation using quasi-maximum likelihood maximization is extremely hard since the Gaussian likelihood typically has many local maxima while the parameter space often is high-dimensional. Additionally estimating a large number of models to test hypotheses on the cointegrating rank at the various unit roots becomes practically impossible for weekly data, for example. This paper shows that in this setting CVA provides consistent estimators of the transfer function generating the data, making it a valuable initial estimator for subsequent quasi-likelihood maximization. Furthermore, the paper proposes new tests for the cointegrating rank at the seasonal frequencies, which are easy to compute and numerically robust, making the method suitable for automatic modeling. A simulation study demonstrates by example that for processes of moderate to large dimension the new tests may outperform traditional tests based on long VAR approximations in sample sizes typically found in quarterly macroeconomic data. Further simulations show that the unit root tests are robust with respect to different distributions for the innovations as well as with respect to GARCH-type conditional heteroskedasticity. Moreover, an application to Kaggle data on hourly electricity consumption by different American providers demonstrates the usefulness of the method for applications. Therefore the CVA algorithm provides a very useful initial guess for subsequent quasi maximum likelihood estimation and also delivers relevant information on the cointegrating ranks at the different unit root frequencies. It is thus a useful tool for example in (but not limited to) automatic modeling applications where a large number of time series involving a substantial number of variables need to be modelled in parallel.

Keywords: cointegration; subspace algorithms; VARMA models; seasonality

JEL Classification: C13; C32

Citation: Bauer, D.; Buschmeier, R. Asymptotic Properties of Estimators for Seasonally Cointegrated State Space Models Obtained Using the CVA Subspace Method. *Entropy* **2021**, *23*, 436. https://doi.org/10.3390/e23040436

Academic Editor: Christian H. Weiss

Received: 19 February 2021
Accepted: 31 March 2021
Published: 8 April 2021

1. Introduction

Many time series show seasonal patterns that, according to [1] for example, cannot be modeled appropriately using seasonal dummies because they exhibit a slowly trending behavior typical for unit root processes.

To model such processes in the vector autoregressive (VAR) framework, Ref. [2] (abbreviated as JS in the following) extend the error correction representation for seasonally integrated autoregressive processes pioneered by [3] to the multivariate case. This vector error correction formulation (VECM) models the yearly differences of a process observed S times per year. The model includes systems having unit roots at some or all of the possible locations $z_j = \exp(\frac{2\pi j}{S} i), j = 0, ..., S-1$ of seasonal unit roots. In JS all unit roots are assumed to be simple such that the process of yearly differences is stationary.

In this setting JS propose an estimator for the autoregressive polynomial subject to restrictions on its rank (the so-called cointegrating rank) at the unit roots z_j based on an iterative scheme focusing on a pair of complex-conjugated unit roots (or the unit roots $z_j = 1$ or $z_j = -1$ respectively) at a time. The main idea here is the reformulation of the model using the so called vector error correction representation. Beside estimators JS also derived likelihood ratio tests for the cointegrating rank at the various unit roots.

Refs. [4,5] propose simpler estimation schemes based on complex reduced rank regression (cRRR in the following). They also show that their numerically simpler algorithm leads to test statistics for the cointegrating rank that are asymptotically equivalent to the quasi maximum likelihood tests of JS. These schemes still typically alternate between cRRR problems corresponding to different unit roots until convergence, although a one step version estimating only once at each unit root exists. Ref. [6] provides updating equations for quasi maximum likelihood estimation in situations where constraints on the parameters prohibit focusing on one unit root at a time.

The leading case here is that of quarterly data ($S = 4$) where potential unit roots are located at ± 1 and $\pm i$, implying that the VECM representation contains four potentially rank restricted matrices. However, increasingly time series of much higher sampling frequency such as hourly, daily or weekly observations are available. In such cases it is unrealistic that all unit roots are present. If a unit root is not present, the corresponding matrix in the VECM is of full rank. Therefore in situations with only a few unit roots being present, the VECM requires a large number of parameters to be estimated. Also in cases with a long period length (such as for example hourly data with yearly cycles) usage of the VECM involves the estimation of all coefficient matrices for lags for at least one year.

In general, for processes of moderate to large dimension the VAR framework involves estimation of a large number of parameters which potentially can be avoided by using the more flexible vector autoregressive moving average (VARMA) or the—in a sense—equivalent state space framework. This setting has been used in empirical research for the modeling of electricity markets, see the survey [7] for a long list of contributions. In particular, ref. [8] use the model described below without formal verification of the asymptotic theory for the quasi maximum likelihood estimation.

Recently, ref. [9] show that in the setting of dynamic factor models, typically used for observation processes of high dimension, the common assumption that the factors are generated using a vector autoregression jointly with the assumption that the idiosyncratic component is white noise (or more generally generated using a VAR or VARMA model independent of the factors) leads to a VARMA process. Also a number of papers (see for example [10–12]) show that in their empirical application the usage of VARMA models instead of approximations using the VAR model leads to superior prediction performance. This, jointly with the fact that the linearization of dynamic stochastic general equilibrium models (DSGE) leads to state space models, see e.g., [13], has fuelled recent interest in VARMA—and thus state space—modeling in particular in macroeconomics, see for example [14].

In this respect, quasi maximum likelihood estimation is the most often used approach for inference. Due to the typically highly non-convex nature of the quasi likelihood function (using the Gaussian density) in the VARMA setting, the criterion function shows many local maxima where the optimization can easily get stuck. Randomization alone does not solve the problem efficiently, as typically the parameter space is high-dimensional causing problems of the curse of dimensionality type.

Moreover, VARMA modeling requires a full specification of the state space unit root structure of the process, see [15]. The state space unit root structure specifies the number of common trends at each seasonal frequency (see below for definitions). For data of weekly or higher sampling frequency it is unlikely that the state space unit root structure is known prior to estimation. Testing all possible combinations is numerically infeasible in many cases.

As an attractive alternative in this respect the class of subspace algorithms is investigated in this paper. One particular member of this class, the so called canonical variate analysis (`CVA`) introduced by [16] (in the literature the algorithm is often called canonical correlation analysis; `CCA`), has been shown to provide system estimators which (under the assumption of known system order) are asymptotically equivalent to quasi maximum likelihood estimation (using the Gaussian likelihood) in the stationary case [17]. `CVA` shares a number of robustness properties in the stationary case with VAR estimators: [18] shows that `CVA` produces consistent estimators of the underlying transfer function in situations where the innovations are conditionally heteroskedastic processes of considerable generality. Ref. [19] shows that `CVA` provides consistent estimators of the transfer function even for stationary fractionally integrated processes, if the order of the system tends to infinity as a function of the sample size at a sufficient rate.

In the I(1) case [20] introduce a heuristic adaptation of the algorithm using the assumption of known cointegrating rank in order to show consistency for the corresponding transfer function estimators. However, the specification of the cointegrating rank is no easy task in itself. In case of misspecification of the cointegrating rank the properties of this approach are unclear. Ref. [21] states without proof that also the original `CVA` algorithm delivers consistent estimates in the I(1) case without the need to impose the true cointegrating rank.

Furthermore for I(1) processes [20] proposed various tests for the cointegrating rank and compared them to tests in the Johansen framework showing superior finite sample performance in particular for multivariate data sets with a large dimension of the modeled process.

This paper builds on these results and shows that `CVA` can also be used in the seasonally integrated case. The main contributions of the paper are:

(i) It is shown that the original `CVA` algorithm in the seasonally integrated case provides strongly consistent system estimators under the assumption of known system order (thus delivering the currently unpublished proof of the claim in the I(1) case in [21]).
(ii) Upper bounds for the order of convergence for the estimated system matrices are given, establishing the familiar superconsistency for the estimation of the cointegrating spaces at all unit roots.
(iii) Several tests for separate (that is for each unit root irrespective of the specification at the other potential unit roots) determination of the seasonal cointegrating ranks are proposed which are based on the estimated systems and are simple to implement.

The derivation of the asymptotic properties of the estimators is complemented by a simulation study and an application, both demonstrating the potential of `CVA` and one of the suggested tests. Jointly our results imply that `CVA` constitutes a very reasonable initial estimate for subsequent quasi likelihood maximization in the VARMA case. Moreover the method provides valuable information on the number of unit roots present in the process, which can be used for subsequent investigation at the very least by providing upper bounds on the number of common trends present at each unit root frequency. Contrary to the JS approach in the VAR framework these tests can be performed in parallel for all unit roots, eliminating the interdependence of the results inherent in the VECM representation. Moreover, they do not use the VECM representation involving a large number of parameters in the case of high sampling rates.

These properties make `CVA` a useful tool in automatic modeling of multivariate (with a substantial number of variables) seasonally (co-)integrated processes.

The paper is organized as follows: in the next section the model set and the main assumptions of the paper are presented. The estimation methods are described in Section 3. Section 4 states the consistency results. Inference on the cointegrating ranks is proposed in Section 5. Data preprocessing is discussed in Section 6. The simulations are contained in Section 7, while Section 8 discusses an application to real world data. Section 9 concludes the paper. Appendix A contains supporting material, Appendix C provides the proofs of the main results of this paper, which are based on preliminary results presented in Appendix B.

Throughout the paper we will use the symbols $o(g_T)$ and $O(g_T)$ to denote orders of almost sure convergence where T denotes the sample size, i.e., $x_T = o(g_T)$ if $x_T/g_T \to 0$ almost surely and $x_T = O(g_T)$ if x_T/g_T is bounded almost surely for large enough T (that is there exists a constant $M < \infty$ such that $\limsup_{T\to\infty} x_T/g_T \leq M$ a.s.). Furthermore, $o_P(g_T), O_P(g_T)$ denote the corresponding in probability versions.

2. Model Set and Assumptions

In this paper state space processes $(y_t)_{t\in\mathbb{Z}}, y_t \in \mathbb{R}^s$, are considered which are defined as the solutions to the following equations for given white noise $(\varepsilon_t)_{t\in\mathbb{Z}}, \varepsilon_t \in \mathbb{R}^s, \mathbb{E}\varepsilon_t = 0, \mathbb{E}\varepsilon_t\varepsilon_t' = \Omega > 0$:

$$\begin{array}{rcl} x_{t+1} & = & Ax_t + K\varepsilon_t, \\ y_t & = & Cx_t + \varepsilon_t. \end{array} \tag{1}$$

Here $x_t \in \mathbb{R}^n$ denotes the unobserved state and $A \in \mathbb{R}^{n\times n}, C \in \mathbb{R}^{s\times n}$ and $K \in \mathbb{R}^{n\times s}$ define the state space system typically written as the tuple (A, C, K).

In this paper we consider without restriction of generality only minimal state space systems in innovations representation. For a minimal system the integer n is called the order of the system. As is well known (cf. e.g., [22]) minimal systems are only identified up to the choice of the basis of the state space. Two minimal systems (A, C, K) and $(\tilde{A}, \tilde{C}, \tilde{K})$ are observationally equivalent if and only if there exists a nonsingular matrix $\mathcal{T} \in \mathbb{R}^{n\times n}$ such that $A = \mathcal{T}\tilde{A}\mathcal{T}^{-1}, C = \tilde{C}\mathcal{T}^{-1}, K = \mathcal{T}\tilde{K}$. For two observationally equivalent systems the impulse response sequences $k_0 = I_s, k_{j+1} = CA^jK = \tilde{C}\tilde{A}^j\tilde{K}, j = 0, 1, ...$ coincide.

Ref. [15] shows that the structure of the Jordan normal form of the matrix A determines the properties (such as stationarity) of the solutions to (1) for $t \in \mathbb{Z}$. Eigenvalues of A on the unit circle lead to unit root processes in the sense of [15] who also define a *state space unit root structure* indicating the location and multiplicity of unit roots. A process $(y_t)_{t\in\mathbb{Z}}$ with state space unit root structure $\Omega_S = \{(0, (c_0)), (2\pi/S, (c_1)), ..., (\pi, (c_{S/2}))\}$ for some even integer S is called multi frequency I(1) (in short MFI(1)). Even S is chosen because it simplifies the notation by implying that $S/2$ also is an integer and $z = -1$ is a seasonal unit root. By adjusting the notation appropriately all results hold true for odd S as well).

If, moreover, such a process is observed for S periods per year, it is called *seasonal MFI(1)*. In this case the canonical form in [15] takes the following form:

$$\begin{array}{rcl} A & = & \mathrm{diag}(A_0, A_1, \ldots, A_{S/2}, A_\bullet), \\ A_0 & = & I_{c_0}, \\ A_j & = & \begin{bmatrix} \cos(\omega_j)I_{c_j} & \sin(\omega_j)I_{c_j} \\ -\sin(\omega_j)I_{c_j} & \cos(\omega_j)I_{c_j} \end{bmatrix}, \quad 0 < j < S/2, \\ A_{S/2} & = & -I_{c_{S/2}}, \\ C & = & \left[\begin{array}{c|cc|c|cc|c|c} C_{0,R} & C_{1,R} & C_{1,I} & \ldots & \ldots & C_{S/2-1,R} & C_{S/2-1,I} & C_{S/2} & C_\bullet \end{array}\right] \\ & = & \left[\begin{array}{c|c|c|c|c|c} C_0 & C_1 & \ldots & C_{S/2-1} & C_{S/2} & C_\bullet \end{array}\right], \\ K & = & \left[\begin{array}{c|cc|c|cc|c|c} K_{0,R}' & K_{1,R}' & K_{1,I}' & \ldots & \ldots & K_{S/2-1,R}' & K_{S/2-1,I}' & K_{S/2}' & K_\bullet' \end{array}\right]' \end{array} \tag{2}$$

where $\omega_j = 2\pi j/S, j = 0, \ldots, S/2$ denote the unit root frequencies and $C_{j,R} \in \mathbb{R}^{s\times c_j}, C_{j,I} \in \mathbb{R}^{s\times c_j}, K_{j,R} \in \mathbb{R}^{c_j\times s}, K_{j,I} \in \mathbb{R}^{c_j\times s}$ where $0 \leq c_j \leq s, 0 \leq j \leq S/2$. Furthermore for $C_{j,\mathbb{C}} := C_{j,R} - iC_{j,I}$ it holds that $C_{j,\mathbb{C}}'C_{j,\mathbb{C}} = I_{c_j}$ and $K_{j,\mathbb{C}} = K_{j,R} + iK_{j,I}$ is of full row rank and positive upper triangular ($C_{0,I} = C_{S/2,I} = 0, K_{0,I} = K_{S/2,I} = 0$), see [15] for details. Finally

$|\lambda_{max}(A_\bullet)| < 1$, where $\lambda_{max}(\mathcal{A})$ denotes an eigenvalue of the matrix $\mathcal{A}$ with maximal modulus. The stable subsystem $(A_\bullet, C_\bullet, K_\bullet)$ is assumed to be in echelon canonical form (see [22]).

Using this notation the assumptions on the data generating process (dgp) in this paper can be stated as follows:

Assumption 1. *$(y_t)_{t\in\mathbb{Z}}$ has a minimal state space representation $(\mathcal{A}_\circ, \mathcal{C}_\circ, \mathcal{K}_\circ), \mathcal{A}_\circ \in \mathbb{R}^{n\times n}$ of the form (2) with minimal $(\mathcal{A}_{\circ,\bullet}, \mathcal{C}_{\circ,\bullet}, \mathcal{K}_{\circ,\bullet}), \mathcal{A}_{\circ,\bullet} \in \mathbb{R}^{n_\bullet \times n_\bullet}$ in echelon canonical form where $c = n - n_\bullet > 0$.*

Furthermore the stability assumption $|\lambda_{max}(\mathcal{A}_{\circ,\bullet})| < 1$ and the strict minimum-phase condition $\rho_0 := |\lambda_{max}(\mathcal{A}_\circ - \mathcal{K}_\circ\mathcal{C}_\circ)| < 1$ hold.

The state at time $t = 1$ is given by $x_1 = [x'_{1,0}, ..., x'_{1,S/2}, x'_{1,\bullet}]'$ where $x_{1,j} \in \mathbb{R}^{\delta_j c_j}$ (for $\delta_j = 2, 0 < j < S/2$ and $\delta_j = 1$ else) is deterministic and $x_{1,\bullet} = \sum_{j=1}^{\infty} \mathcal{A}_{\circ,\bullet}^{j-1} K_{\circ,\bullet}\varepsilon_{1-j}$ is such that $(x_{t,\bullet})_{t\in\mathbb{Z}}$ is stationary.

The noise process $(\varepsilon_t)_{t\in\mathbb{Z}}$ is assumed to be a strictly stationary ergodic martingale difference sequence with respect to the filtration $\mathcal{F}_t$ with zero conditional mean $\mathbb{E}(\varepsilon_t|\mathcal{F}_{t-1}) = 0$, deterministic conditional variance $\mathbb{E}(\varepsilon_t\varepsilon'_t|\mathcal{F}_{t-1}) = \Omega > 0$ and finite fourth moments.

Due to the block diagonal form of A the state equations are in a convenient form such that partitioning the state vector accordingly as

$$x_t = \begin{pmatrix} x_{t,0} \\ x_{t,1} \\ \vdots \\ x_{t,S/2} \\ x_{t,\bullet} \end{pmatrix}, \tag{3}$$

the blocks $(x_{t,j})_{t\in\mathbb{Z}}, x_{t,j} \in \mathbb{R}^{\delta_j c_j}$ for $c_j > 0$ are unit root processes with state space unit root structure $\{(\omega_j, (c_j))\}$. Finally $(x_{t,\bullet})_{t\in\mathbb{Z}}$ is assumed to be stationary due to the assumptions on $x_{1,\bullet}$. If $(\tilde{y}_t)_{t\in\mathbb{N}}$ denotes a different solution to the state space equations corresponding to $\tilde{x}_1$ then (for $t > 1$)

$$\tilde{y}_t - y_t = CA^{t-1}(\tilde{x}_1 - x_1) = \sum_{j=0}^{S/2} C_j A_j^{t-1}(\tilde{x}_{1,j} - x_{1,j}) + C_\bullet A_\bullet^{t-1}(\tilde{x}_{1,\bullet} - x_{1,\bullet}).$$

Note that $C_j A_j^{t-1} z_{12} = \cos(\omega_j t)z_1 + \sin(\omega_j t)z_2, 0 < j < S/2$ (for appropriate vectors z_{12}, z_1, z_2),

$$C_0 A_0^{t-1} = C_0, \quad C_{S/2} A_{S/2}^{t-1} = (-1)^{t-1} C_{S/2}.$$

Therefore the sum $\sum_{j=0}^{S/2} C_j A_j^{t-1}(\tilde{x}_{1,j} - x_{1,j})$ can be modeled using a constant and seasonal dummies. The term $C_\bullet A_\bullet^{t-1}(\tilde{x}_{1,\bullet} - x_{1,\bullet})$ tends to zero with an exponential rate as $t \to \infty$ and hence does not influence the asymptotics.

Assumption 1 implies that the yearly difference

$$\begin{aligned} y_t - y_{t-S} &= CA^S x_{t-S} + \varepsilon_t + \textstyle\sum_{i=1}^{S} CA^{i-1}K\varepsilon_{t-i} - Cx_{t-S} - \varepsilon_{t-S} \\ &= (CA^S - C)x_{t-S} + v_t = (C_\bullet A_\bullet^S - C_\bullet)x_{t-S,\bullet} + v_t \end{aligned}$$

is a stationary VARMA process where $v_t = \varepsilon_t + \sum_{i=1}^{S} CA^{i-1}K\varepsilon_{t-i} - \varepsilon_{t-S}$ since $A_j^S = I_{\delta_j c_j}$. Thus the process according to Assumption 1 is a unit root process in the sense of [15]. Note that we do not assume that all unit roots are contained such that the spectral density of the stationary process $(y_t - y_{t-S})_{t\in\mathbb{Z}}$ may contain zeros due to overdifferentiation and hence the process potentially is not stably invertible. The special form of A_0 implies that $I(1)$ processes are a special case of our dgp while $I(d), d > 1, d \in \mathbb{N}$, processes are not allowed for.

3. Canonical Variate Analysis

The main idea of CVA is that, given the state, the system equations (1) are linear in the system matrices. Therefore, based on an estimate of the state sequence, the system can be estimated using least squares regression. The estimate of the state is based on the following equation (for details see for example [17]):

Let $Y_{t,f}^{+} := [y_t', y_{t+1}', \ldots, y_{t+f-1}']'$ denote the vector of stacked observations for some integer $f \geq n$ and let $E_{t,f}^{+} := [\varepsilon_t', \varepsilon_{t+1}', \ldots, \varepsilon_{t+f-1}']'$. Further define $Y_{t,p}^{-} := [y_{t-1}', \ldots, y_{t-p}']'$. Then (for $t > p$)

$$\begin{array}{rcc} Y_{t,f}^{+} & = & \mathcal{O}_f x_t + \mathcal{E}_f E_{t,f}^{+} = \mathcal{O}_f \mathcal{K}_p Y_{t,p}^{-} + \mathcal{O}_f (\mathcal{A}_\circ - \mathcal{K}_\circ \mathcal{C}_\circ)^p x_{t-p} + \mathcal{E}_f E_{t,f}^{+} \\ & = & \beta_1 Y_{t,p}^{-} + N_{t,f}^{+} \end{array} \tag{4}$$

where $\mathcal{K}_p := [\mathcal{K}_\circ, \bar{\mathcal{A}}_\circ \mathcal{K}_\circ, \bar{\mathcal{A}}_\circ{}^2 \mathcal{K}_\circ, \ldots, \bar{\mathcal{A}}_\circ{}^{p-1} \mathcal{K}_\circ]$ for $\bar{\mathcal{A}}_\circ := \mathcal{A}_\circ - \mathcal{K}_\circ \mathcal{C}_\circ$ and $\mathcal{O}_f := [\mathcal{C}_\circ', \mathcal{A}_\circ' \mathcal{C}_\circ', \ldots, (\mathcal{A}_\circ^{f-1})' \mathcal{C}_\circ']'$. The strict minimum-phase assumption implies $\bar{\mathcal{A}}_\circ{}^p \to 0$ for $p \to \infty$.

Let $\langle a_t, b_t \rangle := T^{-1} \sum_{t=p+1}^{T-f+1} a_t b_t'$ for sequences $(a_t)_{t \in \mathbb{N}}$ and $(b_t)_{t \in \mathbb{N}}$. Then an estimate of β_1 is obtained from the reduced rank regression (RRR) $Y_{t,f}^{+} = \beta_1 Y_{t,p}^{-} + N_{t,f}^{+}$ under the rank constraint $\operatorname{rank}(\beta_1) = n$. This results in the estimate $\hat{\mathcal{O}}_f \hat{\mathcal{K}}_p := [(\hat{\Xi_f^{+}})^{-1} \hat{U}_n \hat{S}_n][\hat{V}_n' (\hat{\Xi_p^{-}})^{-1}]$ of β_1 using the singular value decomposition (SVD)

$$\hat{\Xi_f^{+}} \hat{\beta}_1 \hat{\Xi_p^{-}} = \hat{U} \hat{S} \hat{V}' = \hat{U}_n \hat{S}_n \hat{V}_n' + \hat{R}_n.$$

Here $\hat{\beta}_1 = \langle Y_{t,f}^{+}, Y_{t,p}^{-} \rangle \langle Y_{t,p}^{-}, Y_{t,p}^{-} \rangle^{-1}$ denotes the unrestricted least squares estimate of β_1 and

$$\hat{\Xi_f^{+}} := \langle Y_{t,f}^{+}, Y_{t,f}^{+} \rangle^{-1/2}, \quad \hat{\Xi_p^{-}} := \langle Y_{t,p}^{-}, Y_{t,p}^{-} \rangle^{1/2}. \tag{5}$$

Here the symmetric matrix square root is used. The definition is, however, not of importance and other square roots such as Cholesky factors could be used. $\hat{U}_n \in \mathbb{R}^{fs \times n}$ denotes the matrix whose columns are the left singular vectors to the n largest singular values which are the diagonal entries in $\hat{S}_n := \operatorname{diag}(\hat{\sigma}_1, \hat{\sigma}_2, \ldots, \hat{\sigma}_n), \hat{\sigma}_1 \geq \cdots \geq \hat{\sigma}_n > 0$ and $\hat{V}_n \in \mathbb{R}^{ps \times n}$ contains the corresponding right singular vectors as its columns. $\hat{R}_n$ denotes the approximation error.

The system estimate $(\hat{A}, \hat{C}, \hat{K})$ is then obtained using the estimated state $\hat{x}_t := \hat{\mathcal{K}}_p Y_{t,p}^{-}, t = p+1, \ldots, T+1$ via regression in the system equations.

In the algorithm a specific decomposition of the rank n matrix $\hat{\mathcal{O}}_f \hat{\mathcal{K}}_p$ into the two factors $\hat{\mathcal{O}}_f$ and $\hat{\mathcal{K}}_p$ is given such that $\hat{\mathcal{K}}_p \hat{\Xi_p^{-}} (\hat{\Xi_p^{-}})' \hat{\mathcal{K}}_p{}' = I_n$. It is obvious that every other decomposition of $\hat{\mathcal{O}}_f \hat{\mathcal{K}}_p$ produces an estimated state sequence in a different coordinate system, leading to a different observationally equivalent representation of the same transfer function estimator. Therefore, with respect to consistency of the transfer function estimator it is sufficient to show that there exists a factorization of $\hat{\mathcal{O}}_f \hat{\mathcal{K}}_p$ leading to convergent system matrix estimators $(\tilde{A}, \tilde{C}, \tilde{K})$, even if this factorization cannot be used in actual computations, as it requires information not known at the time of estimation.

In order to generate a consistent initial guess for subsequent quasi likelihood optimization in the set of all state space systems corresponding to processes with state space unit root structure $\Omega_S := \{(\omega_0, (c_0)), ..., (\omega_{S/2}, (c_{S/2}))\}$, however, we will derive a realizable (for known integers c_j and matrices E_j such that $E_j' \mathcal{C}_{\circ,j,\mathbb{C}} = I_{c_j}$) consistent system estimate. To this end note that consistency of the transfer function implies (see for example [23]) that the eigenvalues $\tilde{\lambda}_l$ of $\hat{A}$ converge (in a specific sense) to the eigenvalues λ_j of $\mathcal{A}_\circ$. Therefore transforming $\hat{A}$ into complex Jordan normal form (where $\hat{A}$ is almost surely diagonalizable), ordering the eigenvalues such that groups of eigenvalues $\tilde{\lambda}_l(j), l = 1, ..., c_j$, converging to λ_j are grouped together, we obtain a realizable system $(\check{A}, \check{C}, \check{K})$ where the diagonal blocks of the block diagonal matrix $\check{A}$ corresponding to the unit roots converge to a diagonal matrix with the eigenvalues z_j on the diagonal:

$$\check{A}_{j,\mathbb{C}} = \begin{bmatrix} \tilde{\lambda}_1(j) & 0 & \dots & 0 \\ 0 & \tilde{\lambda}_2(j) & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \tilde{\lambda}_{c_j}(j) \end{bmatrix} \to A_{j,\mathbb{C}} = \begin{bmatrix} z_j & 0 & \dots & 0 \\ 0 & z_j & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & z_j \end{bmatrix}.$$

Replacing $\check{A}_{j,\mathbb{C}}$ by the limit $A_{j,\mathbb{C}}$ and transforming the estimates to the real Jordan normal form, we obtain estimates $(\check{A}, \check{C}, \check{K})$ corresponding to unit root processes with state space unit root structure Ω_S.

Note, however, that this representation not necessarily converges as perturbation analysis only implies convergence of the eigenspaces. Therefore in the final step the estimate $(\check{A}, \check{C}, \check{K})$ is converted such that we obtain convergence of the system matrix estimates. In the class of observationally equivalent systems with the matrix

$$\check{A}_{\mathbb{C}} = \text{diag}(A_{0,\mathbb{C}}, A_{1,\mathbb{C}}, \overline{A_{1,\mathbb{C}}}..., \overline{A_{S/2-1,\mathbb{C}}}, A_{S/2,\mathbb{C}}, \check{A}_\bullet), \quad A_{j,\mathbb{C}} = I_{c_j} z_j,$$

block diagonal transformations of the form $\mathcal{T} = \text{diag}(\mathcal{T}_0, \mathcal{T}_1, \overline{\mathcal{T}_1}, ..., \mathcal{T}_{S/2}, I)$ do not change the matrix $\check{A}_{\mathbb{C}}$. Here the basis of the stable subsystem can be chosen such that the corresponding transformed $(\check{A}_\bullet, \check{C}_\bullet, \check{K}_\bullet)$ is uniquely defined using an overlapping echelon form (see [22], Section 2.6). The impact of such transformations on the blocks of C is given by $\check{C}_{j,\mathbb{C}}\mathcal{T}_j^{-1}$. Therefore, if for each $j = 0, ..., S/2$ a matrix $E_j \in \mathbb{C}^{s \times c_j}$ is known such that $E_j' \mathcal{C}_{\circ,j,\mathbb{C}} \in \mathbb{C}^{c_j \times c_j}$ is nonsingular, the restriction $E_j' \check{C}_{j,\mathbb{C}} = I_{c_j}$ picks a unique representative $(\check{A}, \check{C}, \check{K})$ of the class of systems observationally equivalent to $(\check{A}, \check{C}, \check{K})$.

Note that this estimate $(\check{A}, \check{C}, \check{K})$ is realizable if the integers c_j (needed to identify the c_j eigenvalues of $\hat{A}$ closest to z_j), the matrices E_j (needed to fix a basis for $x_{t,j}$) and the index corresponding to the overlapping echelon form for the stable part are known. Furthermore, this estimate corresponds to a process with state space unit root structure Ω_S and hence can be used as a starting value for quasi likelihood maximization.

Finally in this section it should be noted that the estimate of the state $\hat{x}_t$ here mainly serves the purpose of obtaining an estimator for the state space system. Based on this estimate, Kalman filtering techniques can be used to obtain different estimates of the state sequence. The relation between these different estimates is unclear and so is their usage for inference. For this paper the state estimates $\hat{x}_t$ are only an intermediate step in the `CVA` algorithm.

4. Asymptotic Properties of the System Estimators

As follows from the last section, the central step in the `CVA` procedure is a RRR problem involving stationary and nonstationary components. The asymptotic properties of the solution to such RRR problems are derived in Theorem 3.2. of [24]. Using these results the following theorem can be proved (see Appendix C.1):

Theorem 1. *Let the process $(y_t)_{t\in\mathbb{Z}}$ be generated according to Assumption 1. Let $(\hat{A}, \hat{C}, \hat{K})$ denote the* `CVA` *estimators of the system matrices using the assumption of correctly specified order n with $f \geq n$ not depending on the sample size and finite and $p = o((\log T)^{\bar{a}})$ for some real $0 < \bar{a} < \infty$, $p \geq -d \log T / \log \rho_0$ for some $d > 1$ where $0 < \rho_0 = |\lambda_{max}(\mathcal{A}_\circ - \mathcal{K}_\circ \mathcal{C}_\circ)| < 1$. Let $(\mathcal{A}_\circ, \mathcal{C}_\circ, \mathcal{K}_\circ)$ be in the form given in (2) where $(\mathcal{A}_{\circ,\bullet}, \mathcal{C}_{\circ,\bullet}, \mathcal{K}_{\circ,\bullet})$ is in echelon canonical form and for each $j = 0, ..., S/2$ there exists a row selector matrix $E_j \in \mathbb{R}^{s \times c_j}$ such that $E_j' \mathcal{C}_{\circ,j,\mathbb{C}}$ is non-singular. Then for some integer a:*

(I) $\hat{C}\hat{A}^j\hat{K} - \mathcal{C}_\circ \mathcal{A}_\circ^j \mathcal{K}_\circ = O_P((\log T)^a / \sqrt{T})$ *for each* $j \geq 0$.

(II) Using $D_x = diag(T^{-1} I_c, T^{-1/2} I_{n-c})$ *where* $c = \sum_{j=0}^{S/2} c_j \delta_j$ *we have*

$$(\check{A} - \mathcal{A}_\circ) D_x^{-1} = O_P((\log T)^a), \sqrt{T}(\check{K} - \mathcal{K}_\circ) = O_P((\log T)^a), (\check{C} - \mathcal{C}_\circ) D_x^{-1} = O_P((\log T)^a)$$

for some integer $a < \infty$.

(III) *If the noise is assumed to be an iid sequence, then results (I) and (II) hold almost surely.*

Beside stating consistency in the seasonal integration case, the theorem also improves on the results of [20] in the I(1) case by showing that no adaptation of `CVA` is needed in order to obtain consistent estimators of the impulse response sequence or the system matrices. Note that this consistency result for the impulse response sequence concerns both the short and the long-run dynamics. In particular it implies that short-run prediction coefficients are consistent. Moreover the theorem establishes strong consistency rather than weak consistency as opposed to [20]. (II) establishes orders of convergence which, however, apply only to a transformed system that requires knowledge of the integers c_j and matrices E_j to be realized. No tight bounds for the integer a are derived, since they do not seem to be of much value.

Note that the assumptions on the innovations rule out conditionally heteroskedastic processes. However, since the proof mostly relies on convergence properties for covariance estimators for stationary processes and continuous mapping theorems for integrated processes, it appears likely that the results can be extended to conditionally heteroskedastic processes as well. For the stationary cases this follows directly from the arguments in [18], while for integrated processes results (using different assumptions on the innovations) given for example in [25] can be used. The conditions of [25] hold for example in a large number of GARCH type processes, see [26]. The combination of the different sets of assumptions on the innovations is not straightforward, however, and hence would further complicate the proofs. We refrain from including them.

It is worth pointing out that due to the block diagonal structure of $\mathcal{A}_\circ$ the result $(\check{\mathcal{C}} - \mathcal{C}_\circ)D_x^{-1} = O_P((\log T)^a)$ implies consistency of the blocks $\check{\mathcal{C}}_j$ corresponding to unit root z_j (or the corresponding complex pair) of order almost T^{-1}. Using the complex valued canonical form this implies consistent estimation of $\mathcal{C}_{\circ,j,\mathbb{C}}$ by the corresponding $\check{\mathcal{C}}_{j,\mathbb{C}}$. In the canonical form this matrix determines the cointegrating relations (both the static as well as the dynamic ones, for details see [15]) as the unitary complement to this matrix. It thus follows that `CVA` delivers estimators for the cointegrating relations at the various unit roots that are (super-)consistent. In fact, the proof can be extended to show convergence in distribution of $(\check{\mathcal{C}} - \mathcal{C}_\circ)D_x^{-1}$. This distribution could be used in order to derive tests for cointegrating relations. However, preliminary simulations indicate that these estimates and hence the corresponding tests are not optimal and can be improved upon by quasi maximum likelihood estimation in the VARMA setting initialized by the `CVA` estimates. Therefore we refrain from presenting these results.

Note that the assumptions impose the restriction $\rho_0 > 0$ excluding VAR systems. This is done solely for stating a uniform lower bound on the increase of p as a function of T. This bound is related to the lag length selection achieved using BIC, see [27]. In the VAR case the lag length estimator using BIC will converge to the true order and thus remain finite. All results hold true if in the VAR case a fixed (that is independent of the sample size) $p \geq n$ is used.

5. Inference Based on the Subspace Estimators

Beside consistency of the impulse response sequence also the specification of the integers $c_0, ..., c_{s/2}$ is of interest. First, following [20] this information can be obtained by detecting the unity singular values in the RRR step of the procedure. Second, from the system representation (2) it is clear that the location of the unit roots is determined by the eigenvalues of $\mathcal{A}_\circ$ on the unit circle: The integers c_j denote the number of eigenvalues at the corresponding locations on the unit circle (provided the eigenvalues are simple). Due to perturbation theory (see for example Lemma A2) we know that the eigenvalues of $\hat{A}$ will converge (for $T \to \infty$) to the eigenvalues of $\mathcal{A}_\circ$ and the distribution of the mean of all eigenvalues of $\hat{A}$ converging to an eigenvalue of $\mathcal{A}_\circ$ can be derived based on the distribution of the estimation error $\hat{A} - \mathcal{A}_\circ$. This can be used to derive tests on the number

of eigenvalues at a particular location on the unit circle. Third, if $n \leq s$ the state process is a VAR(1) process and hence in some cases allows for inference on the number of cointegrating relations and thus also on the integers c_j as outlined in [4]. Tests based on these three arguments will be discussed below.

Theorem 2. *Under the assumptions of Theorem 1 the test statistic* $T\sum_{i=1}^{c}(1-\hat{\sigma}_i^2)$ *converges in distribution to the random variable*

$$Z = tr\left[\mathbb{E}(\tilde{\varepsilon}_{t,\perp}\tilde{\varepsilon}_{t,\perp}')\left(\int_0^1 W(r)W(r)'\right)^{-1}\right]$$

where $\tilde{\varepsilon}_{t,\perp} = \tilde{\varepsilon}_{t,1} - \mathbb{E}\tilde{\varepsilon}_{t,1}\tilde{\varepsilon}_{t,\bullet}'(\mathbb{E}\tilde{\varepsilon}_{t,\bullet}\tilde{\varepsilon}_{t,\bullet}')^{-1}\tilde{\varepsilon}_{t,\bullet}$ *(for definition of* $\tilde{\varepsilon}_{t,1}$ *and* $\tilde{\varepsilon}_{t,\bullet}$ *see the proof in Appendix C.2) and where* $W(r)$ *denotes a c-dimensional Brownian motion with variance*

$$\sum_{i=0}^{S-1}\mathcal{A}_u^i\mathcal{K}_u\Omega\mathcal{K}_u'(\mathcal{A}_u^i)'$$

with $\mathcal{A}_u$ *denoting the* $c \times c$ *heading submatrix of* $\mathcal{A}$ *and* $\mathcal{K}_u$ *denoting the submatrix of* $\mathcal{K}$ *composed of the first c rows such that* $(\mathcal{A}_u, \mathcal{C}_u, \mathcal{K}_u)$ *denotes the unstable subsystem.*

The theorem is proved in Appendix C.2, where also the many nuisance parameters of the limiting random variable are explained and defined. The proof also corrects an error in Theorem 4 of [20], where the wrong distribution is given since the second order terms were neglected.

As the distribution is not pivotal and in particular contains information that is unknown when performing the RRR step, it is not of much interest for direct application. Nevertheless the order of convergence allows for the derivation of simple consistent estimators of the number of common trends: Let $\hat{c}_T$ denote the number of singular values calculated in the RRR that exceed $\sqrt{1-h(T)/T}$ for arbitrary $h(T) \to \infty, h(T) < T, h(T)/T \to 0$, for $T \to \infty$. Then it is a direct consequence of Theorem 2 in combination with $\hat{\sigma}_j \to \sigma_j < 1, j > c$, that $\hat{c}_T \to c$ in probability, implying consistent estimation of c. Based on these results also estimators for c could be derived , for example along the lines of [28]. However, as [29] shows, such estimators have not performed well in simulations and thus are not considered subsequently.

The singular values do not provide information on the location of the unit roots. This additional information is contained in the eigenvalues of the matrix $\mathcal{A}_\circ$:

Theorem 3. *Under the assumptions of Theorem 1 let* $\hat{\lambda}_i(m), i = 1, ..., c_m$ *denote the* c_m *eigenvalues of* $\hat{\mathcal{A}}$ *closest to the unit root* $z_m, |z_m| = 1$. *Then defining* $\hat{\mu}_m = \sum_{i=1}^{c_m}(\hat{\lambda}_i(m) - z_m)$ *we obtain*

$$T\hat{\mu}_m \xrightarrow{d} tr\left[\left(\int B(r)B(r)dr\right)^{-1}\int B(r)dB(r)'\right]$$

where $B(r)$ *denotes a* c_m*-dimensional Brownian motion with zero expectation and variance* I_{c_m} *for* $z_m = \pm 1$ *and a complex Brownian motion with expectation zero and variance equal to the identity matrix else.*

Further if $\tilde{\mathcal{A}} := \langle x_{t+1}, x_t\rangle\langle x_t, x_t\rangle^{-1}$ *using the true state* x_t *and* $\tilde{\mu}_m = \sum_{i=1}^{c_m}(\tilde{\lambda}_i(m) - z_m)$ *where* $\tilde{\lambda}_i(m), i = 1, ..., c_m$ *denote the* c_m *eigenvalues of* $\tilde{\mathcal{A}}$ *closest to* z_m*, then* $T(\hat{\mu}_m - \tilde{\mu}_m) = o_P(1)$.

Therefore the estimated eigenvalues can be used in order to obtain a test on the number of common trends at a particular frequency for each frequency separately. The test distribution is obtained as the limit to

$$T\text{tr}[\langle\mathcal{K}_{\circ,m,\mathbb{C}}\varepsilon_t, x_{t,m,\mathbb{C}}\rangle\langle x_{t,m,\mathbb{C}}, x_{t,m,\mathbb{C}}\rangle^{-1}]$$

where $x_{t,m,\mathbb{C}} = \overline{z_m} x_{t-1,m,\mathbb{C}} + \mathcal{K}_{\circ,m,\mathbb{C}} \varepsilon_{t-1}, x_{1,m,\mathbb{C}} = 0$. The distribution thus does not depend on the presence of other unit roots or stationary components of the state. Furthermore it can be seen that it is independent of the noise variance or the matrix $\mathcal{K}_{\circ,m,\mathbb{C}}$. Hence critical values are easily obtained from simulations. Also note that the limiting distribution is identical for all complex unit roots.

Therefore, for each seasonal unit root location z_m we can order the eigenvalues of the estimated matrix $\hat{\mathcal{A}}$ with increasing distance to z_m. Then starting from the assumption of $H_0 : c_m = \bar{c}$ (for a reasonable $\bar{c}$ obtained, e.g., from a plot of the eigenvalues) one can perform the test with statistic $T\hat{\mu}_m$. If the test rejects, then the hypothesis $H_0 : c_m = \bar{c} - 1$ is tested, until the hypothesis is not rejected anymore, or $H_0 : c_m = 1$ is reached. This is then the last test. If H_0 is rejected again, no unit root is found at this location. Otherwise we do not have evidence against $c_m = 1$. In any case, the system needs to be estimated only once and the calculation of the test statistics is easy even for all seasonal unit roots jointly.

The third option for obtaining tests is to use the tests derived in [4] based on the JS framework for VARs. In the case $n \leq s$ the state process $x_{t+1} = \mathcal{A}x_t + \mathcal{K}\varepsilon_t$ is a seasonally integrated VAR(1) process (for $n > s$ the noise variance is singular). The corresponding VECM representation equals

$$p(L)x_t = \sum_{m=1}^{S} (I_n - \mathcal{A}z_m) X_{t-1}^{(m)} + \mathcal{K}\varepsilon_{t-1} = \sum_{m=1}^{S} \alpha_m \beta_m' X_{t-1}^{(m)} + \mathcal{K}\varepsilon_{t-1}$$

where $z_m = \exp(\frac{2\pi m}{S} i), m = 1, ..., S$ and

$$p(L) = 1 - L^S \quad , \quad p_t = p(L)x_t = x_t - x_{t-S},$$
$$p_m(L) = \frac{p(L)}{1 - \overline{z_m} L} \quad , \quad X_t^{(m)} = -\frac{p_m(L)}{p_m(z_m) z_m} x_t.$$

Note that in this VAR(1) setting no additional stationary regressors of the form $p(L)x_{t-j}$ occur. Also no seasonal dummies are needed but could be added to the equation. In this setting [4] suggests to use the eigenvalues $\hat{\lambda}_i$ (ordered with increasing modulus) of the matrix (the superscript $(.)^\pi$ denotes the residuals with respect to the remaining regressors $X_{t-1}^{(j)}, j \neq m$)

$$\langle X_{t-1}^{(m),\pi}, p_t^\pi \rangle \langle p_t^\pi, p_t^\pi \rangle^{-1} \langle p_t^\pi, X_{t-1}^{(m),\pi} \rangle \langle X_{t-1}^{(m),\pi}, X_{t-1}^{(m),\pi} \rangle^{-1}$$

as the basis for a test statistic

$$\tilde{C}_m := -\delta_m \sum_{i=1}^{c_m} \log(1 - \hat{\lambda}_i).$$

where $\delta_m = 2$ for complex unit roots and $\delta_m = 1$ for real unit roots. In the $I(1)$ case this leads to the familiar Johansen trace test, for seasonal unit roots a different asymptotic distribution is obtained.

Theorem 4. *Under the assumptions of Theorem 1 let $\hat{C}_m$ be calculated based on the estimated state and let $\tilde{C}_m$ denote the same statistic based on the true state. Then for $n \leq s$ it holds that $\hat{C}_m - \tilde{C}_m = o_P(T^{-1})$ and*

$$T\hat{C}_m \xrightarrow{d} tr\left[\int dB(r) B(r)' \left(\int B(r) B(r) dr \right)^{-1} \int B(r) dB(r)' \right]$$

where $B(r)$ is a real Brownian motion for $z_m = \pm 1$ or a complex Brownian motion else.

Thus again under the null hypothesis the test statistic based on the estimated state and the one based on the true state reject jointly asymptotically with probability one. Therefore

for $n \leq s$ the tests of JS can be used to obtain information on the number of common cycles, ignoring the fact that the estimated state is used in place of the true state process.

After presenting three disjoint ideas for providing information on the number and location of unit roots, the question arises, which one to use in practice. In the following a number of ideas are given in this respect.

The criterion based on the singular values given in Theorem 2 is of limited information as it only provides the overall number of unit roots. Since the limiting distribution is not pivotal it cannot be used for tests and the choice of the cutoff value $h(T)$ is somewhat arbitrary. Nevertheless, using a relatively large value one obtains a useful upper bound on c which can be included in the typical sequential procedures for tests for c_j.

Using the results of Theorem 4 has the advantage of using a framework that is well known to many researchers. It is remarkable that in terms of the asymptotic distributions there is no difference involved in using the estimated state in place of the true state. The assumption $n \leq s$, however, is somewhat restrictive except in situations with a large s.

Finally the results of Theorem 3 provide simple to use tests for all unit roots , independently of the specification of the model for the remaining unit roots. Again it is remarkable that, under the null, inference is identical for known and for estimated state.

Since our estimators are not quasi maximum likelihood estimators the question of a comparison with the usual likelihood ratio tests arises. For VAR models simulation exercises documented in Section 7 below demonstrate that there are situations where the proposed tests outperform tests in the VAR framework. Comparisons with tests in the state space framework (or equivalently in the VARMA framework) are complicated by the fact that no results are currently available in the literature of this framework. One difference, however, is given by the fact that quasi likelihood ratio tests in the VARMA setting require a full specification of the c_j values for all unit roots. This introduces interdependencies such that the tests for one unit root depend on the specification of the cointegrating rank at the other roots. The interdependencies can be broken by performing tests based on alternative specifications for each unit root. The test based on Theorem 3 does not require this but can be performed based on the same estimate $\hat{A}$. This is seen as an advantage.

The question of the comparison of the empirical size in finite samples as well as power to local alternatives between the `CVA` based tests and tests based on quasi-likelihood ratios is left as a research question.

6. Deterministic Terms

Up to now it has been assumed that no deterministic terms appear in the model contrary to common practice. In the VAR framework dealing with trends is complicated by the usage of the VECM representation, see e.g., [30]. In the state space framework used in this paper, however, deterministic terms are easily incorporated.

Theorem 5. *Let the process $(y_t)_{t\in\mathbb{Z}}$ be generated according to Assumption 1 and assume that the process $(\tilde{y}_t)_{t\in\mathbb{Z}}$ is observed where $\tilde{y}_t = y_t + \Phi d_t$ with*

$$d_t = \left[\begin{array}{cccccc} 1, & \cos(\frac{2\pi}{S}t), & \sin(\frac{2\pi}{S}t), & \cdots & (-1)^t & \end{array} \right]' \in \mathbb{R}^S$$

and $\Phi \in \mathbb{R}^{s\times S}$.

Then if the `CVA` *estimation is applied to*

$$\tilde{y}_t^{\pi} := y_t - \left(\sum_{t=1}^{T} y_t d_t'\right)\left(\sum_{t=1}^{T} d_t d_t'\right)^{-1} d_t, \quad t = 1, ..., T,$$

the results of Theorem 1 hold, i.e., the system is estimated consistently and the orders of convergence for the transformed system $(\check{A}, \check{C}, \check{K})$ hold true.

Furthermore the convergence in distribution results in Theorems 2–4 hold true where in the limits the Brownian motions $B(r)$ occurring in the distributions must be replaced by their demeaned versions $B(r) - \int_0^1 B(s)ds$.

In this sense the results are robust to some operations typically termed preprocessing of data such as demeaning and deseasonalizing using seasonal dummies. More general preprocessing steps such as detrending or the extraction of more general deterministic terms analogous to [30] can be investigated along the same lines.

7. Simulations

The estimation of the seasonal cointegration ranks and spaces is usually carried out via quasi maximum likelihood methods that originated from the VAR model class. Typical estimators in this setting are those of [2,4,5,31]. In the first two experiments we focus on the estimation of the cointegrating spaces and the specification of the cointegration ranks in the classical situation of quarterly data and show that there are certain situations in which `CVA` estimators and the test in Theorem 3 possess finite sample properties superior to those of the methods above. In a third experiment the test performance is evaluated for a daily sampling rate. Moreover, the prediction accuracy of `CVA` is investigated as well as its robustness to innovations exhibiting behaviors often encountered at such higher sampling rates. All simulations are carried out using 1000 replications.

To investigate the practical usefulness of the proposed procedures we generate quarterly data using two VAR dgps of dimension $s = 2$ first and then two more general VARMA dgps with $s = 8$. Each pair contains dgps with different state space unit root structures

$$\{(0,(1)),(\pi/2,(c_{\pi/2})),(\pi,(1))\}, \quad c_{\pi/2} = 1,2.$$

From all four dgps samples of size $T \in \{50, 100, 200, 500\}$ are generated with initial values set to zero. Although none of the dgps contains deterministics, the data is adjusted for a constant and quarterly seasonal dummies as in [5]. For reasons of comparability, the adjustment for deterministic terms is done before estimation.

In the third experiment we generate daily data with dimension $s = 4$ from a state space system including unit roots corresponding to weekly frequencies (that is a period length of seven days). In the simulations we use several years of data (excluding new year's day to account for 52 weeks of seven days each). The first 200 observations are discarded to include the effects of different starting values. In this example the focus lies on a comparison of the prediction accuracy. Furthermore we investigate the robustness of the test procedures to conditional heteroskedasticity of the GARCH type as well as to non-normality of the innovations.

To assess the performance of specifying the cointegrating rank at unit root z using `CVA`, the following test statistic is constructed from the results in Theorem 3

$$\Lambda(c) = T|(\frac{1}{c}\sum_{i=1}^{c}\hat{\lambda}_i) - z| \,. \tag{6}$$

Here $\hat{\lambda}_1, \ldots, \hat{\lambda}_n$ are the eigenvalues of $\hat{A}$ ordered increasingly according to the distance from z. Note that a similar test in [20] only uses the c-th largest eigenvalue, whereas here the average over the nearest c eigenvalues is taken. Critical values have been obtained by simulation using large sample sizes (sample size 2000 (JS) and 5000 (`CVA`), 10,000 replications).

In our first two experiments usage of $\Lambda(c)$ is compared with variants of the likelihood ratio test from [2] (JS), [4] (Q_1), and [5] (Q_2, Q_3). Q_1 is Cubadda's trace test for complex-valued data, Q_2 takes the information at frequency $\pi/2$ into account when the analysis is carried out at frequency $3\pi/2$, and Q_3 iterates between $\pi/2$ and $3\pi/2$ in the alternating reduced rank regression (ARR) of [5]. For the procedure of [2] the likelihood maximization

at frequency $\pi/2$ is carried out using numerical optimization (BFGS) with initial values obtained from an unrestricted regression.

All tests are evaluated by comparing the percentages of correctly detected common trends, or *hit rates*, with 0.95, the hit rate to be expected from a nominal significance level of 0.05. The testing procedure employed for all tests is the same: at each of the frequencies it is started from a null hypothesis of s unit roots against less than s unit roots. In case of rejection, $s-1$ unit roots are tested versus less than $s-1$ and so on, until there are zero unit roots under the alternative.

For the first two experiments the estimation performance of `CVA` for the simultaneous estimation of the seasonal cointegrating spaces is compared with the maximum likelihood estimates of [2,4,31] (cRRR), and also with an iterative procedure (Generalized ARR or GARR) of [5]. The comparison is carried out by means of the gap metric, measuring the distance between the true and the estimated cointegrating space as in [32]. The smaller the mean gap over all replications, the better is the estimation performance. Throughout a difference between two mean gaps or two hit rates is considered statistically significant if it is larger than twice the Monte Carlo standard error.

For all procedures used in this section, an AR lag length has to be chosen first. For `CVA` this can be done using the AIC as in ([33], Section 5), as is done in the third experiment.

In the first two experiments where sample sizes are rather small, we estimate the lag length via minimization of the corrected AIC (AICc) ([34], p. 432), $\hat{k}_{AICc}$, benefitting the simulation results. For larger sample sizes the two criteria lead to the same choices. Due to the quarterly data we work with, the lag length is then chosen to be $\hat{k} = \max\{\hat{k}_{AICc}, 4\}$.

Other information criteria could be chosen here. An anonymous referee also suggested the application of the Modified Akaike Information Criterion (MAIC) of [35], proposed there for the I(1)-case. In an attempt to apply it to the seasonally integrated case considered here, it performed considerably worse than the AICc. Thus we refrain from using the MAIC in the following and also omit the results of that attempt. They can be obtained from the authors upon request.

For `CVA` the truncation indices f and p are chosen as $\hat{f} = \hat{p} = 2\hat{k}$ ([33], Section 5). The system order n is estimated by minimizing ([33], Section 5)

$$SVC(n) = \hat{\sigma}_{n+1}^2 + 2ns\frac{\log T}{T}\,. \tag{7}$$

Here $\hat{\sigma}_i$ denotes the i-th largest singular value from the singular value decomposition of $\hat{\Xi}_f^+ \hat{\beta}_1 \hat{\Xi}_p^-$ (Step 2 of `CVA`). Note that selecting the number of states by SVC is made less influential insofar as $\hat{n} = \max\{c_0 + 2c_{\pi/2} + c_\pi, \hat{n}_{SVC}\}$, where $\hat{n}_{SVC}$ denotes the SVC estimated system order.

In Section 7.1 we start with the two VAR dgps and find that the likelihood-based procedures are mostly superior. Continuing with the VARMA dgps in Section 7.2, `CVA` performs better and is superior for the smaller sample sizes in terms of size and gap and better for all sample sizes in terms of power. Section 7.3 evaluates the performance of the tests for unit roots for larger sample sizes together with the prediction performance in this setting. We find that the tests are robust to the distribution of the innovations as well as to conditional heteroskedasticity of the GARCH type. Furthermore the empirical size of the tests lies close to the size already for moderate sample sizes, where the tests also show almost perfect power properties.

7.1. VAR Processes

The VAR dgps considered in this paper are given by,

$$X_t = \Pi_1 X_{t-1} + \Pi_2 X_{t-2} + \Pi_3 X_{t-3} + \Pi_4 X_{t-4} + \varepsilon_t, \qquad \varepsilon_t \sim \mathrm{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}\right) \tag{8}$$

where $(\varepsilon_t)_{t\in\mathbb{Z}}$ is white noise and the coefficient matrices are

$$\begin{aligned}\Pi_1 &= \begin{bmatrix}\gamma & 0\\ 0 & 0\end{bmatrix}, \Pi_2 = \begin{bmatrix}-0.4 & 0.4-\gamma\\ 0 & 0\end{bmatrix},\\ \Pi_3 &= \begin{bmatrix}-\gamma & 0\\ 0 & 0\end{bmatrix}, \Pi_4 = \begin{bmatrix}0.6-(\gamma/10) & 0.4+\gamma\\ 0 & 1\end{bmatrix}.\end{aligned}$$

This dgp is adopted from [5] with a slight adjustment to Π_4. The corresponding VECM representation in the notation of [5] equals

$$\begin{aligned}X_{0,t} = &\begin{bmatrix}-0.2\\ 0\end{bmatrix}\begin{bmatrix}1+\gamma/8 & -1\end{bmatrix}X_{1,t-1} + \begin{bmatrix}0.2\\ 0\end{bmatrix}\begin{bmatrix}1+\gamma/8 & -1\end{bmatrix}X_{2,t-1} +\\ &\begin{bmatrix}\gamma\\ 0\end{bmatrix}\begin{bmatrix}1+0.05L & -L\end{bmatrix}X_{3,t-1}+\varepsilon_t.\end{aligned}$$

As can be seen from Table 1, the dgps possess unit roots at frequencies 0, π, and $\pi/2$, where $c_{\pi/2} = 2[1]$ for $\gamma = 0[0.2]$, respectively. Note that in all cases the order of integration equals 1, while the number of common cycles at $\pi/2$ is varied.

Table 1. Eigenvalues of the coefficient matrix of the companion form.

		j							
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
$\gamma = 0.2$	z_j	-1	1	i	$-$i	0.126 + i0.99	0.126 $-$ i0.99	-0.790	0.737
	$\lvert z_j\rvert$	1	1	1	1	0.998	0.998	0.790	0.737
$\gamma = 0$	μ_j	-1	i	$-$i	1	i	$-$i	0.775	-0.775
	$\lvert\mu_j\rvert$	1	1	1	1	1	1	0.775	0.775

Table 2 exhibits the hit rates from the application of the different test statistics. At frequencies 0 and π, Λ is compared with the trace test of Johansen (J; based on [31] for unit roots $z = -1$), whereas at $\pi/2$ it is competing with JS, Q_1, Q_2, and Q_3. All competitors are likelihood-based tests which is the term we are referring to when we compare Λ to them as a whole.

Table 2. Hit rates for the different tests (VAR dgp). Twice the maximum (over all entries) Monte Carlo standard error is 0.005.

		0		**π/2**					**π**	
	T	**Λ**	**J**	**Λ**	**JS**	**Q1**	**Q2**	**Q3**	**Λ**	**J**
$\gamma = 0$	50	0.685	0.348	0.351	0.903	0.844	0.851	0.844	0.681	0.343
	100	0.841	0.732	0.490	0.925	0.900	0.902	0.900	0.831	0.724
	200	0.897	0.951	0.841	0.934	0.925	0.924	0.925	0.876	0.936
	500	0.931	0.938	0.916	0.949	0.941	0.942	0.941	0.927	0.948
$\gamma = 0.2$	50	0.550	0.367	0.811	0.796	0.777	0.778	0.788	0.604	0.297
	100	0.711	0.801	0.087	0.920	0.913	0.908	0.908	0.799	0.806
	200	0.907	0.922	0.855	0.954	0.949	0.948	0.947	0.854	0.939
	500	0.944	0.953	0.927	0.939	0.938	0.938	0.936	0.924	0.942

The results for 0 and π are very similar for both dgps in that Λ scores more hits than the likelihood-based tests when the sample size is small, $T \in \{50, 100\}$. Convergence of its finite sample distribution is slower than for the other test statistics, however, as J is closer to 0.95 from $T = 200$ on. For $T = 500$ the distribution of Λ only seems to have converged

to its asymptotic distribution when $c_{\pi/2} = 2$ at frequency 0, whereas convergence of the likelihood-based tests has occurred in all cases.

At $\pi/2$ the likelihood ratio test of JS strictly dominates all implementations of [5] for all sample sizes and both dgps. It strictly dominates the CVA-based test procedure as well, except for one case, it seems: when $c_{\pi/2} = 1$ and $T = 50$ Λ scores slightly, but significantly, more hits than the likelihood ratio test of JS. Surprisingly, Λ is drastically worse when $T = 100$ with only 8.7%, only to be up at 85% for $T = 200$.

The behavior of Λ is explained by z_5 and z_6 being close to $\pm i$ when $c_{\pi/2} = 1$, cf. Table 1. For future reference we will call the corresponding roots *false unit roots*.

For $T = 50$ the estimates of eigenvalues corresponding to actual unit roots are rather not very close to $\pm i$ in contrast to the false unit roots. Thus the latter are mistaken for actual unit roots (cf. the first panel in Figure 1), leading to a hit rate of 81.1%, one that is even larger than the rates at 0 and π. As the sample size increases, the eigenvalue estimates of the true unit roots become more and more accurate, visible from the second and third panel in Figure 1. Accordingly they can be detected correctly more often. Unfortunately however, for $T = 100$ the false unit roots remain to be detected such that often two instead of just one unit root are found by Λ, resulting in a hit rate of only 8.7%. For $T \in \{200, 500\}$ Λ is able to distinguish the false unit roots from the true ones and the detection rate is getting closer to the asymptotic rate, 85.5% and 92.7%, respectively.

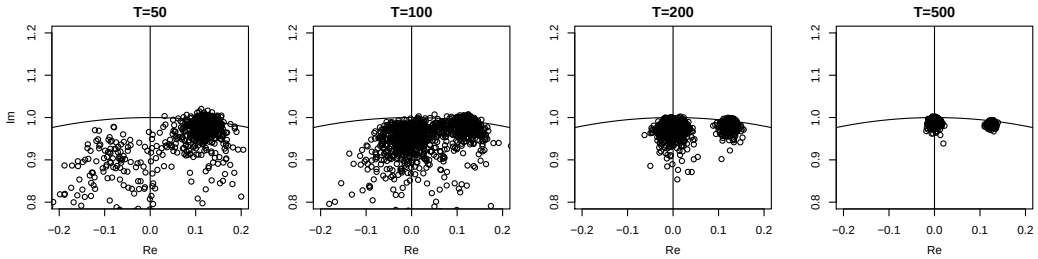


Figure 1. Eigenvalues around $z = i$ of 1000 replications when $\gamma = 0.2$ ($c_{\pi/2} = 1$).

When the VAR dgp without false unit roots and $c_{\pi/2} = 2$ is considered, it is visible that the hit rates of Λ at $\pi/2$ are monotonously increasing in the sample size again. The rates are smaller than those of the likelihood-based tests, however, and also clearly worse than those of Λ at 0 and π, cf. Table 2 again.

Taken together, at frequencies 0 and π which correspond to real-valued unit roots, the use of Λ was advantageous for $T = 50$. It also scored more hits for $T = 100$ and $c_{\pi/2} = 1$. For higher sample sizes the likelihood-based tests clearly dominate Λ at these two frequencies. At $\pi/2$ this superiority of the likelihood-based tests for all sample sizes and both dgps continues. The example also points to a general weakness: if the sample size is low and *false unit roots* are present, it can be difficult for Λ to distinguish them from actual unit roots.

7.2. VARMA Processes

The second setup consists of VARMA data generated by a state space system (A_r, C_r, K_r), $r = 1, 2$, as in (1), where the matrices A_1 and A_2 are constructed as in (2) and are taken to be

$$A_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \end{bmatrix}, \qquad A_2 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & -1 & 0 \end{bmatrix}. \tag{9}$$

These two choices yield the same state space unit root structures as those of the two VAR dgps with $c_{\pi/2} = 1$ and $c_{\pi/2} = 2$ for A_1 and A_2, respectively. The other two system

matrices $K_r \in \mathbb{R}^{(2+2r)\times s}$ and $C_r \in \mathbb{R}^{s\times(2+2r)}$ with $s = 8$ are drawn randomly from a standard normal distribution in each replication and $(\varepsilon_t)_{t\in\mathbb{Z}}$ is multivariate normal white noise with an identity covariance matrix.

Note that these systems are within the VARMA model class such that the dgp is contained in the VAR setting only by increasing the lag length as a function of the sample size. While superiority of the `CVA` approach in such a setting might be expected, this is far from obvious. Moreover, using a long VAR approximation is the industry norm in such situations.

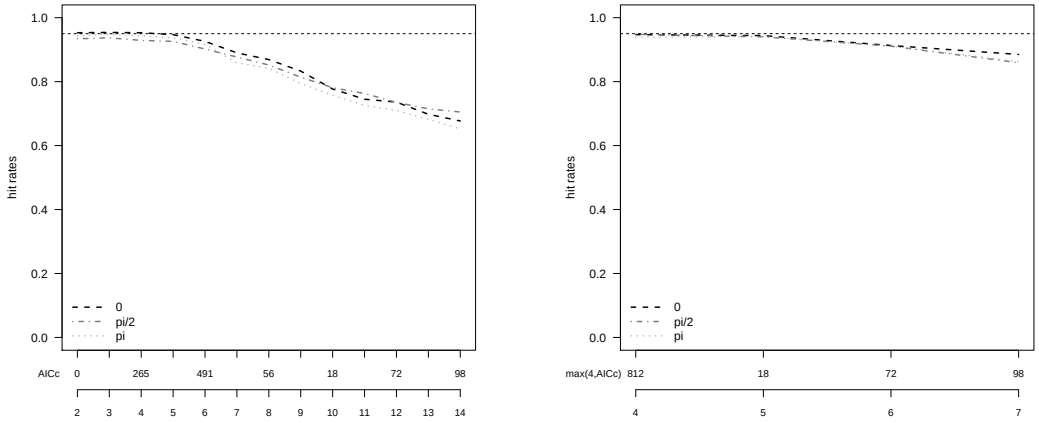
From the hit rates in Table 3 it can be seen that the combination of large s, small T, and a minimal lag length of four render the likelihood-based tests useless at all frequencies with hit rates below ten percent for $T = 50$. Λ in contrast does not suffer from this problem and is already close to 95% for this sample size. Only when $T = 200$ do the likelihood-based tests appear to work, exhibiting hit rates close to 95%.

Table 3. Hit rates for the different tests (VARMA dgp). Twice the maximum (over all entries) Monte Carlo standard error is 0.005.

		0				**π/2**			**π**	
	T	**Λ**	**J**	**Λ**	**JS**	**Q1**	**Q2**	**Q3**	**Λ**	**J**
A_1	50	0.890	0.003	0.906	0.024	0.027	0.032	0.025	0.897	0.008
	100	0.928	0.434	0.944	0.755	0.783	0.783	0.761	0.930	0.440
	200	0.936	0.937	0.923	0.925	0.915	0.916	0.915	0.925	0.924
	500	0.852	0.901	0.853	0.919	0.906	0.904	0.904	0.853	0.894
A_2	50	0.863	0.008	0.785	0.062	0.047	0.063	0.039	0.867	0.006
	100	0.917	0.500	0.880	0.582	0.596	0.596	0.571	0.916	0.518
	200	0.931	0.927	0.882	0.908	0.915	0.913	0.911	0.919	0.922
	500	0.824	0.882	0.786	0.878	0.860	0.859	0.861	0.812	0.865

For all tests alike, however, it is striking that hit rates move away from 95% when $T = 500$. This behavior is most pronounced for Λ, e.g., from $T = 200$ to $T = 500$ its hit rate drops from 93.1% to 82.4% at 0 when A_2 is used. This phenomenon is a consequence of the fact that f and k in the algorithm are chosen data dependent. An inspection of how the hit rates depend on f and k and a comparison with the actually selected $\hat{f}, \hat{k}$ reveals that for $T = 500$ too large values of f and k are chosen too often and leave room for improvement in the hit rates, cf. Figure 2. The figure stresses an important point: The performance of the unit root tests is heavily influenced by the selected lag lengths for all procedures. We tested a number of different information criteria in this respect. `AICc` turned out to be the best criterion overall, but not uniformly. Figure 2 indicates advantages for this example of `BIC` over `AIC` as it on average selects smaller lag lengths, associated here with higher hit rates.

To study the power of the different procedures, the transition dynamics A_r in (9) are multiplied by $\rho \in \{0.8, 0.85, 0.9, 0.95\}$ so that the systems do not contain unit roots at any of the frequencies. Here empirical power is defined as the frequency of choosing zero common trends. This is why for $\rho = 1$, when there are in fact common trends present in our specifications, the empirical power values plotted in Figure 3 are not equal to the actual size we could define as one minus the hit rate: our measure of empirical power in this situation only counts the false test conclusion of zero common trends, but there are of course multiple ways the testing procedure could conclude falsely.



(**a**) Λ,$T = 500$ (**b**) Q_1, $T = 500$

Figure 2. Relationship between hit rates and chosen values of f and k, illustration for the VARMA dgp using A_2. The lower x-axes show f or k, above are the choice frequencies of the selection criteria.

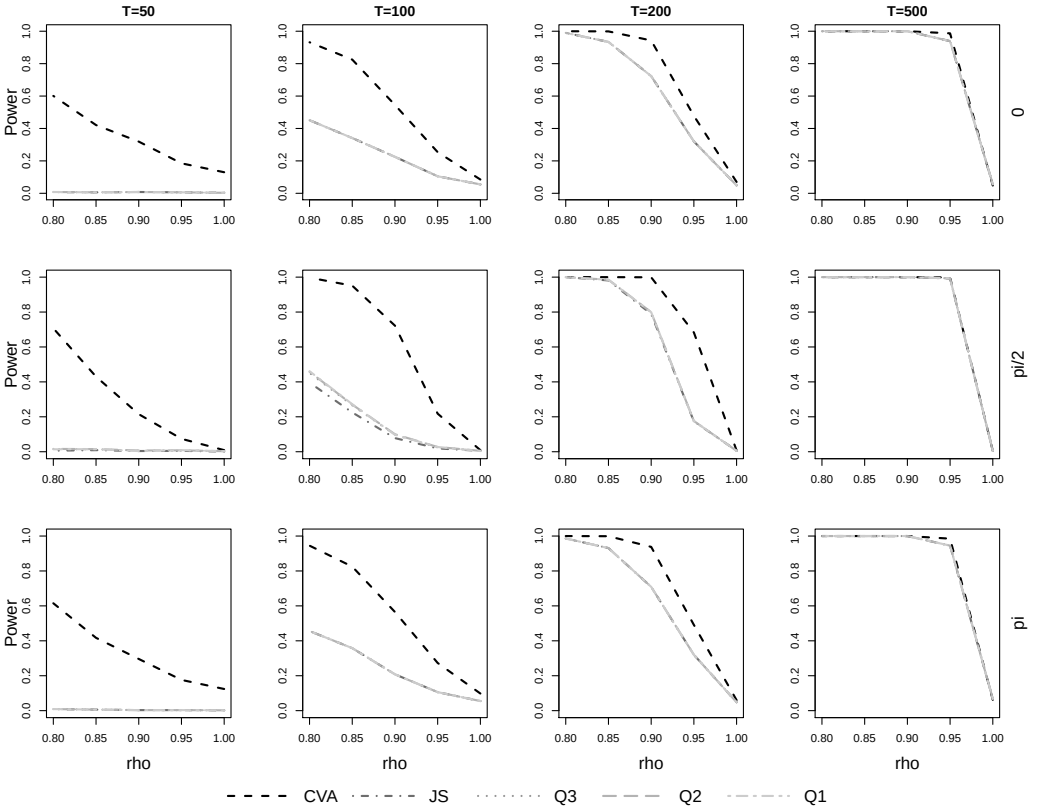


Figure 3. Empirical power of the different test procedures (VARMA dgp with A_2). Twice the Monte Carlo standard error is 0.005.

As expected, rejection of the null hypothesis is easiest when ρ is small and is very difficult when it is close to 1, cf. Figure 3 for the case of A_2.

Further, there are almost no differences among the likelihood-based tests over all combinations of sample size and frequency, only for $T = 100$ is JS significantly worse than the $Q_i, i = 1, 2, 3$ at $\pi/2$. It is also clearly visible at all frequencies that the likelihood-based tests possess no or only very limited power when $T = 50$ and $T = 100$, respectively. Λ, in contrast, is clearly more powerful in these cases. As the sample size increases to $T = 200$, the power of each test improves, still Λ remains the most powerful option. Only for $T = 500$ have the differences almost vanished with small, but significant, advantages for Λ at 0 and π.

The results are the same when A_1 is used and $c_{\pi/2} = 1$ and all of the differences described here are statistically significant.

Next the estimation performance of CVA is evaluated by calculation of the gaps between the true and the estimated cointegrating spaces. At all frequencies these gaps are compared with the GARR procedure of [5] which cycles through frequencies. At $\pi/2$ CVA and GARR are also compared with our implementation of JS and cRRR of [4], whereas it is also compared with the usual Johansen procedure at 0 and π. All estimates are conditional on the true state space unit root structure in the sense that the minimal number of states used is larger or equal to the number of unit roots over all frequencies. Other than imposing a minimum state dimension, the estimation of the order using *SVC* is not influenced. The likelihood-based procedures, on the other hand, take the unit root structure as given, i.e., do not perform CI rank testing for this estimation exercise.

From the results in Table 4 it can be noted first that the likelihood-based procedures show mostly equal mean gaps. Only for $\pi/2$ and $T = 50$ and both dgps does JS possess significantly larger gaps than cRRR and GARR and other differences are not statistically significant. Thus it does not matter in our example whether the iterative procedure is used or not.

Second, CVA is again superior for $T = 50$ where it exhibits mean gaps that are significantly smaller than those of the other estimators at all frequencies. This advantage is turned around for higher sample sizes, though: mean gaps are smaller for the likelihood-based procedures when $T \in \{100, 200, 500\}$ and A_2 is used, if only slightly. When A_1 is used instead, mean gaps do not differ significantly from each other at $\pi/2$ when $T > 50$ and at $0, \pi$ when $T = 100$ and those of CVA are only very modestly worse when $T \in \{200, 500\}$ at $0, \pi$.

Table 4. Mean gaps between estimated and true cointegrating spaces (VARMA dgp). 2*MCse* denotes twice the maximal Monte Carlo standard error for the corresponding row.

			0			**π/2**				**π**		
	T	**2MCse**	**CVA**	**J**	**GARR**	**CVA**	**JS**	**cRRR**	**GARR**	**CVA**	**J**	**GARR**
A_1	50	0.016	0.116	0.189	0.192	0.091	0.147	0.130	0.130	0.111	0.192	0.197
	100	0.004	0.047	0.048	0.048	0.039	0.035	0.035	0.035	0.047	0.046	0.046
	200	0.003	0.023	0.019	0.019	0.019	0.016	0.016	0.016	0.024	0.019	0.019
	500	0.003	0.012	0.007	0.007	0.008	0.008	0.006	0.006	0.011	0.007	0.007
A_2	50	0.016	0.174	0.245	0.242	0.250	0.349	0.331	0.331	0.165	0.231	0.234
	100	0.004	0.072	0.061	0.061	0.098	0.080	0.078	0.078	0.069	0.060	0.060
	200	0.003	0.031	0.026	0.026	0.047	0.036	0.034	0.034	0.032	0.027	0.027
	500	0.003	0.016	0.011	0.010	0.021	0.015	0.013	0.013	0.017	0.011	0.011

Thus, when it comes to estimating the cointegrating spaces, CVA is superior for $T = 50$ and equally good or only slightly worse than the likelihood-based procedures for higher sample sizes. For the systems analyzed, decreasing $c_{\pi/2}$ leads to gaps that are smaller for all methods and these improvements are slightly larger for CVA than for the other estimators.

7.3. Robustness of Unit Root Tests for Daily Data

In this last simulation example we examine the robustness of the proposed procedures with regard to test performance and prediction accuracy with respect to the innovation distribution and the existence of conditional heteroskedasticity of the GARCH-type, as these features are often observed in data of higher frequency, for example in financial applications. While our asymptotic results do not depend on the distribution of the innovations (subject to the assumptions), the assumptions do not include GARCH effects. Nevertheless, the theory in [25,26] suggests that the tests might be robust also in this respect.

We generate a state space system of order $n = 8$ using the matrix $A = [A_{i,j}]_{i,j=1,...,8}$ where $A_{i,i+1} = 1, i = 1, ..., 6, A_{7,1} = 1, A_{8,8} = 0.8$ and $A_{i,j} = 0$ else. This implies that the eigenvalues of this matrix are $\lambda_j = \exp(2\pi ij/7), j = 1, ..., 7, \lambda_8 = 0.8$. Therefore the corresponding process has state space unit root structure

$$((0, (1)), (2\pi/7, (1)), (4\pi/7, (1)), (6\pi/7, (1))).$$

The entries of the matrices C and K are chosen as independent standard normally distributed random variables as before.

A process $(y_t)_{t=1,...,T}$ is generated from filtering an independent identically distributed innovation process $(\varepsilon_t)_{t=-199,...,T+1}$ through the system (A, C, K). The first 200 observations are discarded, the last are used for validation purposes. A total of 1000 replications are generated where in each replication a different system is chosen.

With the generated data three different estimates are obtained: An autoregressive model (called `AR` in the following) is estimated with lag length chosen using `AIC` of maximal lag length equal to $\lfloor\sqrt{T}\rfloor$. Second, an autoregressive model with large lag length (called `ARlong`) is estimated. This estimate is used to hint at the behavior of an autoregression using the lag length equal to a full year. This would correspond to estimating a VECM without rank restrictions, when accounting for yearly differences. The third method consists of the `CVA` estimates, where $f = p = 2\hat{k}_{AIC}$ is chosen. The order is estimated by minimizing `SVC`. However, we correct for orders smaller than $n = 7$ which would limit the possibilities of finding all unit roots.

First, we compare the prediction accuracy for the three methods for two different distributions of the innovations: Beside the standard normal distribution also the student t-distribution with $v = 5$ degrees of freedom (scaled to unit variance) is used. This distribution shows considerably heavier tails than the normal distribution but nevertheless is covered by our assumptions.

Figure 4 provides the results for out-of-sample one day ahead mean absolute prediction error (over all coordinates) for the sample sizes $T = 364$ days (one year), $T = 1092$ (3 years) and $T = 3276$ (nine years). The long AR model is estimated with lag lengths of 8 weeks for the smallest sample size, 10 weeks for the medium sample size and 12 weeks for the largest sample size.

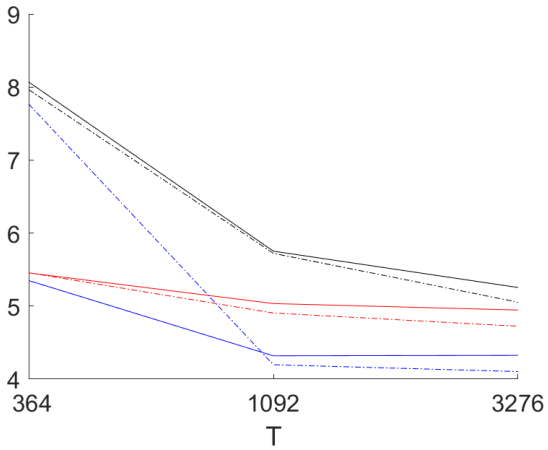


Figure 4. Mean of absolute value of one day ahead prediction error over all four components. CVA (blue), AR (red) and long AR (black). Dash-dot lines refer to the t-distribution.

In the figure the results for the normally distributed innovations are presented as well as the ones for the t-distributed residuals (scaled to unit variance). It can be seen that for the two larger sample sizes the mean absolute error for the residuals for `CVA` is smaller in all cases. For the smallest sample size, by contrast, results are mixed. For `CVA` the results for the heavy tailed distribution in this case are much worse than for the normal

distribution. For the larger sample sizes the differences are small. The maximal standard error of the estimated means over 1000 replications for $T = 1092$ and $T = 3276$ amounts to 0.05. This allows the conclusion that `CVA` performs better for the two larger sample sizes. For $T = 364$ there are no statistically significant differences between the performance of the three methods: `CVA` seems to suffer more from few very large errors (using the root mean square errors the `CVA` results are worse for $T = 364$ in comparison; if one uses the 95% percentiles `CVA` performs best also for the smallest sample size). This results in a standard error over the replications of the mean absolute error for $T = 364$ of 0.18 for normally distributed innovations and 3.4 for t-distributed innovations.
The long AR models are clearly worse than the two other approaches. This happens even if we are still far from using a full year as the lag length.

With regard to the unit root tests we investigate results for the tests of the hypotheses $H_0 : c_m = 1$ versus $H_1 : c_m = 0$ at all frequencies $2\pi m/364, m = 0, ..., 363$. The data generating process features unit roots with $c_m = 1$ at the seven frequencies $2\pi k/7, k = 0, ..., 6$. Therefore the tests should not reject at these frequencies, but should reject at all others.

Consequently we compare the minimum of the non-rejection rates for the seven unit roots (called empirical size below) as well as the maximum of the non-rejection rates for the non-unit root frequencies $\omega_j = 2\pi j/364, j \neq 52k, k = 0, 1, 2, ..., 6$ (called empirical power below).

For the larger sample sizes the empirical size is practically 95% while the empirical power is 100%. For $T = 364$ we obtain an empirical size of 90% for the normal distribution and 91.5% for the t-distribution. The worst empirical power equals 89.3% (normal) and 87.6% (t-distribution). Hence even for one year of data the discrimination properties of the unit root tests are good and we do not observe differences between the normal distribution for the innovations and the heavy tailed t-distribution.

Finally we compare the empirical size and power of the tests for the various unit roots for smaller sample sizes $T \in \{104, 208, 312, 416, 520\}$. For the experiments we consider univariate GARCH models of the form

$$\varepsilon_{t,i} = h_{t,i}\eta_{t,i}, \quad h_{t,i}^2 = 1 + \alpha\varepsilon_{t-1,i}^2 + \beta h_{t-1,i}^2, \quad i = 1, .., 4,$$

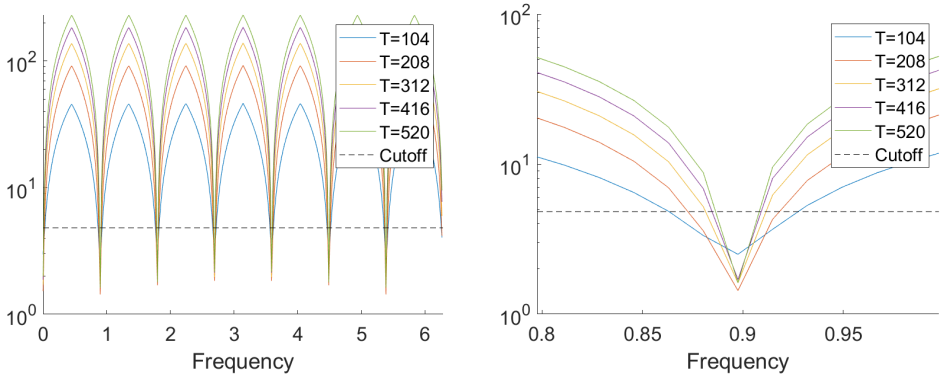
where $(\eta_{t,i})_{t\in\mathbb{Z}}$ is independent and identically standard normally distributed. $\alpha, \beta \geq 0$ are reals. It follows that the component processes $(\varepsilon_{t,i})_{t\in\mathbb{Z}}$ show conditional heteroskedasticity, the persistence of which is governed by $\alpha + \beta$. Here $0 < \alpha + \beta < 1$ implies stationarity while $\alpha + \beta = 1$ implies persistent conditional heteroskedasticity usually termed I-GARCH. We include five different processes for the innovations:

1. norm: $\alpha = \beta = 0$, no GARCH effects
2. G1: $\alpha = 0.8, \beta = 0.1$
3. IG1: $\alpha = 0.8, \beta = 0.2$
4. IG2: $\alpha = 0.5, \beta = 0.5$
5. IG3: $\alpha = 0.2, \beta = 0.8$

For the five different sample sizes 1000 replications of the estimates using the `CVA` algorithm are obtained. For each estimate we calculate the test statistic for testing $H_0 : c_m = 1$ versus $H_0 : c_m = 0$ for $m = 0, ..., 363$ corresponding to the unit roots $z_m = \exp(2\pi im/364)$. This set of unit roots contains all seven unit roots $\exp(2\pi ik/7), k = 0, ..., 6$.

Figure 5 provides the mean over the 1000 replications of the test statistics $\Lambda(1)$ for $z_j, j = 0, ..., 363$ and the five sample sizes. It can be seen that the test $\Lambda(1)$ is able to pinpoint the seven unit roots present in the data generating process fairly accurately even for sample size $T = 104$. The zoom on the region around the unit root frequency $2\pi/7$ shows that the

mean value is larger than the cutoff value of the test (the dashed horizontal line) for the adjacent frequency $2\pi\frac{53}{364}$ already for $T = 312$.



(**a**) Mean of unit root test statistics. (**b**) Zoom of mean unit root tests.

Figure 5. Results of the unit root tests for all seasonal unit roots jointly.

Table 5 lists the minimum of the achieved percentages of non-rejections of the test statistic for the seven unit root frequencies as well as the maximum over all non-unit root frequencies. It can be seen that for all GARCH models for $T = 312$ the test rejects unit roots at all non unit root frequencies every time, while the empirical size is close to the nominal 5%. For small sample sizes the tests are slightly undersized while for $T = 208$ a slight oversizing is observed. The two larger sample sizes are omitted as the tests perform perfectly there.

Table 5. Percentage of accept (minimum for all unit root frequencies) and reject (maximum for non unit root frequencies) of $\Lambda(1)$ test statistic.

	Unit Root Frequencies					**Non Unit Root Frequencies**				
T	**norm**	**G1**	**IG1**	**IG2**	**IG3**	**norm**	**G1**	**IG1**	**IG2**	**IG3**
104	0.94	0.89	0.87	0.88	0.87	0.87	0.82	0.79	0.82	0.79
208	0.98	0.96	0.95	0.94	0.96	0.78	0.75	0.72	0.72	0.69
312	0.97	0.96	0.96	0.95	0.95	0.00	0.00	0.00	0.00	0.00

It follows from the examples presented in this subsection that the test is robust also in small samples with respect to heavy tailed distributions of the innovations (subject to the assumptions). Furthermore also a remarkable robustness with respect to GARCH-type conditional heteroskedasticity is observed.

8. Application

In this section we apply `CVA` to the modeling of electricity consumption using a data set from [36]. The dataset contains hourly consumption data (in megawatts) from a number of US regions, scraped from the webpage of PJM Interconnection LLC, a regional transmission organization. The number of regions have changed over time, thus the data set contains many missing values. It also contains data aggregated into regions called east and west, which are not used subsequently.

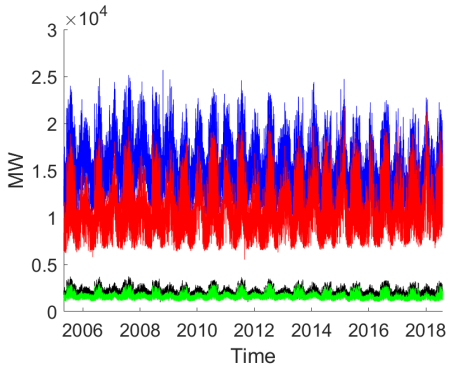
In order to avoid problems with missing values, we restrict the analysis to four regions, for which data over the same time period is available: American Electric Power (AEP; in the following printed in blue), the Dayton Power and Light Company (DAYTON; black), Dominion Virginia Power (DOM; red) and Duquesne Light Co. (DUQ; green). We use data from 1 May 2005 until 31 July 2018. In this period only 3 data points are missing for the four regions and their imputation is handled by interpolation of the corresponding previous values. One observation in this sample is an obvious outlier which is corrected for analogously.

The data is split into an estimation sample covering observations up to the end of 2016 (102,291 observations on 4263 days) and a validation sample containing data in 2017 and 2018 (13,845 observations on 577 days). Data is equally sampled, but contains two hour segments when switching from winter to summer time or back. Table 6 contains some summary statistics.

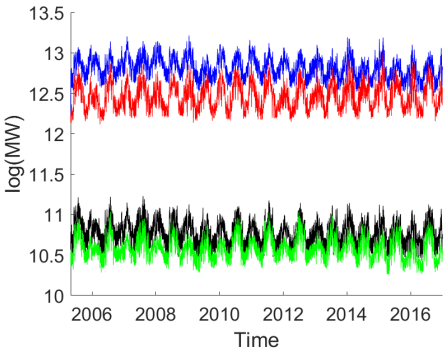
Table 6. Summary of data sets.

Region	Daily Obs. (4263 est., 577 val.)					Hourly Obs. (102,291 est., 13,845 val.)			
	Mean	Mean(log)	Std.(log)	AIC	BIC	Mean(log)	Std.(log)	AIC	BIC
AEP	371,844	12.82	0.127	43	12	9.63	0.168	782	532
DAYTON	48,897	10.79	0.144	43	3	7.60	0.193	772	531
DOM	262,727	12.47	0.158	17	3	9.28	0.215	795	554
DUQ	39,837	10.58	0.130	23	7	7.40	0.177	800	529

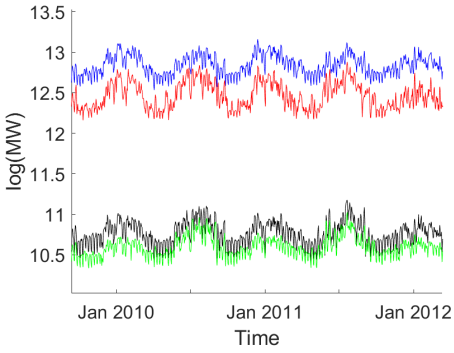
Figure 6 provides an overview of the data: Panel (a) shows the full data on an hourly basis, while (b) presents aggregation to daily frequency. Panel (c) zooms in on a two year stretch of daily consumption. Panel (d) finally provides hourly data for the first month in the validation data. The figures clearly document strong daily, weekly and yearly patterns. From these figures it appears that these seasonal fluctuations are somewhat regular with changes throughout time. It is hence not clear whether a fixed seasonal pattern is appropriate. Also note that the sampling frequency is on an hourly basis such that a year roughly covers 8760 observations.



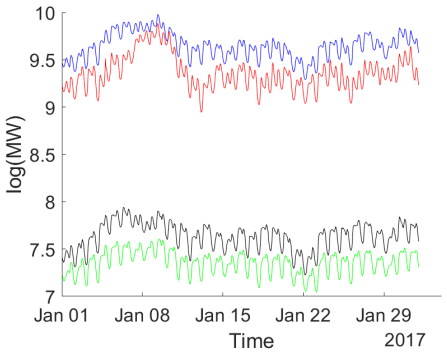
(**a**) Full hourly data



(**b**) Log of daily consumption on estimation set



(**c**) Log of daily consumption from 2010 to 2012



(**d**) Log of hourly consumption on first month of validation set

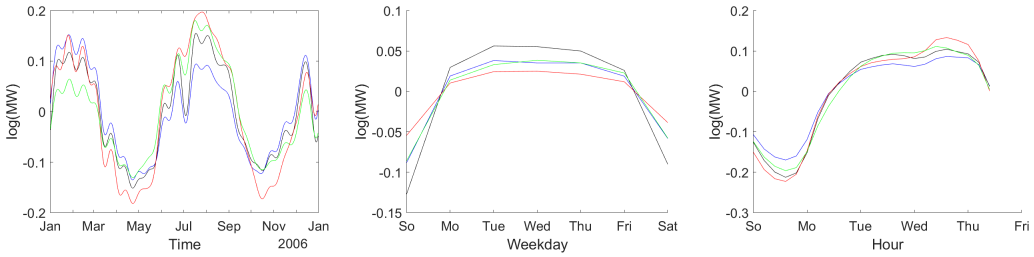
Figure 6. Electricity consumption data.

In the following we estimate (on the estimation part) and compare (on the validation part) a number of different models, first for the full hourly data set and afterwards for

the aggregated daily data. As a benchmark we will use univariate AR models including deterministic seasonal patterns for daily, weekly and yearly variations. Subsequently we estimate models using `CVA` including different sets of such seasonal patterns.

First in the analysis using dummy variables fixed periodic patterns have been estimated. We model the natural logarithm of consumption (to reduce problems due to heteroskedasticity) and include dummies for weekdays, hours and sine and cosine terms corresponding to the first 20 Fourier frequencies with respect to annual periodicity. The corresponding results can be viewed in Figure 7. It is obvious that there is quite some periodic variation. Also the four data sets show very similar patterns as expected.

After the extraction of these deterministic terms the next step is univariate autoregressive (AR) modeling. Figure 8 shows the BIC values of AR models of lag lengths zero to 800 for the four series as well as the BIC of a multivariate AR model for the same number of lags. The chosen values are given in Table 6.



(**a**) Yearly fluctuation (**b**) Weekly fluctuation (**c**) Daily fluctuation

Figure 7. Periodic patterns from dummy variables.

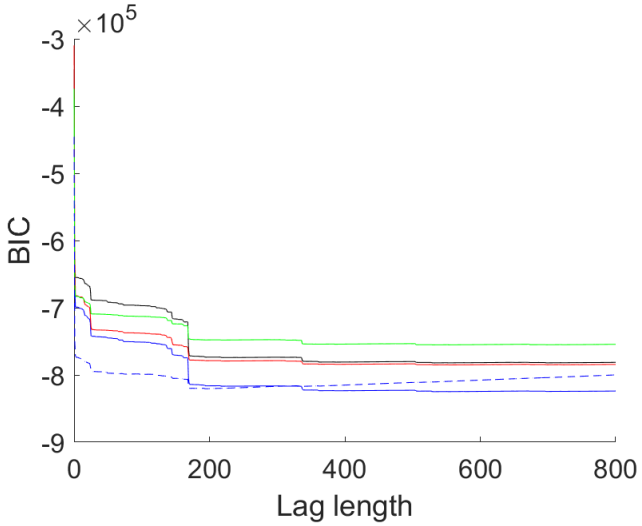


Figure 8. BIC values for univariate models and multivariate model (dashed line; divided by four to fit).

The BIC curve is extremely flat for the univariate models. Noticeable drops in BIC occur around lag 24 (one day), 144 (six days), 168 (one week), 336 (two weeks), 504 (three weeks). BIC selects large lag lengths from 529 (DUQ) up to 554 (DOM). AIC selects lag lengths close to the maximum allowed with a minimum at 772 lags. The BIC pattern of the multivariate model differs in that the two drops at two and three weeks are missing. Instead, the optimal BIC value is obtained at lag 194, well below the optimal lag lengths in the univariate cases. AIC here opts for lag length 531, just over 22 days.

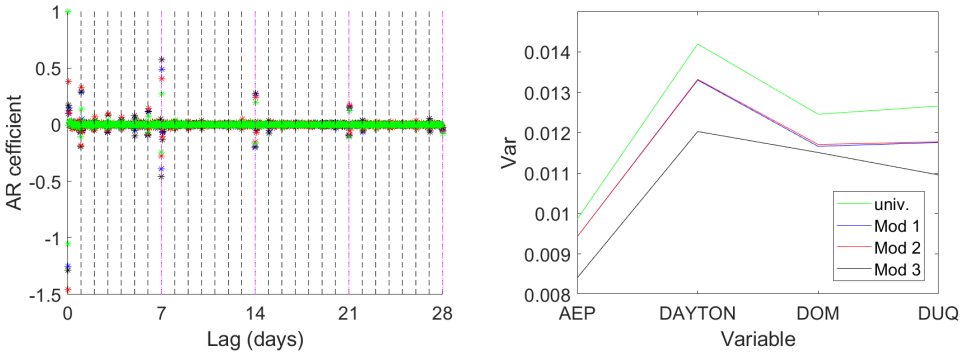
Subsequently `CVA` is applied with $f = \hat{k}_{BIC}$, $p = \hat{k}_{AIC}$ as estimated for the multivariate model. This differs from the usual recommendation of $f = p = 2\hat{k}_{AIC}$ in order to avoid numerical problems with huge matrices. The order is chosen according to `SVC`, resulting in $\hat{n} = 240$. The corresponding model is termed `Mod 1` in the following. Note that this

configuration of $f, \hat{n}$ does not fulfill the requirements of our asymptotic theory. The bound $f \geq n$ ensures that the matrix $\mathcal{O}_f$ has full column rank. Generically this will be the case for $fs \geq n$ leading to a less restrictive assumption. In practice too low values of f will be detected by $\hat{n}$ estimated close to the maximum, which is not the case here.

As a second model we only use weekday dummies but neglect the other deterministics. Again AIC ($\hat{k}_{AIC} = 531$) and BIC ($\hat{k}_{BIC} = 195$) are used to determine the optimal lag length in the multivariate AR model. The corresponding `CVA` estimated model uses $\hat{n} = 245$ according to `SVC`, resulting in `Mod 2`.

The third model uses only a constant as deterministic term. Again similar AIC (555) and BIC (195) values are selected. A state space model, `Mod 3`, using `CVA` is estimated with $\hat{n} = 209$.

Figure 9 provides information on the results. Panel (a) shows the coefficients of the univariate AR models. It can be seen that lags around one day and one to three weeks play the biggest role for all four datasets. Panel (b) shows that the multivariate models lead to better one step ahead predictions in terms of the root mean square error (RMSE). `Mod 1` and `Mod 2` show practically equivalent out of sample prediction error for all four data sets, while `Mod 3` delivers the best out of sample fit for all four regions.



(**a**) AR coefficients (**b**) RMSE on validation data set

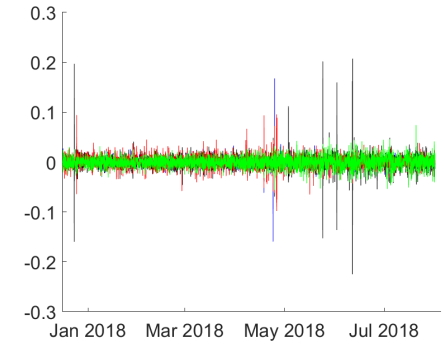
Figure 9. Results for the hourly datasets.

In particular in financial applications data of high sampling frequency shows persistent behaviour, also in terms of conditional heteroskedasticity, as well as heavy tailed distributions of the innovations. For our data sets Figure 10 below provides some information in this respect for the residuals according to `Mod 3`. Panel (a) provides a plot of the residuals in the year 2018 (contained in the validation period). It can be seen that large deviations occur occasionally, while else residuals vary in a tight band around 0. The kernel density estimates for the normalized (to unit variance) residuals on the full validation data set in panel (b) show the typical heavy tailed distributions. Panel (c) contains an ACF plot for the four regions again calculated using the full validation sample. It demonstrates that the model successfully eliminates all autocorrelations with only a few ACF values occurring outside the confidence interval. Panel (d) provides the ACF plot for the squared innovations to examine GARCH-type effects. While GARCH-effects are clearly visible, the ACF drops to zero fast with occasional positive values (except maybe for the Duquesne data).

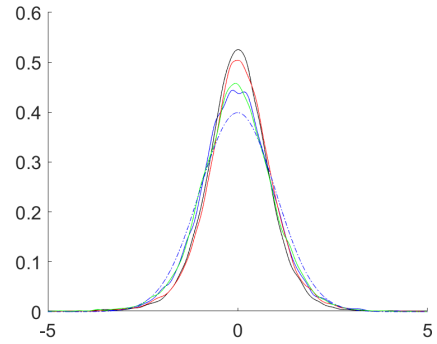
Applying the eigenvalue based test $\Lambda(1)$ for $c = 1$ and all Fourier frequencies $\omega_j = 2\pi j/(365 * 24)$ we find that for `Mod 2` and `Mod 3` the largest p-value is obtained for ω_{365} corresponding to a period length of one day with 0.0187 for `Mod 2` (test statistic 6.6) and 0.02 for `Mod 3` (with a statistic of 6.5). This implies that the unit root at frequency ω_{365} is not rejected for a significance level of 1%, but is rejected for 5%. All other unit roots are rejected at every usual significance level. For `Mod 1` the test statistic for ω_{365} equals 41.2 corresponding to a p-value of practically 0. This implies that on top of a deterministic daily pattern the series show strong persistence at the daily period. Excluding the hourly dummies pulls the roots closest to ω_{365} closer to the unit circle resulting in insignificant

unit root tests and improves the one step ahead forecasts. Including the dummies weakens the evidence of a unit root while leading to worse predictions.

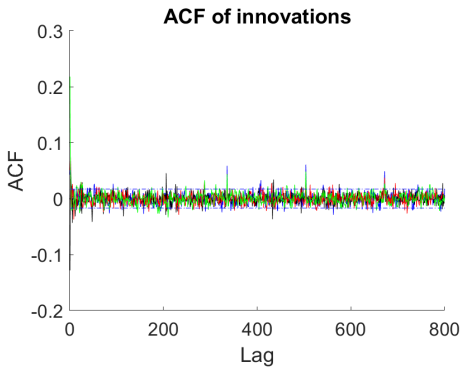
The analysis is repeated with data aggregated to daily sampling frequency. The aggregation reduces the required lag lengths, as is visible from Table 6 in the univariate case, and hence we use `CVA` with the recommended $f = p = 2\hat{k}_{AIC}$. Beside the univariate models, in this case also a naive model of predicting the consumption for today as yesterday's consumption is used. Three multivariate models are estimated: `Mod 1` contains weekday dummies and sine and cosine terms for the first twenty Fourier frequencies corresponding to a period of one year. `Mod 2` only contains the weekday dummies, while `Mod 3` only uses the constant. Figure 11 provides the out-of-sample RMSE for one day ahead predictions (panel (a)) and seven days ahead predictions (panel (b)).



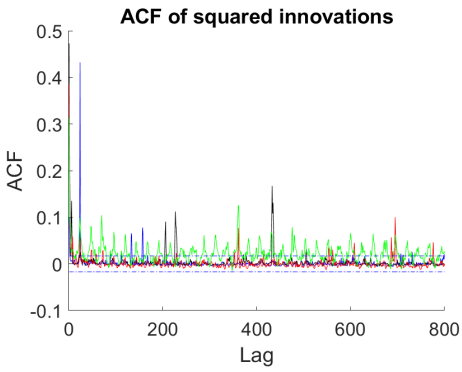
(**a**) Residuals for 2018.



(**b**) Kernel density estimate for normalized residuals. Blue dashed line refers to standard normal distribution.

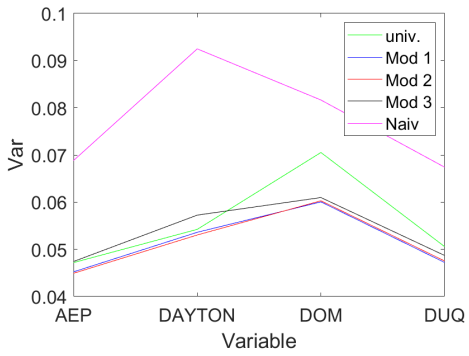


(**c**) ACF of residuals.

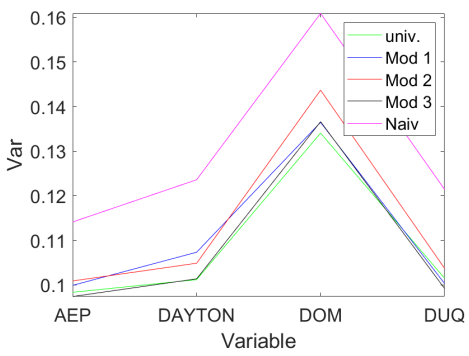


(**d**) ACF of squared residuals.

Figure 10. Residual analysis.



(**a**) RMSE of one day ahead predictions

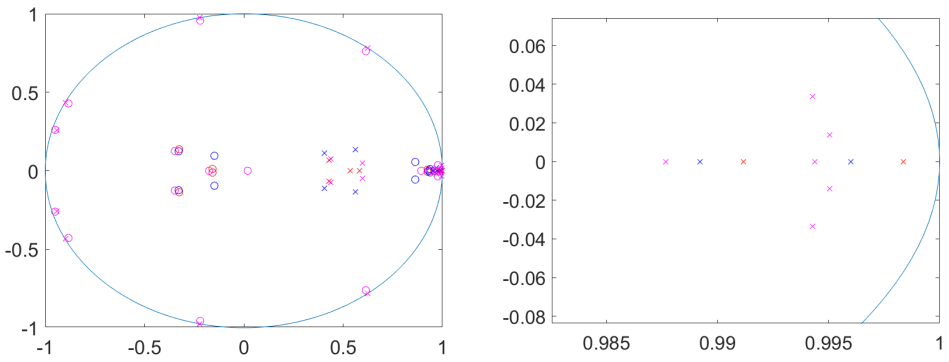


(**b**) RMSE of seven day ahead predictions

Figure 11. Results for the hourly datasets.

It can be seen that both `Mod 1` and `Mod 2` beat the univariate AR models in terms of one step ahead prediction error, while `Mod 3` performs better for seven days ahead prediction. `Mod 1` performs on par with `Mod 2` for one step ahead prediction but performs better in predicting seven steps ahead. In Figure 12 poles and zeros for the three estimated state space models are plotted. Here the poles (marked with 'x') are the eigenvalues of the matrix A. These are the inverses of the determinantal roots of the autoregressive matrix polynomial in the equivalent VARMA representation. The zeros are the inverses of zeros of the determinant of the MA polynomial. We can see that for `Mod 3` with only a constant, poles close to $2\pi j/7, j = 1, ..., 6$ arise to capture the weekly pattern. The other two models only show one pole close to the unit circle, a real pole of almost $z = 1$. The pole corresponding to `Mod 1` is closer to the unit circle than the one for `Mod 2` (see (b)).

For `Mod 3` we obtain p-values for the tests of three complex unit roots of 0.05 ($\omega = 2\pi/7$), 0.165 ($4\pi/7$) and 0.01 ($6\pi/7$), which are hence all not statistically significant for significance level $\alpha = 0.01$. The corresponding test for $z = 1$ shows a p-value of 0.004. This provides evidence against the null hypothesis of the root being present. For `Mod 1` the p-value for the test of $z = 1$ is 0.28 and hence we cannot reject the null. `Mod 2` provides a p-value of 0.023 and hence weak evidence for the presence of the unit root. This can be seen from the distance of the nearest pole from the point $z = 1$ in Figure 12.



(**a**) Poles and zeros of the three models. (**b**) Zoom around $z = 1$.

Figure 12. Poles (x) and zeros (o) of the transfer functions corresponding to the three models: `Mod 1` (red), `Mod 2` (blue), `Mod 3` (magenta).

Jointly this indicates that the location and strength of persistence due to the estimated roots is influenced by the presence of deterministic terms: if the deterministic terms are not included in the model, the cyclical patterns are generated by poles situated close to the unit circle.

The decision whether on top of the deterministic seasonality unit roots exist, is not easy in all cases: for the daily data the locations of the poles indicate that deterministic seasonality is enough to capture weekly fluctuations while a unit root at $z = 1$ appears to be needed to capture yearly variations. For hourly data there is evidence that the daily cycle is best captured with a unit root at frequency ω_{365}. This leads to the best predictive fit. Finally note that temporal aggregation from hourly data to daily data implies that the frequency ω_{365} for hourly data aliases to the frequency $\omega = 0$ in the daily data. Therefore the higher evidence of a unit root at $z = 1$ found in daily data might be a consequence of the unit root at frequency ω_{365} found for hourly data, compare [37].

The system matrix estimates as well as the evidence in support of unit roots at ω_{365} for hourly data and $z = 1$ for daily data that we obtain from the `CVA` modeling can be taken as starting points in subsequent quasi maximum likelihood estimation.

9. Conclusions

In this paper the asymptotic properties of `CVA` estimators for seasonally integrated unit root processes are investigated. The main results can be summarized as follows:

- `CVA` provides consistent estimators for long-run and short-run dynamics without knowledge of the location and number of unit roots. Hence the algorithm is robust with respect to the presence of trending components at frequency zero as well as at the other seasonal unit root frequencies.
- The singular values calculated in the RRR step reveal information on the total number of unit roots. The distance of the singular values to one can be used to construct a consistent estimator of this quantity.
- The eigenvalues of $\hat{A}$ can be used in order to test for the number of common trends. Under the null hypothesis these tests are asymptotically equivalent to the corresponding tests using the true state, making the derivation of asymptotic results and the simulation of the test distribution simple.
- An analogous statement holds for the Johansen trace test in the $I(1)$ case and analogous tests in the $MFI(1)$ case calculated on the basis of the estimated state in the restrictive setting of $n \leq s$. Under the null hypothesis these tests reject and accept asymptotically jointly with the corresponding tests calculated using the true state.
- From the simulation exercises we conclude that `CVA` performs best when the dgp is of the more general VARMA type, the process dimension is moderate to large and the sample size is small. Then it is superior to the likelihood-based procedures based on VAR approximations in terms of the estimation performance and the size and power of Λ, the test developed from `CVA`. For higher sample sizes the likelihood-based procedures are clearly superior when it comes to the size of the corresponding tests, whereas Λ remains the best test choice in terms of empirical power. The estimation performance is about equal for all procedures when the sample size is high with slight advantages for the likelihood-based procedures.
- The simulations also demonstrate that the unit root test results are robust with respect to the distribution of the innovation sequence as well as some forms of conditional heteroskedasticity of the GARCH-type.

Because of the promising performance of `CVA` and in particular its robustness it can be recommended as a simple way to extract information on the number of common trends from the estimated matrix of transition dynamics. This information can be used in order to reduce the uncertainty in a subsequent likelihood ratio analysis where quasi maximum likelihood estimates can be obtained starting from the `CVA` estimates. Since the `CVA` estimates can be obtained for a range of orders numerically fast they are seen as a valuable starting point for the empirical modeling of time series potentially including seasonal cointegration. Moreover they can also be used in situations where the number of seasons is large or even unclear as in hourly data sets as demonstrated in the case study.

Author Contributions: Conceptualization, D.B. and R.B.; methodology, D.B.; software, R.B.; formal analysis, D.B. and R.B.; writing—original draft preparation, D.B. and R.B.; writing—review and editing, D.B. and R.B.; visualization, D.B. and R.B.; supervision, D.B. Both authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation—Projektnummer 276051388) which is gratefully acknowledged. We acknowledge support for the publication costs by the Deutsche Forschungsgemeinschaft and the Open Access Publication Fund of Bielefeld University.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Supporting Material

Appendix A.1. Complex Valued Canonical Form

Additionally to the real valued canonical form (2) we will also use the corresponding complex valued representation obtained by transforming each block corresponding to unit root $z_j = \cos(\omega_j) + i\sin(\omega_j)$ with the transformation matrix

$$\mathcal{T}_j = \begin{bmatrix} I_{c_j} & iI_{c_j} \\ I_{c_j} & -iI_{c_j} \end{bmatrix}$$

leading to the triple of system matrices in the j-th block as:

$$\mathcal{A}_{j,\mathbb{C}} = \begin{bmatrix} \overline{z_j} I_{c_j} & 0 \\ 0 & z_j I_{c_j} \end{bmatrix}, \quad \mathcal{K}_{j,\mathbb{C}} = \begin{bmatrix} K_{j,\mathbb{C}} \\ \overline{K_{j,\mathbb{C}}} \end{bmatrix}, \quad \mathcal{C}_{j,\mathbb{C}} = \begin{bmatrix} C_{j,\mathbb{C}}/2 & \overline{C_{j,\mathbb{C}}}/2 \end{bmatrix},$$

such that

$$x_{t+1,j,\mathbb{C}} = \overline{z_j} x_{t,j,\mathbb{C}} + K_{j,\mathbb{C}}\varepsilon_t, \quad x_{t,j} = \mathcal{T}_j^{-1} \begin{bmatrix} x_{t,j,\mathbb{C}} \\ \overline{x_{t,j,\mathbb{C}}} \end{bmatrix}.$$

Lemma A1. *Let $x_t = [x'_{t,0}, x'_{t,1}, \ldots, x'_{t,S/2}, x'_{t,\bullet}]'$ where $x_{t,j}$ is generated according to $x_{t+1,j} = A_j x_{t,j} + K_j \varepsilon_t, t \in \mathbb{N}$ with A_j as in (2) and $K_j = [K'_{j,R}, K'_{j,I}]' \in \mathbb{R}^{\delta_j c_j \times s}$ using iid white noise process $(\varepsilon_t)_{t\in\mathbb{N}}$ where $x_{0,j}$ is deterministic. Further let $(x_{t,\bullet})_{t\in\mathbb{N}}$ denote the stationary solution to the equation $x_{t+1,\bullet} = A_\bullet x_{t,\bullet} + K_\bullet \varepsilon_t$ such that $M_\bullet = \mathbb{E} x_{t,\bullet} x'_{t,\bullet} > 0$.*
(I) Then using $Q_T = \sqrt{(\log\log T)/T}$ for $u_t = \sum_{i=0}^{q} \varphi_i \varepsilon_{t+i}$ for arbitrary $q \in \mathbb{N}, q < \infty$, and coefficients $\varphi_i, i = 0, ..., q$ we have

$$\begin{array}{rcl} \langle x_{t,\bullet}, u_t \rangle = O(Q_T) & , & \langle u_{t-j}, u_t \rangle - \mathbb{E} u_{t-j} u'_t = O(Q_T), \\ \langle x_{t,j}, x_{t,\bullet} \rangle = O(\log T) & , & \langle x_{t,j}, u_t \rangle = O(\log T) \\ \langle x_{t,j}, x_{t,k} \rangle / T = O(\log\log T) & , & j,k = 0, ..., S/2. \end{array}$$

If $(\varepsilon_t)_{t\in\mathbb{Z}}$ only fulfills Assumptions 1 then the order bounds hold in probability rather than almost surely.
(II) Furthermore for $0 < j, k < S/2$

$$\begin{array}{rcl} \langle x_{t,j,\mathbb{C}}, \varepsilon_t \rangle & \xrightarrow{d} & \frac{1}{2}\int_0^1 W_j dB'_{j,\mathbb{C}} =: M_j, \\ \langle x_{t,j,\mathbb{C}}, x_{t,k,\mathbb{C}} \rangle / T & \xrightarrow{d} & \begin{cases} \frac{1}{2}\int_0^1 W_j W'_j := N_j & , \quad j = k, \\ 0 & , \quad j \neq k \end{cases} \\ \langle x_{t,j}, \varepsilon_t \rangle & \xrightarrow{d} & \begin{bmatrix} \frac{1}{2}\int_0^1 (W_{j,R} dB'_{j,R} + W_{j,I} dB'_{j,I}) \\ \frac{1}{2}\int_0^1 (W_{j,I} dB'_{j,R} - W_{j,R} dB'_{j,I}) \end{bmatrix}, \\ \langle x_{t,k}, x_{t,j} \rangle / T & \xrightarrow{d} & \begin{cases} \frac{1}{2}\begin{bmatrix} \int_0^1 (W_{k,R} W'_{k,R} + W_{k,I} W'_{k,I}) & \int_0^1 (W_{k,R} W'_{k,I} - W_{k,I} W'_{k,R}) \\ -\int_0^1 (W_{k,R} W'_{k,I} - W_{k,I} W'_{k,R}) & \int_0^1 (W_{k,R} W'_{k,R} + W_{k,I} W'_{k,I}) \end{bmatrix} & , \quad j = k \\ 0 & , \quad j \neq k \end{cases} \end{array}$$

where $W_j = W_{j,R} + iW_{j,I} = K_{j,\mathbb{C}} B_{j,\mathbb{C}}, K_{j,\mathbb{C}} = K_{j,R} + iK_{j,I}, B_{j,\mathbb{C}} = B_{j,R} + iB_{j,I}$ and $B_{j,R}, B_{j,I}$ are two independent Brownian motions with covariance matrix Ω. For $j = 0$ and $j = S/2$ the results hold analogously:

$$\begin{array}{rcl} \langle x_{t,0}, \varepsilon_t \rangle \xrightarrow{d} \int_0^1 W_{0,R} dW'_{0,R} & , & \langle x_{t,0}, x_{t,0} \rangle / T \xrightarrow{d} \int_0^1 W_{0,R} W'_{0,R}, \\ \langle x_{t,S/2}, \varepsilon_t \rangle \xrightarrow{d} \int_0^1 W_{S/2,R} dW'_{S/2,R} & , & \langle x_{t,S/2}, x_{t,S/2} \rangle / T \xrightarrow{d} \int_0^1 W_{S/2,R} W'_{S/2,R}. \end{array}$$

Proof. Most evaluations in (I) are standard, see for example Lemma 4 in [38].
(II) follows from the results in Section 4 of [2] for the complex valued representations or [39] for the corresponding real case. □

Appendix A.2. Perturbation of Eigendecompositions

Lemma A2 (Rayleigh-Schrödinger expansion)**.** *Let $\hat{A}_t = A - \delta A_t$ where $\|\delta A_t\| \to 0$ and where $A = U\Lambda U^{-1} \in \mathbb{R}^{n\times n}, \Lambda = diag(\lambda_1 I_{c_1}, ..., \lambda_J I_{c_J}), \sum_{j=1}^{J} c_j = n$ is diagonalizable. $U = [U_1, ..., U_J] \in \mathbb{C}^{n\times n}$ is a nonsingular matrix such that for $U^{-1} = [V_1, ..., V_J]'$ we have $V_j' U_j = I_{c_j}$.*

Then for each circle $B(\lambda_j, \delta)$ around λ_j not containing any other eigenvalue of A there exist from some t onwards

- *c_j eigenvalues of $\hat{A}_t$ in the circle $B(\lambda_j, \delta)$ around λ_j*
- *a basis $\hat{U}_{t,j}$ for the space spanned by the eigenspaces to these c_j eigenvalues such that $V_j' \hat{U}_{t,j} = I_{c_j}$,*
- *a sequence of matrices $\hat{B}_{t,j} = V_j' \hat{A}_t \hat{U}_{t,j} \in \mathbb{C}^{c_j \times c_j}$.*

Then $\hat{U}_{t,j} = \sum_{k=0}^{\infty} Z_k$, $\hat{B}_{t,j} = \sum_{k=0}^{\infty} C_k$ where

$$Z_0 = U_j \quad , \quad C_0 = \lambda_j I_{c_j},$$
$$Z_k = \Sigma(\delta A_t Z_{k-1} + \sum_{i=1}^{k-1} Z_{k-i} C_i) \quad , \quad C_k = -V_j' \delta A_t Z_{k-1}.$$

Here $\Sigma = U(\Lambda - I_n \lambda_j)^+ U^{-1}$ where $diag(s_1, ..., s_n)^+ = diag(s_1^+, ..., s_n^+)$ and $x^+ = 1/x, x \neq 0$ and zero else, that is $(\Lambda - I_n\lambda_j)^+$ denotes a quasi-inverse.
Furthermore for $\rho = \|\delta A_t\| < 1$ we obtain: $\|C_k\| \leq \mu_C \rho^k, \|Z_k\| \leq \mu_Z \rho^k, \qquad k \geq 0.$

The results follow directly from Section 2.9 of [23], see in particular Proposition 2.9.1 and the discussion below this proposition. Further note that the results hold for each root separately and hence the restriction $\ell_j = 1$ needs to hold only for the investigated root for the results to apply. Finally note that a second order approximation $\hat{U}_{t,j} = Z_0 + Z_1 + Z_2$ and $\hat{B}_{t,j} = C_0 + C_1 + C_2$ is accurate to the order $o(\|\delta A_t\|^2)$.

Appendix A.3. Random Transformation of Systems

Lemma A3. *Let the assumptions of Theorem 1 hold and use the same notation as given there. Let $(\tilde{\mathcal{A}}, \tilde{\mathcal{C}}, \tilde{\mathcal{K}})$ denote a sequence of systems converging a.s. to $(\mathcal{A}, \mathcal{C}, \mathcal{K})$ such that $(\tilde{\mathcal{A}} - \mathcal{A})D_x^{-1} = O((\log T)^a)$, $\sqrt{T}(\tilde{\mathcal{K}} - \mathcal{K}) = O((\log T)^a)$, $(\tilde{\mathcal{C}} - \mathcal{C})D_x^{-1} = O((\log T)^a)$ and let $\mathcal{A}_0 = S_0 \mathcal{A} S_0^{-1} = diag(\mathcal{A}_{0,11}, \mathcal{A}_{0,22})$, $\mathcal{K}_0 = S_0\mathcal{K}$, $\mathcal{C}_0 = \mathcal{C}S_0^{-1}$. Further let*

$$S_T = \begin{bmatrix} S_{T,11} & S_{T,12} \\ 0 & S_{T,22} \end{bmatrix} \to S_0$$

such that $(S_T - S_0)D_x^{-1} = O((\log T)^a)$. Let $\Delta S = (S_T - S_0)D_x^{-1}, \Delta\mathcal{A} = (\tilde{\mathcal{A}} - \mathcal{A})D_x^{-1}$ and denote the sequence of transformed systems as $(\hat{\mathcal{A}}, \hat{\mathcal{C}}, \hat{\mathcal{K}}) = (S_T \tilde{\mathcal{A}} S_T^{-1}, \tilde{\mathcal{C}} S_T^{-1}, S_T \tilde{\mathcal{K}})$. Let the block entries of S_0 be denoted as S_{ij} and the blocks of ΔS be denoted as ΔS_{ij}. Then:

$$\begin{aligned}
T(\hat{\mathcal{A}}_{11} - \mathcal{A}_{0,11}) &= (\Delta S_{11}\mathcal{A}_{11} - \mathcal{A}_{0,11}\Delta S_{11} + S_{11}\Delta\mathcal{A}_{11} + S_{12}\Delta\mathcal{A}_{21})S_{11}^{-1} + o(1), \\
\sqrt{T}(\hat{\mathcal{A}}_{12} - \mathcal{A}_{0,12}) &= (S_{11}\Delta\mathcal{A}_{12} + S_{12}\Delta\mathcal{A}_{22})S_{22}^{-1} + \Delta S_{12}S_{22}^{-1}\mathcal{A}_{0,22} - \mathcal{A}_{0,11}\Delta S_{12}S_{22}^{-1} + o(1), \\
T(\hat{\mathcal{A}}_{21} - \mathcal{A}_{0,21}) &= S_{22}\Delta\mathcal{A}_{21}S_{11}^{-1} + o(1), \\
\sqrt{T}(\hat{\mathcal{A}}_{22} - \mathcal{A}_{0,22}) &= \Delta S_{22}S_{22}^{-1}\mathcal{A}_{0,22} + S_{22}\Delta\mathcal{A}_{22}S_{22}^{-1} - \mathcal{A}_{0,22}\Delta S_{22}S_{22}^{-1} + o(1), \\
\sqrt{T}(\hat{\mathcal{K}} - \mathcal{K}_0) &= \begin{bmatrix} \Delta S_{12}\mathcal{K}_2 + S_{11}\sqrt{T}(\tilde{\mathcal{K}}_1 - \mathcal{K}_1) + S_{12}\sqrt{T}(\tilde{\mathcal{K}}_2 - \mathcal{K}_2) \\ \Delta S_{22}\mathcal{K}_2 + S_{22}\sqrt{T}(\tilde{\mathcal{K}}_2 - \mathcal{K}_2) \end{bmatrix} + o(1), \\
(\hat{\mathcal{C}} - \mathcal{C}_0)D_x^{-1} &= (\tilde{\mathcal{C}} - \mathcal{C})D_x^{-1}\begin{bmatrix} S_{11}^{-1} & 0 \\ 0 & S_{22}^{-1} \end{bmatrix} - \mathcal{C}_0\begin{bmatrix} \Delta S_{11}S_{11}^{-1} & \Delta S_{12}S_{22}^{-1} \\ 0 & \Delta S_{22}S_{22}^{-1} \end{bmatrix} + o(1).
\end{aligned}$$

Proof. The proof follows from straightforward computations using the various orders of convergence by neglecting higher order terms. □

Appendix B. Reduced Rank Regression with Integrated Variables

The main results of this paper are based on a more general result documented in [24] (henceforth called BRRR). BRRR uses a slightly different setting and in particular a different dgp. The following lemma provides the essence of the results of BRRR that will be used below.

Lemma A4. *Let* $(y_t)_{t\in\mathbb{N}}, (z_t^r)_{t\in\mathbb{N}}, (z_t^u)_{t\in\mathbb{N}}, y_t \in \mathbb{R}^s, z_t^r \in \mathbb{R}^m, z_t^u \in \mathbb{R}^l$ *be three processes related via*

$$y_t = b_r z_t^r + b_u z_t^u + u_t$$

where the zero mean stationary process $(u_t)_{t\in\mathbb{N}}$ *is such that* $\mathbb{E}u_t(z_t^r)' = 0, \mathbb{E}u_t(z_t^u)' = 0, \mathbb{E}u_t u_t' > 0$ *and where* $n = rank(b_r) < \min(s, m)$*, that is* b_r *is of reduced rank.*

Further assume that there exist square nonsingular matrices $\mathcal{T}_y \in \mathbb{R}^{s\times s}, \mathcal{T}_r \in \mathbb{R}^{m\times m}, \mathcal{T}_u \in \mathbb{R}^{n\times n}$ *such that*

$$\tilde{y}_t = \mathcal{T}_y y_t = (\mathcal{T}_y b_r \mathcal{T}_r^{-1})(\mathcal{T}_r z_t^r) + (\mathcal{T}_y b_r \mathcal{T}_r^{-1})(\mathcal{T}_r z_t^r) + \mathcal{T}_y u_t = \tilde{b}_r \tilde{z}_t + \tilde{b}_u \tilde{z}_t^u + \tilde{u}_t$$

such that with $c_\bullet = n - c$ *we have*

$$\tilde{b}_r = \begin{bmatrix} I_c & 0 & 0 \\ 0 & 0 & \tilde{b}_{r,\bullet} \end{bmatrix}, \quad \tilde{b}_{r,\bullet} = \tilde{O}_\bullet \Gamma_\bullet', \quad \tilde{O}_\bullet \in \mathbb{R}^{(s-c)\times c_\bullet}, \quad \Gamma_\bullet \in \mathbb{R}^{m_\bullet \times c_\bullet}.$$

Here the partitioning corresponds to $\tilde{z}_t' = [\tilde{z}_{t,1}', \tilde{z}_{t,2}', \tilde{z}_{t,\bullet}']$ *where* $\tilde{z}_{t,1} \in \mathbb{R}^c, \tilde{z}_{t,2} \in \mathbb{R}^{m-c-m_\bullet}$ *are MFI(1) processes and* $(\tilde{z}_{t,\bullet})_{t\in\mathbb{N}}, \tilde{z}_{t,\bullet} \in \mathbb{R}^{m_\bullet}$ *is stationary,* $\tilde{z}_t^u = [(\tilde{z}_{t,1}^u)', (\tilde{z}_{t,\bullet}^u)']'$ *where* $(\tilde{z}_{t,1}^u)_{t\in\mathbb{N}}$ *is a MFI(1) process and* $(\tilde{z}_{t,\bullet}^u)_{t\in\mathbb{N}}$ *is stationary and where the following bounds hold* ($\tilde{z}_{t,:} = [\tilde{z}_{t,1}', \tilde{z}_{t,2}']'$)*:*

$$\begin{array}{lll}
\langle \tilde{u}_t, \tilde{u}_t \rangle = O(1) & , \ \langle \tilde{u}_t, \tilde{z}_{t,\bullet} \rangle = O(Q_T) & , \ \langle \tilde{u}_t, \tilde{z}_{t,\bullet}^u \rangle = O(Q_T), \\
\langle \tilde{u}_t, \tilde{u}_t \rangle - \mathbb{E}\tilde{u}_t \tilde{u}_t' = O(Q_T) & , \ \langle \tilde{u}_t, \tilde{z}_{t,:} \rangle = O(\log T) & , \ \langle \tilde{u}_t, \tilde{z}_{t,1}^u \rangle = O(\log T), \\
\hat{M}_\bullet = \langle \begin{pmatrix} \tilde{z}_{t,\bullet} \\ \tilde{z}_{t,\bullet}^u \end{pmatrix}, \begin{pmatrix} \tilde{z}_{t,\bullet} \\ \tilde{z}_{t,\bullet}^u \end{pmatrix} \rangle & , \ \hat{M}_\bullet^{-1} = O(1) & , \ \hat{M}_\bullet = O(1), M_\bullet > 0 \\
\hat{M}_1 = \langle \begin{pmatrix} \tilde{z}_{t,:} \\ \tilde{z}_{t,1}^u \end{pmatrix}, \begin{pmatrix} \tilde{z}_{t,:} \\ \tilde{z}_{t,1}^u \end{pmatrix} \rangle & , \ \hat{M}_1 / T = O(\log\log T) & , \ (\hat{M}_1)^{-1} = O(Q_T^2), \\
\langle \begin{pmatrix} \tilde{z}_{t,\bullet} \\ \tilde{z}_{t,\bullet}^u \end{pmatrix}, \begin{pmatrix} \tilde{z}_{t,:} \\ \tilde{z}_{t,1}^u \end{pmatrix} \rangle = O(\log T) & , \ \hat{M}_\bullet - M_\bullet = O(Q_T). &
\end{array}$$

Then the reduced rank regression estimator $\hat{b}_{RRR} = [\hat{b}_{r,RRR}, \hat{b}_{u,RRR}]$ *maximizing the Gaussian likelihood subject to* $rank(\beta_r) = n = c + c_\bullet$ *is consistent:* $\hat{b}_{RRR} - b = O((\log T)^a / \sqrt{T})$ *for some* $a < \infty$*. Furthermore* $\tilde{b}_{RRR,r} - \tilde{b}_r = [O((\log T)^a / T), O((\log T)^a / \sqrt{T})]$ *with* $\tilde{b}_{RRR,r} = \mathcal{T}_y \hat{b}_{RRR,r} \mathcal{T}_r^{-1}$*, where the second block has* $m_\bullet$ *columns and corresponds to the stationary components of the regressor vector.*

Proof. The theorem slightly extends the results of BRRR by adding high level assumptions instead of low level assumptions on the data generating process. The proof hence consists in adjusting the proof in BRRR. In the following we only indicate where arguments in BRRR need to be replaced. A detailed proof would replicate much of the arguments in BRRR and hence is omitted.

The representation of Theorem 3.1 in BRRR is contained in the assumptions. Then consistency follows from examining the proof of the first part of Theorem 3.2 in BRRR: essential for the norm bounds are Lemma A.1 (I) and (III). The norm bounds stated under point (I) are directly assumed in this lemma except for the filtered version using n_t in place of x_t. Instead, here the results for n_t which are needed in the proof of Theorem 3.2 of BRRR are directly assumed. (III) then follows. Lemmas A.3–A.5 in BRRR do not depend on the assumptions on the various processes and hence continue to hold. Then the proof for consistency in Appendix A.3.1 of BRRR only uses these norm bounds referring also to [38] (which is also only based on the norm bounds contained in the assumptions of this lemma) and hence continues to hold. □

Appendix C. Proofs of the Theorems

Appendix C.1. Proof of Theorem 1

For proving consistency of the transfer function estimators it is sufficient to find a (possibly) random matrix $\tilde{\mathcal{S}}_T$ such that the least squares estimates $(\tilde{\mathcal{A}}, \tilde{\mathcal{C}}, \tilde{\mathcal{K}})$ of one representation $(\mathcal{A}, \mathcal{C}, \mathcal{K})$ of the true system obtained using $\tilde{x}_t := \tilde{\mathcal{S}}_T \hat{x}_t$ converges (a.s.) to $(\mathcal{A}, \mathcal{C}, \mathcal{K})$. This will be done in two steps: First a particular basis (which is not realizable in practice) will be chosen such that $\tilde{\mathcal{K}}_p - \mathcal{K}_p = o(1)$ sufficiently fast such that in the second step the regressions in the system equations based on the resulting state estimator $\tilde{x}_t$ are consistent. The derivation of the first step will also provide an approximation of the error term which can be used in order to derive the asymptotic distribution.

Appendix C.1.1. Proof of Theorem 1 (I)

The central step in `CVA` is the solution to the RRR problem. The following proof heavily draws on the results contained in [24] (henceforth called BRRR) collected in Lemma A4 for easier reference. As in BRRR, in order to derive the asymptotic properties we first transform the vectors in order to separate stationary and nonstationary terms. In order to achieve the separation let $Z_t = [y'_{t-1}, y'_{t-2}, ..., y'_{t-S}]' \in \mathbb{R}^{sS}$. Then for $p = kS$ we obtain

$$Y^-_{t,p} = \begin{pmatrix} y_{t-1} \\ y_{t-2} \\ \vdots \\ y_{t-S} \\ y_{t-S-1} \\ \vdots \\ y_{t-kS} \end{pmatrix} = \begin{pmatrix} Z_t \\ Z_{t-S} \\ \vdots \\ Z_{t-(k-1)S} \end{pmatrix}.$$

It is easy to see that for each j the process $(Z_{rS-j})_{r\in\mathbb{N}}$ is an $I(1)$ process. Moreover the strict minimum-phase condition for $(\mathcal{A}_\circ, \mathcal{C}_\circ, \mathcal{K}_\circ)$ implies that also for the system corresponding to $(Z_{rS-j})_{r\in\mathbb{N}}$ the strict minimum-phase condition holds.

Define the transformation $\mathcal{T}_S := [\mathcal{O}_{S,1}, \mathcal{O}_{S,\perp}]'$ where $\mathcal{O}_{S,1} \in \mathbb{R}^{sS\times c}$ denotes the matrix containing the first c columns of $\mathcal{O}_S$ for the system $(\mathcal{A}_\circ, \mathcal{C}_\circ, \mathcal{K}_\circ)$ in the canonical form. Further $\mathcal{O}_{S,\perp}$ is a block column of an orthonormal matrix such that $\mathcal{O}'_{S,\perp}\mathcal{O}_{S,1} = 0$. Then the argument of [20] shows that in $\mathcal{T}_S Z_t$ the first c components are integrated while the remaining $sS - c$ components are stationary. Then consider for $p = kS < t \le T - f + 1$ (using $\mathcal{O}^\dagger_{S,1} = (\mathcal{O}'_{S,1}\mathcal{O}_{S,1})^{-1}\mathcal{O}'_{S,1}$)

$$\tilde{z}_{t,p} := \begin{bmatrix} \mathcal{O}^\dagger_{S,1}\mathcal{O}_S(x_t - \bar{\mathcal{A}}^p_\circ x_{t-p}) \\ \mathcal{O}'_{S,\perp}Z_t \\ \mathcal{O}^\dagger_{S,1}(Z_t - Z_{t-S}) \\ \mathcal{O}'_{S,\perp}Z_{t-S} \\ \vdots \\ \mathcal{O}^\dagger_{S,1}(Z_{t-(k-2)S} - Z_{t-(k-1)S}) \\ \mathcal{O}'_{S,\perp}Z_{t-(k-1)S} \end{bmatrix}, \quad \tilde{y}_t := \begin{bmatrix} \mathcal{O}^\dagger_{f,1} \\ \mathcal{O}'_{f,\perp} \end{bmatrix} Y^+_{t,f}. \tag{A1}$$

Here $\mathcal{O}_{f,\perp}$ is a matrix such that $\mathcal{O}'_{f,\perp}\mathcal{O}_{f,1} = 0, \mathcal{O}'_{f,\perp}\mathcal{O}_{f,\perp} = I$. Obviously $\tilde{z}_{t,p}$ is a linear transformation of $Y^-_{t,p}$ and $\tilde{y}_t$ of $Y^+_{t,f}$. It can be shown that the linear transformation is nonsingular such that there is a one-one relation between $Y^-_{t,p}$ and $\tilde{z}_{t,p}$. In $\tilde{z}_{t,p}$ and $\tilde{y}_t$ only the first c components are unit root processes, the remaining components being stationary.

For $p \neq kS$ the final $p - kS$ block rows of $\tilde{z}_{t,p}$ are defined as $y_{t-(k-1)S-j} - y_{t-kS-j}, j = 1, ..., p - kS$. Clearly also these components are stationary.

Partition $\tilde{z}_{t,p} = [\tilde{z}'_{t,1}, \tilde{z}'_{t,\bullet}]', \tilde{z}_{t,1} \in \mathbb{R}^c$, into its first c and the remaining coordinates (omitting the subscript $_p$ on the right hand side for notational convenience). Similarly

partition $\tilde{y}_t = [\tilde{y}'_{t,1}, \tilde{y}'_{t,\bullet}]', \tilde{y}_{t,1} \in \mathbb{R}^c$. Using these transformed matrices, $Y^+_{t,f} = \beta_1 Y^-_{t,p} + N^+_{t,f}$ can be written as

$$\tilde{y}_t = \tilde{b}_1 \tilde{z}_{t,p} + \tilde{N}^+_{t,f,p} = \begin{bmatrix} \tilde{y}_{t,1} \\ \tilde{y}_{t,\bullet} \end{bmatrix} = \begin{bmatrix} I_c & 0 \\ 0 & \tilde{b}_{\bullet,p} \end{bmatrix} \begin{bmatrix} \tilde{z}_{t,1} \\ \tilde{z}_{t,\bullet} \end{bmatrix} + \tilde{\mathcal{O}}_f \bar{\mathcal{A}}^p_\circ x_{t-p} + \begin{bmatrix} \tilde{\varepsilon}_{t,1} \\ \tilde{\varepsilon}_{t,\bullet} \end{bmatrix} \tag{A2}$$

where

$$\tilde{b}_1 = \begin{bmatrix} I_c & 0 \\ 0 & \tilde{b}_{\bullet,p} \end{bmatrix}, \quad \tilde{b}_{\bullet,p} = \mathbb{E}\tilde{y}_{t,\bullet}\tilde{z}'_{t,\bullet}\left(\mathbb{E}\tilde{z}_{t,\bullet}\tilde{z}'_{t,\bullet}\right)^{-1} = O_{\bullet,p}\Gamma'_{\bullet,p}, \quad \tilde{b}_1 = O_p\Gamma'_p$$

and where $\tilde{b}_{\bullet,p}$ is of rank $n-c$ providing a representation of the form given in Theorem 3.1 of BRRR except that the error term $\tilde{N}^+_{t,f,p}$ (defined by the equation) is not white. Finally (A2) also defines the sub blocks $\tilde{\varepsilon}_{t,i}$ of $\tilde{N}^+_{t,f,p}$ which are hence linear combinations of $E^+_{t,f}$ and therefore typically MA(f) processes. Note that $\tilde{z}_{t,1}, \tilde{z}_{t,\bullet}, \tilde{y}_{t,\bullet}$ depend on the choice of f and p which is not reflected in the notation.

Here $\left(\mathbb{E}\tilde{z}_{t,\bullet}\tilde{z}'_{t,\bullet}\right)^{-1}$ and $\mathbb{E}\tilde{y}_{t,\bullet}\tilde{z}'_{t,\bullet}$ are worth a remark: for $p = kS$ the results of [20] can be directly used to obtain upper and lower bounds for the norms of these matrices uniformly in $k \in \mathbb{N}$. For $p \neq kS$ the additional rows in $\tilde{z}_{t,\bullet}$ add entries to $\mathbb{E}\tilde{y}_{t,\bullet}\tilde{z}'_{t,\bullet}$ that are of order $O(\lambda^p)$ for some $0 < \lambda < 1$ as $y_t - y_{t-S}$ is a VARMA process. Similarly the smallest eigenvalue of $\mathbb{E}\tilde{z}_{t,\bullet}\tilde{z}'_{t,\bullet}$ can be bounded from below based on arguments for $p = kS$ following [20] which in turn refer to Theorem 6.6.10 of [22]. The additional terms for $p \neq kS$ correspond to backward innovations with non-singular covariance matrix thus also leading to a lower bound of the smallest eigenvalue uniformly in k. (The backward innovations representation for a stationary VARMA process $(y_t)_{t\in\mathbb{Z}}$ equals $y_t = \sum_{j=1}^\infty k^b_j y_{t+j} + \varepsilon^b_t$ and can be obtained from the complex conjugate of the spectral density. Nonsingularity of the spectral density implies that the backward innovation ε^b_t have nonsingular covariance matrix. This implies a lower bound on the accuracy with which components of $y_{t-(k-1)S-j}$ can be predicted based on $y_{t-i}, i \leq (k-1)S$.)

Furthermore the strict minimum-phase assumption for the state space representation $(\mathcal{A}_\circ, \mathcal{C}_\circ, \mathcal{K}_\circ)$ of the process $(y_t)_{t\in\mathbb{Z}}$ implies the strict minimum-phase assumption for the sub-sampled process $(Z_{kS+j})_{k\in\mathbb{Z}}$. Thus the arguments of [20] show that $[\tilde{b}_{\bullet,p}, 0] \to \tilde{b}_{\bullet,\infty}$ where the norm of the difference is of order $O(\|\bar{\mathcal{A}}^p_\circ\|)$. The increase of p as a function of the sample size jointly with the strict minimum-phase assumption implies that $O(\|\bar{\mathcal{A}}^p_\circ\|) = o(T^{-1})$. This also implies that $\tilde{\mathcal{O}}_f \bar{\mathcal{A}}^p_\circ x_{t-p} = o_p(T^{-1/2})$.

Correspondingly there exists a limiting decomposition $\tilde{b}_{\bullet,\infty} = O_\bullet \Gamma'_\bullet$ such that $\Gamma'_\bullet S_\bullet = I_{n-c}$ where $S_\bullet$ denotes a selector matrix whose columns contain the vectors of the canonical basis of $\mathbb{R}^\infty$. Since $[\mathcal{K}_\circ, (\mathcal{A}_\circ - \mathcal{K}_\circ\mathcal{C}_\circ)\mathcal{K}_\circ, (\mathcal{A}_\circ - \mathcal{K}_\circ\mathcal{C}_\circ)^2\mathcal{K}_\circ, ..., (\mathcal{A}_\circ - \mathcal{K}_\circ\mathcal{C}_\circ)^{n-1}\mathcal{K}_\circ]$ is of full row rank it can be assumed that $S_\bullet$ only has nonzero entries in its first ns rows. Denoting the submatrix of the first ps rows by $S_{p,2}$ then also $[\Gamma'_\bullet]_{1:p} S_{p,2} = I_{n-c}$ where $[.]_{1:p}$ denotes the first p block columns of a matrix. This fixes a unique decomposition of $\tilde{b}_\bullet$ and hence $O_\bullet$ and $\Gamma_\bullet$ do not depend on p. Convergence of $\tilde{b}_{\bullet,p}$ to $\tilde{b}_\bullet$ jointly with the lower bound on $p(T)$ then implies convergence of order $o(T^{-1})$ of $O_{\bullet,p}$ to $O_\bullet$ and $\Gamma'_{\bullet,p}$ to $[\Gamma'_\bullet]_{1:p}$ using the decomposition of $\tilde{b}_{\bullet,p}$ such that $\Gamma'_{\bullet,p} S_{p,2} = I_{n-c}$. Correspondingly $O_p \to O$ and $\|\Gamma'_p - [\Gamma']_{1:p}\| \to 0$.

Therefore the reduced rank regression in the `CVA` procedure shows the same structure as investigated in Lemma A4 with the differences that $\tilde{z}_{t,2}$ and $\tilde{z}^u_t$ do not occur, and $\tilde{z}_{t,\bullet}$ has increasing size as a function of the sample size. The next lemma therefore extends the results of BRRR to the RRR used in `CVA`:

In the following we will use a generic $a \in \mathbb{N}$ in statements like $O((\log T)^a)$, not necessarily the same in each occurrence. In this sense e.g., the product of two terms that are $O((\log T)^a)$ is again taken to be $O((\log T)^a)$.

Lemma A5. *Let the assumptions of Theorem 1 hold where additionally $(\varepsilon_t)_{t\in\mathbb{Z}}$ is iid. Introduce the notation*

$$\tilde{D}_z = diag(T^{-1/2}I_c, I_{ps-c}), \quad \tilde{D}_y = diag(T^{-1/2}I_c, I_{fs-c}), \quad \tilde{D}_x = diag(T^{-1/2}I_c, I_{n-c}).$$

Let $\bar{G}_p$ denote a solution to

$$(\tilde{D}_z\langle\tilde{z}_{t,p},\tilde{y}_t\rangle\tilde{D}_y)(\tilde{D}_y\langle\tilde{y}_t,\tilde{y}_t\rangle\tilde{D}_y)^{-1}(\tilde{D}_y\langle\tilde{y}_t,\tilde{z}_{t,p}\rangle\tilde{D}_z)\bar{G}_p = (\tilde{D}_z\langle\tilde{z}_{t,p},\tilde{z}_{t,p}\rangle\tilde{D}_z)\bar{G}_p\bar{R}^2$$

using the notation of (A1) where $\bar{R}^2 \to \Theta^2 = diag(I_c, \Theta_\bullet) \in \mathbb{R}^{n\times n}$ and where $\bar{G}_p$ is normalized such that $\bar{G}_{1,1,p} = I_c$, $\bar{G}'_{\bullet,2,p}S_{p,2} = I_{n-c}$ for a selector matrix $S_{p,2}$. Further let

$$\bar{\Gamma}_p = \begin{bmatrix} I_c & 0 \\ 0 & \bar{\Gamma}_{\bullet,p} \end{bmatrix}, \bar{\Gamma}'_{\bullet,p}S_{p,2} = I_{n-c}$$

denote the solution to the decoupled problem where the stationary and the nonstationary subproblem are separated:

$$\begin{pmatrix} \langle\tilde{z}_{t,1},\tilde{y}_{t,1}\rangle\langle\tilde{y}_{t,1},\tilde{y}_{t,1}\rangle^{-1}\langle\tilde{y}_{t,1},\tilde{z}_{t,1}\rangle\bar{\Gamma}_{1,1,p} \\ \langle\tilde{z}_{t,\bullet},\tilde{y}_{t,\bullet}\rangle\langle\tilde{y}_{t,\bullet},\tilde{y}_{t,\bullet}\rangle^{-1}\langle\tilde{y}_{t,\bullet},\tilde{z}_{t,\bullet}\rangle\bar{\Gamma}_{\bullet,p} \end{pmatrix} = \begin{pmatrix} \langle\tilde{z}_{t,1},\tilde{z}_{t,1}\rangle\bar{\Gamma}_{1,1,p}\bar{\Theta}_1 \\ \langle\tilde{z}_{t,\bullet},\tilde{z}_{t,\bullet}\rangle\bar{\Gamma}_{\bullet,p}\bar{\Theta}_\bullet \end{pmatrix}.$$

(I) Then if $f \geq n$ fixed independent of T and $p \geq -d\log T/\log\rho_0, d > 1$, $p = o((\log T)^{\bar{a}})$ for some $\bar{a} < \infty$ the a.s. results of Lemma A.6 (I)-(III) and Lemma A.7 of [24] hold true for $(\log T)^3$ replaced by $(\log T)^a$ for some integer $a < \infty$. In particular $\bar{G}_p - \bar{\Gamma}_p = O((\log T)^a/T^{1/2})$.
(II) Using the notation $\delta G_p := \bar{G}_p - \bar{\Gamma}_p$ define

$$\tilde{\mathcal{S}}_T := \begin{bmatrix} I_c & -\sqrt{T}\delta G'_{\bullet,1,p}\langle\tilde{z}_{t,\bullet},\tilde{z}_{t,\bullet}\rangle\bar{\Gamma}^\dagger_{\bullet,p} \\ 0 & I_{ps-c} \end{bmatrix}, \quad \bar{\Gamma}^\dagger_{\bullet,p} := \bar{\Gamma}_{\bullet,p}(\bar{\Gamma}'_{\bullet,p}\langle\tilde{z}_{t,\bullet},\tilde{z}_{t,\bullet}\rangle\bar{\Gamma}_{\bullet,p})^{-1}.$$

Then for $\tilde{\Gamma}'_p := \tilde{\mathcal{S}}_T\tilde{D}_x^{-1}\bar{G}'_p\tilde{D}_z$ and

$$\Gamma' = \begin{bmatrix} I & 0 \\ 0 & \Gamma'_\bullet \end{bmatrix}$$

we obtain $\tilde{\Gamma}'_p - [\Gamma']_{1:p} = [O((\log T)^a/T), O((\log T)^a/T^{1/2})]$ where the partitioning corresponds to the partitioning of $\tilde{z}_{t,p}$ into $\tilde{z}_{t,1}$ and $\tilde{z}_{t,\bullet}$. Here $\Gamma'_\bullet$ denotes the right factor of $\tilde{b}_{\bullet,\infty} = O_\bullet\Gamma'_\bullet$ such that $[\Gamma'_\bullet]_{1:p}S_{p,2} = I_{n-c}$ holds.

(III) Let the assumptions of Theorem 1 hold. Then $\hat{Z}_T := Tvec\left((\tilde{\Gamma}'_p - [\Gamma']_{1:p})\begin{bmatrix} I_c \\ 0 \end{bmatrix}\right)$ converges in distribution.

Proof. (I) First consider the entries of the vectors $\tilde{y}_{t,\bullet}$ and $\tilde{z}_{t,\bullet}$ (see (A1)) more closely. Since in

$$\mathcal{O}'_{f,\perp}Y^+_{t,f} = \mathcal{O}'_{f,\perp}(\mathcal{O}_{f,\bullet}x_{t,\bullet} + \mathcal{E}_f E^+_{t,f})$$

the nonstationary directions are filtered by definition, $\tilde{y}_{t,\bullet}$ is stationary and does not depend on T.

Further, also $\tilde{z}_{t,\bullet}$ is stationary for fixed p as the nonstationary directions are either filtered by pre-multiplication with $\mathcal{O}'_{S,\perp}$ or by yearly differencing $Z_t - Z_{t-S}$.

Therefore we obtain from stationary theory for fixed $p = kS$ that

$$\|\mathbb{E}\tilde{y}_{t,\bullet}\tilde{z}'_{t,\bullet}(\mathbb{E}\tilde{z}_{t,\bullet}\tilde{z}'_{t,\bullet})^{-1} - \langle\tilde{y}_{t,\bullet},\tilde{z}_{t,\bullet}\rangle\langle\tilde{z}_{t,\bullet},\tilde{z}_{t,\bullet}\rangle^{-1}\| = o(1).$$

Here $\sup_p\|(\mathbb{E}\tilde{z}_{t,\bullet}\tilde{z}'_{t,\bullet})^{-1}\| < \infty$ has been discussed before. Now $\mathbb{E}\tilde{y}_{t,\bullet}\tilde{z}'_{t,\bullet}(\mathbb{E}\tilde{z}_{t,\bullet}\tilde{z}'_{t,\bullet})^{-1} = \tilde{\beta}_{\bullet,p} + o(T^{-1/2}) = O_{\bullet,p}[\Gamma'_\bullet]_{1:p} + o(T^{-1/2})$ where the $o(T^{-1/2})$ term appears due to neglecting $\tilde{\mathcal{O}}_f\bar{A}^p x_{t-p}$. It follows that $\det[(\tilde{\beta}_{\bullet,p}S_{p,2})'(\tilde{\beta}_{\bullet,p}S_{p,2})] = \det[O'_{\bullet,p}O_{\bullet,p}] > 0$ and

hence $\|\hat{\beta}_{\bullet,p} - \tilde{\beta}_{\bullet,p}\|_{Fr} = o(1)$ implies $\lim_{T\to\infty} \det\left[(\hat{\beta}_{\bullet,p}S_{p,2})'(\hat{\beta}_{\bullet,p}S_{p,2})\right] > 0$ a.s. where $\hat{\beta}_{\bullet,p} := \langle \tilde{y}_{t,\bullet}, \tilde{z}_{t,\bullet}\rangle \langle \tilde{z}_{t,\bullet}, \tilde{z}_{t,\bullet}\rangle^{-1}$. Since $\hat{O}_{\bullet,p}\bar{\Gamma}'_{\bullet,p} - \tilde{\beta}_{\bullet,p} = o(1)$ due to consistency, also

$$\lim_{T\to\infty} \det\left[(\hat{O}_{\bullet,p}\bar{\Gamma}'_{\bullet,p}S_{p,2})'(\hat{O}_{\bullet,p}\bar{\Gamma}'_{\bullet,p}S_{p,2})\right] = \lim_{T\to\infty} \det \hat{O}'_{\bullet,p}\hat{O}_{\bullet,p} \det(\bar{\Gamma}'_{\bullet,p}S_{p,2})^2 > 0 \quad \text{a.s.}$$

Since $\bar{\Gamma}_{\bullet,p} - \Gamma_{\bullet,p} = o(1)$ due to the definition of $\bar{\Gamma}_{\bullet,p}$ and the continuity of the solution of the eigenvalue problem it follows that $\hat{O}_{\bullet,p} - O_{\bullet,p} = o(1)$ and therefore $\limsup_T \det \hat{O}'_{\bullet,p}\hat{O}_{\bullet,p} > 0$. As in Lemma 6 of [40] it can be shown that $\Gamma'_{\bullet,p} - [\Gamma'_{\bullet}]_{1:p} = o(T^{-1})$ and $O_{\bullet,p} = O_{\bullet} + o(T^{-1})$ for the range of p given in Theorem 1 since these matrices correspond to a stationary problem. Hence the chosen normalization of $\bar{\Gamma}_{\bullet,p}$ can be used a.s.

Next in order to obtain the convergence of $\bar{G}$ to $\bar{\Gamma}_p$, Lemma A.6 of BRRR is slightly extended to the current situation (for details and notation see there). Lemma A.6 of BRRR contains three parts: BRRR(I) gives bounds on the error in the matrices (with $l_T = \log T$)

$$\begin{aligned}
\delta_{yz} &= \begin{bmatrix} \frac{1}{T}\langle \tilde{y}_{t,1}, \tilde{z}_{t,1}\rangle & \frac{1}{\sqrt{T}}\langle \tilde{y}_{t,1}, \tilde{z}_{t,\bullet}\rangle \\ \frac{1}{\sqrt{T}}\langle \tilde{y}_{t,\bullet}, \tilde{z}_{t,1}\rangle & \langle \tilde{y}_{t,\bullet}, \tilde{z}_{t,\bullet}\rangle \end{bmatrix} - \begin{bmatrix} \frac{1}{T}\langle \tilde{z}_{t,1}, \tilde{z}_{t,1}\rangle & 0 \\ 0 & \langle \tilde{y}_{t,\bullet}, \tilde{z}_{t,\bullet}\rangle \end{bmatrix} = \begin{bmatrix} O(\frac{1}{T}l_T^a) & O(\frac{1}{\sqrt{T}}l_T^a) \\ O(\frac{1}{\sqrt{T}}l_T^a) & 0 \end{bmatrix}, \\
\delta_{yy} &= \begin{bmatrix} \frac{1}{T}\langle \tilde{y}_{t,1}, \tilde{y}_{t,1}\rangle & \frac{1}{\sqrt{T}}\langle \tilde{y}_{t,\bullet}, \tilde{y}_{t,\bullet}\rangle \\ \frac{1}{\sqrt{T}}\langle \tilde{y}_{t,\bullet}, \tilde{y}_{t,1}\rangle & \langle \tilde{y}_{t,\bullet}, \tilde{y}_{t,\bullet}\rangle \end{bmatrix} - \begin{bmatrix} \frac{1}{T}\langle \tilde{z}_{t,1}, \tilde{z}_{t,1}\rangle & 0 \\ 0 & \langle \tilde{y}_{t,\bullet}, \tilde{y}_{t,\bullet}\rangle \end{bmatrix} = \begin{bmatrix} O(\frac{1}{T}l_T^a) & O(\frac{1}{\sqrt{T}}l_T^a) \\ O(\frac{1}{\sqrt{T}}l_T^a) & 0 \end{bmatrix}, \\
\delta_{zz} &= \begin{bmatrix} \frac{1}{T}\langle \tilde{z}_{t,1}, \tilde{z}_{t,1}\rangle & \frac{1}{\sqrt{T}}\langle \tilde{z}_{t,1}, \tilde{z}_{t,\bullet}\rangle \\ \frac{1}{\sqrt{T}}\langle \tilde{z}_{t,\bullet}, \tilde{z}_{t,1}\rangle & \langle \tilde{z}_{t,\bullet}, \tilde{z}_{t,\bullet}\rangle \end{bmatrix} - \begin{bmatrix} \frac{1}{T}\langle \tilde{z}_{t,1}, \tilde{z}_{t,1}\rangle & 0 \\ 0 & \langle \tilde{z}_{t,\bullet}, \tilde{z}_{t,\bullet}\rangle \end{bmatrix} = \begin{bmatrix} 0 & O(\frac{1}{\sqrt{T}}l_T^a) \\ O(\frac{1}{\sqrt{T}}l_T^a) & 0 \end{bmatrix}.
\end{aligned}$$

BRRR(II) deals with $J = \bar{Q} - \bar{\Phi} =$

$$\tilde{D}_z\langle \tilde{z}_t, \tilde{y}_t\rangle \tilde{D}_y(\tilde{D}_y\langle \tilde{y}_t, \tilde{y}_t\rangle \tilde{D}_y)^{-1}\tilde{D}_y\langle \tilde{y}_t, \tilde{z}_t\rangle \tilde{D}_z - \begin{bmatrix} \frac{1}{T}\langle \tilde{z}_{t,1}, \tilde{z}_{t,1}\rangle & 0 \\ 0 & \langle \tilde{z}_{t,\bullet}, \tilde{y}_{t,\bullet}\rangle\langle \tilde{y}_{t,\bullet}, \tilde{y}_{t,\bullet}\rangle^{-1}\langle \tilde{y}_{t,\bullet}, \tilde{z}_{t,\bullet}\rangle \end{bmatrix}$$

and BRRR(III) shows that there exists a solution $\bar{G}_p$ converging to a solution $\bar{\Gamma}_p$ of the separated problem.

For showing the orders of convergence of δ_{zz} the arguments are unchanged except for noting that in $\langle \tilde{z}_{t,1}, \tilde{z}_{t,\bullet}\rangle$ the number of columns increases as a function of the sample size. Since the a.s. bounds on the entries of this expression hold uniformly (as follows straightforwardly from the arguments of Lemma A.1 of BRRR) this does not change the arguments. With respect to δ_{yz} note that now $\tilde{y}_t = \tilde{\beta}_1\tilde{z}_{t,p} + \tilde{\varepsilon}_t + \tilde{\mathcal{O}}_f\bar{A}^p x_{t-p}$. Due to the increase of p as a function of the sample size, $\bar{A}^p = o(T^{-1-\epsilon})$ for small enough $\epsilon > 0$ and therefore $\tilde{\mathcal{O}}_f\bar{A}^p x_{t-p} = o(T^{-1/2-\epsilon/2})$ since $x_t = o(T^{(1+\epsilon)/2})$ (uniformly in $1 \le t \le T$) whether $(x_t)_{t\in\mathbb{Z}}$ is a unit root process or stationary. Hence $\langle \tilde{\mathcal{O}}_f\bar{A}^p x_{t-p}, \tilde{\mathcal{O}}_f\bar{A}^p x_{t-p}\rangle = o(1)$. Further $\langle \tilde{\mathcal{O}}_f\bar{A}^p x_{t-p}, \tilde{\varepsilon}_t\rangle = o(T^{-1/2})$ follows from $\langle x_{t-p}, \tilde{\varepsilon}_t\rangle = O(\log T)$ (see Lemma A.1 (I)). This shows that the additional term is always of lower order and can be neglected. The remaining arguments follow exactly as in the proof of Lemma A.6 of BRRR. The proof of Lemma A.7 of BRRR only uses the order bounds derived above and hence follows immediately. This shows (I).

(II) Using the definition of $\tilde{\mathcal{S}}_T$ we obtain:

$$\begin{aligned}
\tilde{\Gamma}'_p &= \tilde{\mathcal{S}}_T\tilde{D}_x^{-1}\bar{G}'_p\tilde{D}_z = \tilde{\mathcal{S}}_T\begin{bmatrix} I_c & \sqrt{T}\delta G'_{\bullet,1,p} \\ \delta G'_{1,2,p}/\sqrt{T} & \bar{G}'_{\bullet,2,p} \end{bmatrix} \\
&= \begin{bmatrix} I_c - \delta G'_{\bullet,1,p}\langle \tilde{z}_{t,\bullet}, \tilde{z}_{t,\bullet}\rangle \bar{\Gamma}^{\dagger}_{\bullet,p}\delta G'_{1,2,p} & \sqrt{T}\delta G'_{\bullet,1,p}(I - \langle \tilde{z}_{t,\bullet}, \tilde{z}_{t,\bullet}\rangle \bar{\Gamma}^{\dagger}_{\bullet,p}\bar{G}'_{\bullet,2,p}) \\ \delta G'_{1,2,p}/\sqrt{T} & \bar{G}'_{\bullet,2,p} \end{bmatrix}.
\end{aligned}$$

From (I) and Lemma A.7 of BRRR $\delta G_{\bullet,1,p} = O((\log T)^a/T^{1/2})$, $\delta G_{1,2,p} = O((\log T)^a/T^{1/2})$ and $\bar{G}_{\bullet,2,p} - \bar{\Gamma}_{\bullet,p} = o((\log T)^a/T^{1/2})$. Finally

$$\delta G'_{\bullet,1,p}(I - \langle \tilde{z}_{t,\bullet}, \tilde{z}_{t,\bullet}\rangle \bar{\Gamma}^{\dagger}_{\bullet,p}\bar{G}'_{\bullet,2,p}) = \delta G'_{\bullet,1,p}(I - \langle \tilde{z}_{t,\bullet}, \tilde{z}_{t,\bullet}\rangle \bar{\Gamma}^{\dagger}_{\bullet,p}\bar{\Gamma}'_{\bullet,p}) + O((\log T)^a/T) = O((\log T)^a/T)$$

as in the proof of Lemma A.7 of BRRR. Using Lemma A.5 (III) of BRRR with $\hat{\Xi}_f = \langle \tilde{y}_{t,\bullet}, \tilde{y}_{t,\bullet} \rangle^{-1/2}$ it follows that $\bar{\Gamma}'_{\bullet,p} - \Gamma'_{\bullet,p} = O((\log T)^a T^{-1/2})$. Since $\bar{G}_{\bullet,2,p} - \bar{\Gamma}_{\bullet,p} = o((\log T)^a / T^{1/2})$ the same rate of convergence holds for $\bar{G}'_{\bullet,2,p} - \Gamma'_{\bullet,p} = O((\log T)^a / T^{1/2})$. It follows that $\tilde{\Gamma}'_p - [\Gamma']_{1:p} = [O((\log T)^a / T), O((\log T)^a / T^{1/2})]$.

(III) From above we have

$$T(\tilde{\Gamma}'_p - [\Gamma']_{1:p}) \begin{pmatrix} I_c \\ 0 \end{pmatrix} = \begin{bmatrix} -(\sqrt{T}\delta G'_{\bullet,1,p} \langle \tilde{z}_{t,\bullet}, \tilde{z}_{t,\bullet} \rangle \bar{\Gamma}^{\dagger}_{\bullet,p} \sqrt{T} \delta G'_{1,2,p}) \\ \sqrt{T}\delta G'_{1,2,p} \end{bmatrix} + o_P(1). \tag{A3}$$

Now from the proof of Lemma A.7 of BRRR we obtain

$$[\sqrt{T}\delta G_{\bullet,1,p}]' = \Xi O_\bullet (I - \Theta^2_\bullet)^{-1} \bar{\Gamma}'_{\bullet,p} + o_P(1).$$

Furthermore using the expression given in Lemma A.7 of BRRR:

$$\begin{aligned}
\sqrt{T}\delta G_{1,2,p} &= \sqrt{T} Z_{11}^{-1} [\delta^{1\bullet}_{zz} \Gamma_{\bullet,p} \Theta^2_\bullet - J_{1,\bullet} \Gamma_{\bullet,p}](I - \Theta^2_\bullet)^{-1} + o_P(1) \\
&= \sqrt{T} Z_{11}^{-1} [\delta^{1\bullet}_{zz} \Gamma_{\bullet,p} (\Theta^2_\bullet - I) + [\delta^{1,\bullet}_{zz} - J_{1,\bullet}] \Gamma_{\bullet,p}](I - \Theta^2_\bullet)^{-1} + o_P(1) \\
&= -Z_{11}^{-1} \langle \tilde{z}_{t,1}, x_{t,\bullet} \rangle - Z_{11}^{-1} \sqrt{T} [J_{1,\bullet} - \delta^{1,\bullet}_{zz}] \Gamma_{\bullet,p} (I - \Theta^2_\bullet)^{-1} + o_P(1) \\
&= -Z_{11}^{-1} \langle \tilde{z}_{t,1}, x_{t,\bullet} \rangle - Z_{11}^{-1} \mathbb{E} \tilde{\varepsilon}_{t,1} \tilde{\varepsilon}'_{t,\bullet} (\mathbb{E} \tilde{y}_{t,\bullet} (\tilde{y}_{t,\bullet})')^{-1} \mathbb{E} \tilde{y}_{t,\bullet} x'_{t,\bullet} (I - \Theta^2_\bullet)^{-1} + o_P(1).
\end{aligned}$$

This shows the result. □

The transformations in the representation lead to an estimator $\bar{G}$ taking the place of $\hat{\mathcal{K}}_p$. Using $\tilde{\mathcal{S}}_T$ as defined in Lemma A5 the corresponding estimator $\tilde{\Gamma}'_p = \tilde{\mathcal{S}}_T \tilde{D}_x^{-1} \bar{G}'_p \tilde{D}_z$ fulfills $\tilde{\Gamma}'_p - \Gamma'_p = [O((\log T)^a / T), O((\log T)^a / \sqrt{T})]$.

Based on this result let $(\mathcal{A}, \mathcal{C}, \mathcal{K})$ denote the realization of the true transfer function in the state basis corresponding to Γ'_p where $\Gamma'_p S_p = I_n$ and let $(\tilde{\mathcal{A}}, \tilde{\mathcal{C}}, \tilde{\mathcal{K}})$ denote the (unfeasible) `CVA` estimates using $\tilde{x}_t := \tilde{\Gamma}'_p \tilde{z}_{t,p}$. The next lemma then provides the main ingredients for the rest of the proofs:

Lemma A6. *Let the assumptions of Theorem 1 hold and define* $D_x := diag(I_c T^{-1}, I_{n-c} T^{-1/2})$. *Then there exists an integer* $a < \infty$ *such that*

$$(\tilde{\mathcal{A}} - \mathcal{A}) D_x^{-1} = O((\log T)^a), \quad (\tilde{\mathcal{C}} - \mathcal{C}) D_x^{-1} = O((\log T)^a), \quad (\tilde{\mathcal{K}} - \mathcal{K}) = O((\log T)^a / T^{1/2}).$$

Proof. First note that the regression of $Y^+_{t,f}$ onto $Y^-_{t,p}$ includes time points $t = p+1, ..., T-f+1$ whereas for estimating the system matrices we can use $\hat{x}_t, t = p+1, ..., T+1$ and $y_t, t = p+1, ..., T$. Thus in this proof we use $\langle a_t, b_t \rangle^T_{p+1} := T^{-1} \sum_{t=p+1}^T a_t b'_t$ instead of $\langle a_t, b_t \rangle = T^{-1} \sum_{t=p+1}^{T-f+1} a_t b'_t$.

The following orders of convergence are straightforward to derive using the results of Lemma A1, $\bar{\mathcal{A}}^p = o(T^{-1})$, $(\tilde{\Gamma}'_p - [\Gamma']_{1:p}) D_z^{-1} = O((\log T)^a)$ and $\tilde{x}_t - x_t = (\tilde{\Gamma}'_p - [\Gamma']_{1:p}) \tilde{z}_{t,p} - \bar{\mathcal{A}}^p x_{t-p}, t > p$ according to Lemma A5 and Lemma A1 for the range of p given in Theorem 1:

$$\begin{aligned}
\langle \varepsilon_t, \tilde{x}_t - x_t \rangle^T_{p+1} = O(p(\log T)^a / T) \quad &, \quad \tilde{D}_z \langle \tilde{z}_{t,p}, \tilde{x}_t - x_t \rangle^T_{p+1} = O(p(\log T)^a / T^{1/2}) \\
\tilde{D}_z \langle \tilde{z}_{t+1,p}, \tilde{x}_t - x_t \rangle^T_{p+1} = O(p(\log T)^a / T^{1/2}) \quad &, \quad \tilde{D}_x \langle x_t, \tilde{x}_t - x_t \rangle^T_{p+1} = O(p(\log T)^a / T^{1/2}) \\
\langle \tilde{x}_t - x_t, \tilde{x}_t - x_t \rangle^T_{p+1} = O(p^2 (\log T)^a / T) \quad &.
\end{aligned}$$

Using these orders of convergence we obtain

$$\tilde{D}_x \langle \tilde{x}_t, \tilde{x}_t \rangle^T_{p+1} \tilde{D}_x = \tilde{D}_x \langle x_t, x_t \rangle^T_{p+1} \tilde{D}_x + O(p^2 (\log T)^a / T^{1/2}) > 0 \quad a.s.$$

From Lemma A1 also $(\tilde{D}_x\langle\tilde{x}_t,\tilde{x}_t\rangle_{p+1}^T\tilde{D}_x)^{-1} = (\tilde{D}_x\langle x_t,x_t\rangle_{p+1}^T\tilde{D}_x)^{-1}(1+o(1)) = O((\log T)^a)$.

Therefore

$$\begin{aligned}(\tilde{\mathcal{C}}-\mathcal{C})D_x^{-1} &= \sqrt{T}\Big(\langle\varepsilon_t,\tilde{x}_t\rangle_{p+1}^T - \mathcal{C}\langle\tilde{x}_t-x_t,\tilde{x}_t\rangle_{p+1}^T\Big)\tilde{D}_x(\tilde{D}_x\langle\tilde{x}_t,\tilde{x}_t\rangle_{p+1}^T\tilde{D}_x)^{-1}\\ &= \sqrt{T}\langle\varepsilon_t,x_t\rangle_{p+1}^T\tilde{D}_x(\tilde{D}_x\langle x_t,x_t\rangle_{p+1}^T\tilde{D}_x+o(1))^{-1}\\ &\quad -\sqrt{T}\mathcal{C}\langle\tilde{x}_t-x_t,x_t\rangle_{p+1}^T\tilde{D}_x(\tilde{D}_x\langle x_t,x_t\rangle_{p+1}^T\tilde{D}_x)^{-1}+o(1) = O(p(\log T)^a).\end{aligned} \tag{A4}$$

This in particular establishes consistency for the estimate. Next analogously (using the notation $\delta x_t = \tilde{x}_t - x_t$) we obtain $(\tilde{\mathcal{A}}-\mathcal{A})D_x^{-1} =$

$$\begin{aligned}&\sqrt{T}\langle\tilde{x}_{t+1}-\mathcal{A}\tilde{x}_t,\tilde{x}_t\rangle_{p+1}^T\tilde{D}_x(\tilde{D}_x\langle\tilde{x}_t,\tilde{x}_t\rangle_{p+1}^T\tilde{D}_x)^{-1}\\ =\;&\sqrt{T}\Big(\langle(\tilde{x}_{t+1}-x_{t+1})+(x_{t+1}-\mathcal{A}x_t)+\mathcal{A}(x_t-\tilde{x}_t),\tilde{x}_t\rangle_{p+1}^T\tilde{D}_x\Big)(\tilde{D}_x\langle x_t,x_t\rangle_{p+1}^T\tilde{D}_x+o(1))^{-1}\\ =\;&\sqrt{T}\Big(\langle\delta x_{t+1},x_t\rangle_{p+1}^T-\mathcal{A}\langle\delta x_t,x_t\rangle_{p+1}^T+\langle\mathcal{K}\varepsilon_t,x_t\rangle_{p+1}^T\Big)\tilde{D}_x(\tilde{D}_x\langle x_t,x_t\rangle_{p+1}^T\tilde{D}_x)^{-1}+o(1)\\ =\;&O(p(\log T)^a)\end{aligned} \tag{A5}$$

and therefore consistency for $\tilde{\mathcal{A}}$ is established. Finally note that for

$$\hat{\varepsilon}_t = y_t - \tilde{\mathcal{C}}\tilde{x}_t = \varepsilon_t + \mathcal{C}(x_t-\tilde{x}_t) + (\mathcal{C}-\tilde{\mathcal{C}})\tilde{x}_t$$

it follows that $\langle\hat{\varepsilon}_t,\hat{\varepsilon}_t\rangle_{p+1}^T = \Omega + O(p^2(\log T)^a/T^{1/2})$. Furthermore since $\hat{\varepsilon}_t$ denotes the residuals of the regression of y_t onto $\tilde{x}_t$ it follows that $\langle\hat{\varepsilon}_t,\bar{x}_t\rangle_{p+1}^T = 0$. From this we obtain

$$\begin{aligned}\sqrt{T}(\hat{\mathcal{K}}-\mathcal{K}) &= \sqrt{T}(\langle\tilde{x}_{t+1}-\mathcal{K}\hat{\varepsilon}_t,\hat{\varepsilon}_t\rangle_{p+1}^T(\langle\hat{\varepsilon}_t,\hat{\varepsilon}_t\rangle_{p+1}^T)^{-1}\\ &= \sqrt{T}\Big(\langle(\tilde{x}_{t+1}-x_{t+1})-\mathcal{A}\delta x_t+\mathcal{K}(\varepsilon_t-\hat{\varepsilon}_t),\hat{\varepsilon}_t\rangle_{p+1}^T\Big)(\langle\hat{\varepsilon}_t,\hat{\varepsilon}_t\rangle_{p+1}^T)^{-1}\\ &= \sqrt{T}\Big(\langle\delta x_{t+1}-\mathcal{A}\delta x_t+\mathcal{K}(\varepsilon_t-\hat{\varepsilon}_t),\hat{\varepsilon}_t\rangle_{p+1}^T\Big)\Omega^{-1}(1+o(1))\\ &= \sqrt{T}\Big(\langle\delta x_{t+1}-\mathcal{A}\delta x_t+\mathcal{K}(\varepsilon_t-\hat{\varepsilon}_t),\varepsilon_t\rangle_{p+1}^T\Big)\Omega^{-1}(1+o(1))+o(1)\\ &= \Big(\sqrt{T}\langle\delta x_{t+1},\varepsilon_t\rangle_{p+1}^T\Big)\Omega^{-1}+o(1) = \Big(\sqrt{T}\langle(\tilde{\Gamma}_p'-\Gamma_p')\bar{z}_{t+1,p},\varepsilon_t\rangle_{p+1}^T\Big)\Omega^{-1}+o(1)\\ &= O(p(\log T)^a).\end{aligned} \tag{A6}$$

□

These expressions do not only show consistency of a specific order, but also give the relevant highest order terms for the asymptotic distribution, which are used below.
As $\hat{\mathcal{C}}\hat{\mathcal{A}}^j\hat{\mathcal{K}} = \tilde{\mathcal{C}}\tilde{\mathcal{A}}^j\hat{\mathcal{K}} \to \mathcal{C}\mathcal{A}^j\mathcal{K} = \mathcal{C}_\circ\mathcal{A}_\circ^j\mathcal{K}_\circ$, Lemma A6 establishes consistency for the impulse response sequence $\hat{\mathcal{C}}\hat{\mathcal{A}}^j\hat{\mathcal{K}}$ (thus proofs Theorem 1 (I)) as well as, jointly with $p = O((\log T)^a)$, the rate of convergence $O((\log T)^a/T^{1/2})$ for the not realizable choice of the basis and the impulse response sequence $\mathcal{C}\mathcal{A}^j\mathcal{K}$.

Appendix C.1.2. Proof of Theorem 1 (II)

In order to arrive at the canonical representation $(\check{A},\check{C},\check{K})$ two steps are performed: first the reordered Jordan normal form is calculated, afterwards the matrices $\tilde{C}_{j,\mathbb{C}}$ are transformed such that $E_j'\check{C}_{j,\mathbb{C}} = I_{c_j}$ holds. We will follow these steps below.

In the first step a transformation matrix $\hat{U}$ needs to be found such that $\tilde{A} = \hat{U}\tilde{\mathcal{A}}\hat{U}^{-1}$ is in reordered Jordan normal form. In this respect $\tilde{\mathcal{A}}$ and $\mathcal{A}$ are used in Lemma A2. Accordingly $\hat{U}_t = [\hat{U}_{t,1},...,\hat{U}_{t,S/2},\hat{U}_{t,\bullet}]$ can be defined such that $V_j'\hat{U}_{t,j} = I_{c_j}$ where $U \in \mathbb{R}^{n\times n}$ corresponds to the transformation from $\mathcal{A}$ to $\mathcal{A}_\circ$ as given in the theorem. An appropriate choice of $\tilde{z}_{t,1}$ leads to $U = I_n$. Furthermore the basis in the space spanned by the columns of $\hat{U}_{t,\bullet}$ where $\hat{U}_{t,j}'\hat{U}_{t,\bullet} = 0$ can be chosen such that $[0,I]\hat{U}_{t,\bullet} = I$ for large enough T.

A first order approximation according to Lemma A2 then leads to

$$\hat{U}_{t,j} = U_j + Z_1 + O(\|\hat{\mathcal{A}}-\mathcal{A}\|^2) = U_j - \Sigma(\hat{\mathcal{A}}-\mathcal{A})U_j + O(\|\hat{\mathcal{A}}-\mathcal{A}\|^2)$$

for $j = 0, ..., S/2$. Consequently $\|\hat{U}_{t,j} - U_j\| = O((\log T)^a T^{-1})$ and thus also $\hat{U}_t - U = O((\log T)^a T^{-1})$. Consequently the order of convergence for the transformed system $(\hat{A}, \hat{C}, \hat{K})$ is unchanged. In a second step an upper triangular transformation matrix $\tilde{U}$ can be found transforming $(\hat{A}, \hat{C}, \hat{K})$ such that $\tilde{A}$ corresponds to the reordered Jordan normal form. Due to the upper block triangularity of this transform we can apply Lemma A3 to show that the order of convergence remains identical.

For the second step note that Lemma A3 provides the required terms: An application to the block diagonal transformation $\mathcal{S}_T = \mathrm{diag}(E_1'\tilde{C}_{1,\mathbb{C}}, ..., E_{S/2}'\tilde{C}_{S/2,\mathbb{C}}, \mathcal{S}_{T,\bullet})$, where $\mathcal{S}_{T,\bullet}$ transforms the stationary subsystem to echelon form, concludes the proof.

Appendix C.1.3. Proof of Theorem 1 (III)

The only argument that uses the iid assumption is the almost sure convergence of $(\tilde{D}_x\langle x_t, x_t\rangle\tilde{D}_x)^{-1}$. Weakening the assumptions on the noise implies that this order of convergence still holds in probability while the almost sure version cannot be shown with the tools of this paper. This concludes the proof of Theorem 1.

Appendix C.2. Proof of Theorem 2

Using the notation introduced in (A1),

$$\hat{X} = \tilde{D}_z\langle\tilde{y}_t, \tilde{z}_{t,p}\rangle\tilde{D}_z(\tilde{D}_z\langle\tilde{z}_{t,p}, \tilde{z}_{t,p}\rangle\tilde{D}_z)^{-1}\tilde{D}_z\langle\tilde{z}_{t,p}, \tilde{y}_t\rangle\tilde{D}_z(\tilde{D}_z\langle\tilde{y}_t, \tilde{y}_t\rangle\tilde{D}_z)^{-1} \to X_\circ = \begin{bmatrix} I_c & 0 \\ 0 & X_{\circ,\bullet} \end{bmatrix}$$

for a suitable matrix $X_{\circ,\bullet}$. The eigenvalues of $\hat{X}$ are the squares of the singular values of the RRR problem in the first step of `CVA`. Therefore

$$\begin{aligned} T\sum_{i=1}^{c}(1-\hat{\sigma}_i^2) &= -T\mathrm{tr}\Big[U_c'(\hat{X}-X_\circ)[U_c - (X_\circ - I)^\dagger(\hat{X}-X_\circ)U_c]\Big] + o_P(1) \\ &= -T\mathrm{tr}\Big[\Delta X_{11} - \Delta X_{12}(X_{\circ,\bullet} - I)^\dagger\Delta X_{21}\Big] + o_P(1) \end{aligned}$$

according to a second order approximation in the Rayleigh-Schrödinger expansions (Lemma A2). Now, in the current situation we obtain $(I - \hat{X})\begin{bmatrix} I \\ 0 \end{bmatrix} =$

$$\begin{aligned} &= \Big(\tilde{D}_y\langle\tilde{y}_t, \tilde{y}_t\rangle\tilde{D}_y - \tilde{D}_y\langle\tilde{y}_t, \tilde{z}_{t,p}\rangle\tilde{D}_z(\tilde{D}_z\langle\tilde{z}_{t,p}, \tilde{z}_{t,p}\rangle\tilde{D}_z)^{-1}\tilde{D}_z\langle\tilde{z}_{t,p}, \tilde{y}_t\rangle\tilde{D}_y\Big)(\tilde{D}_y\langle\tilde{y}_t, \tilde{y}_t\rangle\tilde{D}_y)^{-1}\begin{bmatrix} I \\ 0 \end{bmatrix} \\ &= \Big(\tilde{D}_y\langle\tilde{\varepsilon}_t, \tilde{\varepsilon}_t\rangle\tilde{D}_y - \tilde{D}_y\langle\tilde{\varepsilon}_t, \tilde{z}_{t,p}\rangle\tilde{D}_z(\tilde{D}_z\langle\tilde{z}_{t,p}, \tilde{z}_{t,p}\rangle\tilde{D}_z)^{-1}\tilde{D}_z\langle\tilde{z}_{t,p}, \tilde{\varepsilon}_t\rangle\tilde{D}_y\Big)(\tilde{D}_y\langle\tilde{y}_t, \tilde{y}_t\rangle\tilde{D}_y)^{-1}\begin{bmatrix} I \\ 0 \end{bmatrix}. \end{aligned}$$

Furthermore $\langle\tilde{\varepsilon}_t, \tilde{z}_{t,p}\rangle\tilde{D}_z(\tilde{D}_z\langle\tilde{z}_{t,p}, \tilde{z}_{t,p}\rangle\tilde{D}_z)^{-1}\tilde{D}_z\langle\tilde{z}_{t,p}, \tilde{\varepsilon}_t\rangle = O_P(T^{-1})$ and

$$(\tilde{D}_y\langle\tilde{y}_t, \tilde{y}_t\rangle\tilde{D}_y)^{-1}\begin{bmatrix} I \\ 0 \end{bmatrix} = \begin{bmatrix} I \\ -\langle\tilde{y}_{t,\bullet}, \tilde{y}_{t,\bullet}\rangle^{-1}\langle\tilde{y}_{t,\bullet}, \tilde{y}_{t,1}\rangle/\sqrt{T} \end{bmatrix}(\langle\tilde{y}_{t,1}^\pi, \tilde{y}_{t,1}^\pi\rangle/T)^{-1}$$

where $\tilde{y}_{t,1}^\pi = \tilde{y}_{t,1} - \langle\tilde{y}_{t,1}, \tilde{y}_{t,\bullet}\rangle\langle\tilde{y}_{t,\bullet}, \tilde{y}_{t,\bullet}\rangle^{-1}\tilde{y}_{t,\bullet}$.

From this we get using $\mathbb{E}\tilde{\varepsilon}_{t,\bullet}\tilde{\varepsilon}_{t,\bullet}' = \mathbb{E}\tilde{y}_{t,\bullet}\tilde{y}_{t,\bullet}' - X_{\circ,\bullet}\mathbb{E}\tilde{y}_{t,\bullet}\tilde{y}_{t,\bullet}'$:

$$\begin{aligned} T(I_c - \hat{X}_{1,1}) &= \Big(\langle\tilde{\varepsilon}_{t,1}, \tilde{\varepsilon}_{t,1}\rangle - \langle\tilde{\varepsilon}_{t,1}, \tilde{\varepsilon}_{t,\bullet}\rangle\langle\tilde{y}_{t,\bullet}, \tilde{y}_{t,\bullet}\rangle^{-1}\langle\tilde{y}_{t,\bullet}, \tilde{y}_{t,1}\rangle\Big)(\langle\tilde{y}_{t,1}^\pi, \tilde{y}_{t,1}^\pi\rangle/T)^{-1} + o_P(1), \\ \sqrt{T}\Delta X_{2,1} &= (-\langle\tilde{\varepsilon}_{t,\bullet}, \tilde{\varepsilon}_{t,1}\rangle + (I - X_{\circ,\bullet})\langle\tilde{y}_{t,\bullet}, \tilde{y}_{t,1}\rangle)(\langle\tilde{y}_{t,1}^\pi, \tilde{y}_{t,1}^\pi\rangle/T)^{-1} + o_P(1) \\ \sqrt{T}\Delta X_{1,2} &= -\mathbb{E}\tilde{\varepsilon}_{t,1}\tilde{\varepsilon}_{t,\bullet}'(\mathbb{E}\tilde{y}_{t,\bullet}\tilde{y}_{t,\bullet}')^{-1} + o_P(1). \end{aligned}$$

Thus $T\sum_{i=1}^{c}(1-\hat{\sigma}_i^2) =$

$$\begin{aligned} &= \operatorname{tr}\Big[\Big(\langle\tilde{\varepsilon}_{t,1},\tilde{\varepsilon}_{t,1}\rangle - \mathbb{E}\tilde{\varepsilon}_{t,1}\tilde{\varepsilon}'_{t,\bullet}(\mathbb{E}\tilde{y}_{t,\bullet}\tilde{y}'_{t,\bullet})^{-1}(I-X_{0,\bullet})^{-1}\mathbb{E}\tilde{\varepsilon}_{t,\bullet}\tilde{\varepsilon}'_{t,1}\Big)(\langle\tilde{y}_{t,1},\tilde{y}_{t,1}\rangle/T)^{-1}\Big] + o_P(1) \\ &= \operatorname{tr}\Big[\Big(\langle\tilde{\varepsilon}_{t,1},\tilde{\varepsilon}_{t,1}\rangle - \mathbb{E}\tilde{\varepsilon}_{t,1}\tilde{\varepsilon}'_{t,\bullet}(\mathbb{E}\tilde{\varepsilon}_{t,\bullet}\tilde{\varepsilon}'_{t,\bullet})^{-1}\mathbb{E}\tilde{\varepsilon}_{t,\bullet}\tilde{\varepsilon}'_{t,1}\Big)(\langle\tilde{y}_{t,1},\tilde{y}_{t,1}\rangle/T)^{-1}\Big] + o_P(1) \xrightarrow{d} Z. \end{aligned}$$

Appendix C.3. Proof of Theorem 3

The proof of Theorem 3 follows the same path as the proof of Theorem 1. In (A5) it was shown that the asymptotic distribution of $T(\tilde{\mathcal{A}}_{11}-\mathcal{A}_{\circ,11})$ depends on

$$\langle\tilde{x}_{t+1,j}-x_{t+1,j},x_{t,k}\rangle, \langle\tilde{x}_{t,j}-x_{t,j},x_{t,k}\rangle \quad , \quad \langle\varepsilon_t,x_{t,j}\rangle, \langle x_{t,k},x_{t,j}\rangle/T$$

for $j,k=0,...,S/2$. Note that

$$\delta x_{t+i} = \tilde{x}_{t+i}-x_{t+i} = (\tilde{\Gamma}'_p-[\Gamma']_{1:p})\tilde{z}_{t+i,p}+o_P(T^{-1})$$

for $i=0,1$. Then the results of Lemma A5 show that the first c columns of $(\tilde{\Gamma}'_p-[\Gamma']_{1:p})$ converge to a random variable (below denoted as Z_Γ) when multiplied with T while the remaining columns converge in distribution when multiplied with $\sqrt{T}$. Therefore

$$\langle\delta x_{t+i},x_{t,k}\rangle = T(\tilde{\Gamma}'_p-[\Gamma']_{1:p})\frac{\langle\tilde{z}_{t+i,p},x_{t,k}\rangle}{T}+o_P(1) = T(\tilde{\Gamma}'_p-[\Gamma']_{1:p})\begin{bmatrix} I_c \\ 0 \end{bmatrix}\frac{\langle\tilde{z}_{t+i,1},x_{t,k}\rangle}{T}+o_P(1).$$

Due to the definition (A1), $\tilde{z}_{t,1}=[x_{t,j}]_{j=0,...,S/2}+o(T^{-1})$ and hence (using $\mathcal{A}_\circ=\operatorname{diag}(\mathcal{A}_{\circ,u},\mathcal{A}_{\circ,\bullet})$)

$$\langle\tilde{z}_{t+1,1},x_{t,k}\rangle/T = \mathcal{A}_{\circ,u}\langle\tilde{z}_{t,1},x_{t,k}\rangle/T+o(1).$$

Considering now the complex-valued representation and using the notation

$$\Delta\Gamma_1 := T(\tilde{\Gamma}'_p-[\Gamma']_{1:p})\begin{bmatrix} I_c \\ 0 \end{bmatrix}, \quad S_j=[0_{c_j,\sum_{i<j}c_i},I_{c_j},0_{c_j,\sum_{i>j}c_j}]$$

where $S_j\tilde{z}_{t,1}=x_{t,j,\mathbb{C}}$, it follows that the contribution of these two terms to the limiting distribution of the diagonal block corresponding to the unit root z_j amounts to (using $\langle x_{t,j,\mathbb{C}},x_{t,k,\mathbb{C}}\rangle/T\to 0$ for $k\neq j$ and $\delta x_{t,j,\mathbb{C}}=\tilde{x}_{t,j,\mathbb{C}}-x_{t,j,\mathbb{C}}$)

$$\langle\delta x_{t+1,j,\mathbb{C}},x_{t,j,\mathbb{C}}\rangle - \mathcal{A}_{\circ,jj}\langle\delta x_{t,j,\mathbb{C}},x_{t,j,\mathbb{C}}\rangle =$$

$$\begin{aligned} &= S_j\Delta\Gamma_1\mathcal{A}_{\circ,u}\frac{\langle\tilde{z}_{t,1},x_{t,j,\mathbb{C}}\rangle}{T} - \mathcal{A}_{\circ,jj}S_j\Delta\Gamma_1\frac{\langle\tilde{z}_{t,1},x_{t,j,\mathbb{C}}\rangle}{T}+o_P(1) \\ &= S_j\Delta\Gamma_1S'_j\mathcal{A}_{\circ,jj}\frac{\langle x_{t,j,\mathbb{C}},x_{t,j,\mathbb{C}}\rangle}{T} - \mathcal{A}_{\circ,jj}S_j\Delta\Gamma_1S'_j\frac{\langle x_{t,j,\mathbb{C}},x_{t,j,\mathbb{C}}\rangle}{T}+o_P(1) \\ &= S_j\Delta\Gamma_1S'_j\overline{z_j}\frac{\langle x_{t,j,\mathbb{C}},x_{t,j,\mathbb{C}}\rangle}{T} - \overline{z_j}S_j\Delta\Gamma_1S'_j\frac{\langle x_{t,j,\mathbb{C}},x_{t,j,\mathbb{C}}\rangle}{T}+o_P(1) = o_P(1). \end{aligned}$$

Therefore, for the diagonal blocks in (A5) these two terms do not contribute and the asymptotic distribution is determined by

$$T\langle\mathcal{K}_{\circ,j}\varepsilon_t,x_{t,j}\rangle\langle x_{t,j},x_{t,j}\rangle^{-1}$$

for which the asymptotic results are provided in Lemma A1. This also shows that estimating the state does not change the asymptotic distribution in the diagonal blocks as the impact of $\tilde{\Gamma}_p-\Gamma_p$ is of lower order.

In order to derive the distribution of the sum of the eigenvalues note that as in the proof of Theorem 2, according to Lemma A2 the sum of the eigenvalues of $\tilde{\mathcal{A}}$ converging to z_j obeys the following second order approximation:

$$\begin{aligned} T\sum_{i=1}^{c_j}(\hat{\lambda}_i - z_j) &= T\mathrm{tr}\Big[U_j'(\hat{\mathcal{A}} - \mathcal{A}_\circ)[U_j - \mathcal{A}_\circ(z_j)^\dagger(\hat{\mathcal{A}} - \mathcal{A}_\circ)U_j]\Big] + o_P(T^{-1}) \\ &= T\mathrm{tr}\Big[\hat{\mathcal{A}}_{\circ,jj} - z_j I_{c_j}\Big] + o_P(1) \end{aligned}$$

since $(\hat{\mathcal{A}} - \mathcal{A}_\circ)U_j = O((\log T)^a T^{-1})$ in this case implying that the second order terms vanish. Thus we obtain the asymptotic distribution under the null hypothesis as the limiting distribution of

$$T\mathrm{tr}[\langle \mathcal{K}_{\circ,j,\mathbb{C}}\varepsilon_t, x_{t,j,\mathbb{C}}\rangle\langle x_{t,j,\mathbb{C}}, x_{t,j,\mathbb{C}}\rangle^{-1}].$$

It is easy to verify that this test statistic is pivotal for complex and real unit roots. This proves Theorem 3.

Appendix C.4. Proof of Theorem 4

The result for $\tilde{C}_m$ can be shown using the results of [4]. As the eigenvalues are insensitive to changes in the basis we can assume without restriction of generality that the only unit root components in $\mathcal{T}X_t^{(m)}$ are contained in the first c_m rows:

$$c_t^{(m)} := \mathcal{T}X_t^{(m)} = \begin{bmatrix} c_{t,u}^{(m)} \\ c_{t,\bullet}^{(m)} \end{bmatrix}, \quad \tilde{D}_c = \begin{bmatrix} T^{-1}I_{c_m} & 0 \\ 0 & I_{n-c_m} \end{bmatrix}.$$

Due to the filtering, $c_{t,\bullet}^{(m)}$ is stationary while $c_{t,u}^{(m)}$ contains the unit root z_m. Then the relevant matrix $\hat{X}_m$ can be written as

$$\hat{X}_m := \langle c_{t-1}^\pi, p_t^\pi\rangle\langle p_t^\pi, p_t^\pi\rangle^{-1}\langle p_t^\pi, c_{t-1}^\pi\rangle\langle c_{t-1}^\pi, c_{t-1}^\pi\rangle^{-1}.$$

Since $p_t = \mathcal{K}\varepsilon_{t-1} + \sum_{j=1,j\neq m}^S \alpha_j\beta_j' X_{t-1}^{(j)} + [0,\tilde{\alpha}_m]c_{t-1}^{(m)}$, we consequently have $p_t^\pi = \mathcal{K}\varepsilon_{t-1}^\pi + [0,\tilde{\alpha}_m]c_{t-1}^\pi$. Therefore, for the three components of $\hat{X}_m$ we obtain with appropriate definitions of the random variables S_m, T_m and using standard asymptotics

$$\begin{aligned} \langle p_t^\pi, p_t^\pi\rangle = \langle \mathcal{K}\varepsilon_{t-1}^\pi + \tilde{\alpha}_m c_{t-1,\bullet}^\pi, \mathcal{K}\varepsilon_{t-1}^\pi + \tilde{\alpha}_m c_{t-1,\bullet}^\pi\rangle &\to \mathcal{K}(\mathbb{E}\varepsilon_{t-1}\varepsilon_{t-1}')\mathcal{K}' + \tilde{\alpha}_m\mathbb{E}c_{t-1,\bullet}^\Pi(c_{t-1,\bullet}^\Pi)'\tilde{\alpha}_m' > 0, \\ \langle p_t^\pi, c_{t-1}^\pi\rangle = \langle \mathcal{K}\varepsilon_{t-1}^\pi + \tilde{\alpha}_m c_{t-1,\bullet}^\pi, c_{t-1}^\pi\rangle &\xrightarrow{d} [S_m, \tilde{\alpha}_m\mathbb{E}c_{t-1,\bullet}^\Pi(c_{t-1,\bullet}^\Pi)'], \\ \langle p_t^\pi, c_{t-1}^\pi\rangle\langle c_{t-1}^\pi, c_{t-1}^\pi\rangle^{-1}\tilde{D}_c^{-1} &= [0,\tilde{\alpha}_m] + \langle \mathcal{K}\varepsilon_{t-1}^\pi, c_{t-1}^\pi\rangle\langle c_{t-1}^\pi, c_{t-1}^\pi\rangle^{-1}\tilde{D}_c^{-1} \\ \langle \mathcal{K}\varepsilon_{t-1}^\pi, c_{t-1}^\pi\rangle\langle c_{t-1}^\pi, c_{t-1}^\pi\rangle^{-1}\tilde{D}_c^{-1} &\xrightarrow{d} [T_m, 0]. \end{aligned}$$

Correspondingly the first block column $\hat{X}_{m,u}$ of $\hat{X}_m$ converges to zero such that $T\hat{X}_{m,u}$ converges in distribution while the second block column converges in probability without normalization. This shows that

$$T\sum_{i=1}^{c_m}\hat{\lambda}_i = T\mathrm{tr}\Big[U_m'(\hat{X}_m - X_m)[U_m - X_m^\dagger(\hat{X}_m - X_m)U_m\Big] + o_P(1) = \mathrm{tr}\big[T\hat{X}_{m,uu}\big] + o_P(1)$$

converges in distribution. The limit is given in [4].

For the case of the estimated state note that the difference between the estimated and the true state is given as

$$\tilde{x}_t - x_t = \tilde{\Gamma}_p'\tilde{z}_{t,p} - \Gamma_p'\tilde{z}_{t,p} - \bar{\mathcal{A}}^p x_{t-p} = (\tilde{\Gamma}_p - \Gamma_p)'\tilde{z}_{t,p} - \bar{\mathcal{A}}^p x_{t-p}.$$

The strict minimum-phase assumption and the assumption on the increase of $p = p(T)$ implies that the second term can be neglected being $o_P(T^{-1})$. Furthermore

$$(\tilde{\Gamma}_p - \Gamma_p)'\tilde{z}_{t,p} = (\tilde{\Gamma}_p - \Gamma_p)'\tilde{D}_z^{-1}\tilde{D}_z\tilde{z}_{t,p}, \quad (\tilde{\Gamma}_p - \Gamma_p)'\tilde{D}_z^{-1} = O_P(T^{-1/2}).$$

Using this it can be concluded that

$$\begin{aligned} \langle \hat{p}_t, \hat{p}_t \rangle &= \langle p_t, p_t \rangle + O_P(T^{-1/2}) \quad , \quad \langle \hat{p}_t, \hat{c}^{(m)}_{t-1,\bullet} \rangle = \langle p_t, c^{(m)}_{t-1,\bullet} \rangle + O_P(T^{-1/2}), \\ \langle \hat{p}_t, \hat{c}^{(m)}_{t-1,u} \rangle &= \langle p_t, c^{(m)}_{t-1,u} \rangle + O_P(T^{-1/2}) \quad , \quad \langle \hat{c}^{(m)}_{t,u}, \hat{c}^{(k)}_{t,u} \rangle = \langle c^{(m)}_{t,u}, c^{(k)}_{t,u} \rangle + O_P(1). \end{aligned}$$

These equations imply that the difference between the expression using the true state and the one using the estimated state converges to zero, implying that the two tests accept and reject jointly asymptotically under the null hypothesis.

References

1. Rodrigues, P.M.; Taylor, A. Alternative estimators and unit root tests for seasonal autoregressive processes. *J. Econom.* **2004**, *120*, 35–73.
2. Johansen, S.; Schaumburg, E. Likelihood Analysis of Seasonal Cointegration. *J. Econom.* **1999**, *88*, 301–339.
3. Hylleberg, S.; Engle, R.; Granger, C.; Yoo, B. Seasonal Integration and Cointegration. *J. Econom.* **1990**, *44*, 215–238.
4. Cubadda, G. Complex Reduced Rank Models For Seasonally Cointegrated Time Series. *Oxf. Bull. Econ. Stat.* **2001**, *63*, 497–511.
5. Cubadda, G.; Omtzigt, P. Small-sample improvements in the statistical analysis of seasonally cointegrated systems. *Comput. Stat. Data Anal.* **2005**, *49*, 333–348.
6. Ahn, S.K.; Cho, S.; Seong, B. Inference of Seasonal Cointegration: Gaussian Reduced Rank Estimation and Tests for Various Types of Cointegration. *Oxford Bull. Econ. Stat.* **2004**, *66*, 261–284.
7. Vivas, E.; Allende-Cid, H.; Salas, R. A Systematic Review of Statistical and Machine Learning Methods for Electrical Power Forecasting with Reported MAPE Score. *Entropy* **2020**, *22*, 1412.
8. García-Martos, C.; Rodríguez, J.; Sánchez, M.J. Forecasting electricity prices and their volatilities using Unobserved Components. *Energy Econ.* **2011**, *33*, 1227–1239.
9. Dufour, J.M.; Stevanović, D. Factor-augmented VARMA models with macroeconomic applications. *J. Bus. Econ. Stat.* **2013**, *31*, 491–506.
10. Dias, G.; Kapetanios, G. Estimation and forecasting in vector autoregressive moving average models for rich datasets. *J. Econom.* **2018**, *202*, 75–91.
11. Foroni, C.; Marcellino, M.; Stevanović. Mixed-frequency models with moving-average components. *J. Appl. Econom.* **2019**, *34*, 688–706.
12. Kascha, C.; Trenkler, C. Simple Identification and specification of cointegrated VARMA models. *J. Appl. Econom.* **2015**, *30*, 675–702.
13. Ravenna, F. Vector autoregressions and reduced form representations of DSGE models. *J. Monet. Econ.* **2007**, *54*, 2048–2064.
14. Komunjer, I.; Zhu, Y. Likelihood ratio testing in linear state space models: An application to dynamic stochastic general equilibrium models. *J. Econom.* **2020**, *218*, 561–586
15. Bauer, D.; Wagner, M. A State Space Canonical Form for Unit Root Processes. *Econom. Theory* **2012**, *28*, 1313–1349.
16. Larimore, W.E. System Identification, reduced order filters and modeling via canonical variate analysis. In Proceedings of the 1983 American Control Conference, San Francisco, CA, USA, 22–24 June 1983; pp. 445–451.
17. Bauer, D. Comparing the `CCA` subspace method to quasi maximum likelihood methods in the case of no exogenous inputs. *J. Time Ser. Anal.* **2006**, *26*, 631–668.
18. Bauer, D. Using Subspace Methodes for Estimating ARMA models for multivariate time series with conditionally heteroskedastic innovations, *Econom. Theory* **2008** , *24*, 1063–1092.
19. Bauer, D. Using Subspace Methods to Model Long-Memory Processes. In *Theory and Applications of Time Series Analysis. ITISE 2018. Contributions to Statistics*; Valenzuela, O., Rojas, F., Pomares, H., Rojas, I., Eds; Springer: Berlin/Heidelberg, Germany, 2019.
20. Bauer, D.; Wagner, M. Estimating Cointegrated Systems Using Subspace Algorithms. *J. Econom.* **2002**, *111*, 47–84.
21. Bauer, D. Estimating linear dynamical systems using subspace methods. *Econom. Theory* **2005**, *21*, 181–211.
22. Hannan, E.J.; Deistler, M. *The Statistical Theory of Linear Systems*; John Wiley: New York, NY, USA, 1998.
23. Chatelin, F. *Eigenvalues of Matrices*; John Wiley & Sons: Hoboken, NJ, USA, 1993.
24. Bauer, D. Asymptotic Distribution of Estimators in Reduced Rank Regression Settings When the Regressors Are Integrated. Technical Report. 2012. AIT. Available online: http://arxiv.org/abs/1211.1439 (accessed on 26 March 2021).
25. Phillips, P.C.B.; Durlauf, S.N. Multiple Time Series Regression with Integrated Processes. *Rev. Econ. Stud.* **1986**, *LIII*, 473–495.
26. Carrasco, M.; Chen, X. Mixing and Moment Properties of Various GARCH and Stochastic Volatility Models. *Econom. Theory* **2002**, *18*, 17–39.
27. Bauer, D.; Wagner, M. *Autoregressive Approximations to MFI(1) Processes*; Technical Report; Department for Mathematical Methods in Economics: TU Wien, Austria, 2004.

28. Bierens, H. Nonparametric cointegration analysis. *J. Econom.* **1997**, *77*, 379–404.
29. Wagner, M. A Comparison of Johansen's, Bierens' and the Subspace Algorithm Method for Cointegration Analysis. *Oxf. Bull. Econ. Stat.* **2004**, *66*, 399–424.
30. Johansen, S.; Nielsen, M. The cointegrated vector autoregressive model with general deterministic terms. *J. Econom.* **2018**, *202*, 214–229.
31. Lee, H. Maximum Likelihood Inference on Cointegration and Seasonal Cointegration. *J. Econom.* **1992**, *54*, 1–47.
32. Bauer, D.; Wagner, M. Using subspace algorithm cointegration analysis: Simulation performance and application to the term structure. *Comput. Stat. Data Anal.* **2009**, *53*, 1954–1973.
33. Bauer, D. Order Estimation for Subspace Methods. *Automatica* **2001**, *37*, 1561–1573.
34. Brockwell, P.J.; Davis, R.A. *Time Series: Theory and Methods*; Springer Series in Statistics, 2. Ed.; Springer: New York, NY, USA, 2006.
35. Qu, Z.; Perron, P. A Modified Information Criterion for Cointegration Tets Based on a VAR Approximation. *Econom. Theory* **2007**, *23*, 638–658.
36. Mulla, R. Hourly Energy Consumption. Available online: www.kaggle.com/robikscube/hourly-energy-consumption/ (accessed on 22 January 2021).
37. del Barrio, Castro, T.; Rodrigues, P.M.M.; Taylor, A.M.R. Temporal Aggregation of Seasonally Near-Integrated Processes. *J. Time Ser. Anal.* **2019**, *40*, 872–886.
38. Bauer, D. Almost sure bounds on the estimation error for ols estimators when the regressors include certain MFI(1) processes. *Econom. Theory* **2009**, *25*, 571–582.
39. Ahn, S.; Reinsel, G. Estimation of Partially Nonstationary Vector Autoregressive Models with Seasonal Behaviour. *J. Econom.* **1994** *62*, 317–350.
40. Bauer, D.; Deistler, M.; Scherrer, W. Consistency and Asymptotic Normality of some Subspace Algorithms for Systems Without Observed Inputs. *Automatica* **1999**, *35*, 1243–1254.