EUROPEAN UNIVERSITY INSTITUTE
DEPARTMENT OF ECONOMICS

EUI Working Paper **ECO** No. 2005 /9

Autoregressive Approximations of Multiple Frequency I(1) Processes

DIETMAR BAUER
and
MARTIN WAGNER

BADIA FIESOLANA, SAN DOMENICO (FI)

Published in Italy in June 2005
European University Institute
Badia Fiesolana
I-50016 San Domenico (FI)
Italy

Autoregressive Approximations of Multiple Frequency I(1) Processes*

Dietmar Bauer
TU Wien
Institute for Mathematical
Methods in Economics

Martin Wagner†
University of Bern
Department of Economics
and
European University Institute

June 16, 2005

Abstract

We investigate autoregressive approximations of multiple frequency I(1) processes, of which I(1) processes are a special class. The underlying data generating process is assumed to allow for an infinite order autoregressive representation where the coefficients of the Wold representation of the suitably differenced process satisfy mild summability constraints. An important special case of this process class are VARMA processes. The main results link the approximation properties of autoregressions for the nonstationary multiple frequency I(1) process to the corresponding properties of a related stationary process, which are well known (cf. Section 7.4 of Hannan and Deistler, 1988). First, error bounds on the estimators of the autoregressive coefficients are derived that hold uniformly in the lag length. Second, the asymptotic properties of order estimators obtained with information criteria are shown to be closely related to those for the associated stationary process obtained by suitable differencing. For multiple frequency I(1) VARMA processes we establish divergence of order estimators based on the `BIC` criterion at a rate proportional to the logarithm of the sample size.

JEL Classification: C13, C32

Keywords: Unit Roots, Multiple Frequency I(1) Process, Nonrational Transfer Function, Cointegration, VARMA Process, Information Criteria

---

*Part of this work has been carried out while the first author held a post-doc position at the Cowles Foundation, Yale University, financed by the Max Kade Foundation and the second author was visiting the Economics Departments of Princeton University and the European University Institute in Florence. Support from these institutions is gratefully acknowledged.

†Correspondence to: Tel: ++39 055 4685 936, Fax: ++39 055 4685 902, email: `Martin.Wagner@iue.it`.

1 Introduction

This paper considers unit root processes that admit an infinite order autoregressive representation where the autoregression coefficients satisfy mild summability constraints. More precisely the class of multiple frequency I(1) vector processes is analyzed. Following Bauer and Wagner (2004) a process is called multiple frequency I(1), briefly MFI(1), if the integration orders corresponding to all unit roots are equal to one and certain restrictions on the deterministic components are fulfilled (for details see Definition 2 in Section 2). Processes with seasonal unit roots with integration orders equal to one fall into this class, as do I(1) processes (where in both cases certain restrictions on the deterministic terms have to be fulfilled, see below).

VARMA processes are a leading example of the class of processes considered in this paper. However, the analysis is not restricted to VARMA processes, since we do not restrict the analysis to rational transfer functions. On the other hand long memory processes (e.g. fractionally integrated processes) are not contained in the discussion.

Finite order vector autoregressions are probably the most prominent model in time series econometrics and especially so in the analysis of integrated and cointegrated time series. The limiting distribution of least squares estimators for this model class is well known, both for the stationary case as well as for the MFI(1) case, see i.a. Lai and Wei (1982), Lai and Wei (1983), Chan and Wei (1988), Johansen (1995) or Johansen and Schaumburg (1999). Also model selection issues in this context are well understood, see e.g. Pötscher (1989) or Johansen (1995).

In the stationary case finite order vector autoregressions have been extended to more general processes by letting the order tend to infinity as a function of the sample size and certain characteristics of the true system. In this respect the paper of Lewis and Reinsel (1985) is one of the earliest examples. The properties of lag length selection using information criteria in this situation are well understood. Section 7.4 of Hannan and Deistler (1988), referred to as HD henceforth, collects many results in this respect: First, error bounds that hold uniformly in the lag length are presented for the estimated autoregressive coefficient matrices. Second, the asymptotic properties of information criteria in this misspecified situation (in the sense that no finite autoregressive representation exists) are discussed in a rather general setting.

In the I(1) case autoregressive approximations have been studied i.a. in Saikkonen (1992), Saikkonen (1993) and Saikkonen and Lütkepohl (1996). Here the first two papers derive the asymptotic properties of the estimated cointegrating space and the third one develops the asymptotic theory for all autoregressive coefficients. In these three papers, analogously to Lewis and Reinsel (1985), a lower bound on the increase of the lag length is imposed. This lower bound depends on characteristics of the true data generating process. Saikkonen and Luukkonen (1997) show that for the asymptotic validity of the Johansen testing procedures for the cointegrating rank this lower bound is not needed. Instead only convergence to infinity is needed.

For the seasonal integration case analogous results on the properties of autoregressive approximations and the behavior of tests developed for the finite order autoregressive case in the case of approximating an infinite order VAR process do not seem to be available in the literature. It is one aim of this paper to contribute to this area.

In most papers dealing with autoregressive approximations the order of the autoregression is assumed to increase within bounds that are a function of the sample size and typically the lower bounds are dependent upon system quantities that are unknown prior to estimation, see e.g. Assumption (iii) in Theorem 2 of Lewis and Reinsel (1985). In practice the autoregressive order is typically estimated using information criteria. The properties of the corresponding order estimators are well known in the stationary case, see again Section 7.4 of HD. For the I(1) and MFI(1) cases, however, knowledge seems to be sparse and partially incorrect: Ng and Perron (1995) discuss order estimation with information criteria for univariate I(1) ARMA processes. Unfortunately (as noticed in Lütkepohl and Saikkonen, 1999, Section 5) their Lemma 4.2 is not strong enough to support their conclusion that for typical choices of the penalty factor the behavior of the order estimator based on minimizing information criteria is identical to the behavior of the order estimator for the (stationary) differenced process, since they only show that the difference between the two information criteria (for the original data and for the differenced data) for given lag length is of order $o_p(T^{-1/2})$, whereas the penalty term in the information criterion is proportional to $C_T T^{-1}$, where usually $C_T = 2$ (AIC) or $C_T = \log T$ (BIC) is used. The asymptotic properties of order estimators based on information criteria are typically derived by showing that asymptotically the penalty term dominates the estimation error. This allows to write the information criterion as the sum of a deterministic function $\tilde{L}_T$ (that depends upon the order and the penalty term C_T, see p. 333 of HD for a definition) and a comparatively small estimation error. Subsequently, the asymptotic properties of the order estimator are linked to the minimizer of the deterministic function (see HD, Section 7.4, p. 333, for details). In order to show asymptotic equivalence of lag length selection based only upon the deterministic function and the information criterion, therefore an $O_p(C_T T^{-1})$ bound that holds uniformly in the lag length has to be obtained for the estimation error. A similar problem occurs in Lemma 5.1 of Lütkepohl and Saikkonen (1999), where only a bound of order $o_p(K_T/T)$ is derived, with $K_T = o(T^{1/3})$ denoting the upper bound for the autoregressive lag length. Again this bound on the error is not strong enough to show asymptotic equivalence of the order estimator based on the nonstationary process with the order estimator based on the associated stationary process for typical penalty factors C_T.

This paper extends the available theory in two ways: First the estimation error in autoregressive approximations is shown to be of order $O_P((\log T/T)^{1/2})$ uniformly in the lag length for a moderately large upper bound on the lag length given by $H_T = o((T/\log T)^{1/2})$. This result extends Theorem 7.4.5 of HD, p. 331, to the case of MFI(1) processes. Based upon this result we show in a second step that the information criteria applied to the untransformed process have (in probability) the same behavior as the information criteria applied to a suitably differenced stationary process. This on the one hand provides a rigorous proof for the fact already stated for univariate I(1) processes in Ng and Perron (1995) and on the other hand extends the results from the I(1) case to the MFI(1) case. In particular in the VARMA case it follows that the `BIC` order estimate increases proportionally to $\log T$ to infinity.

The paper is organized as follows: In Section 2 some basic definitions, assumptions and the class of processes considered are presented. Section 3 discusses autoregressive approximations for stationary processes. The main results for MFI(1) processes are stated in Section 4 and Section 5 briefly summarizes and concludes. Two appendices follow the main text. In Appendix A several useful lemmata are collected and Appendix B contains the proofs of the theorems.

Throughout the paper we use the notation $F_T = o(g_T)$ for a random matrix sequence

$F_T \in \mathbb{R}^{a_T \times b_T}$ if $\lim_{T\to\infty} \max_{1\leq i\leq a_T, 1\leq j\leq b_T} |F_{i,j,T}|/g_T = 0$ a.s., where $F_{i,j,T}$ denotes the (i,j)-th entry of F_T. Also $F_T = O(g_T)$ means $\limsup_{T\to\infty} \max_{1\leq i\leq a_T, 1\leq j\leq b_T} |F_{i,j,T}|/g_T < M < \infty$ a.s. for some constant M. Analogously $F_T = o_P(g_T)$ means that $\max_{1\leq i\leq a_T, 1\leq j\leq b_T} |F_{i,j,T}|/g_T$ converges to zero in probability and $F_T = O_P(g_T)$ means that for each $\varepsilon > 0$ there exists a constant $M(\varepsilon) < \infty$ such that $\mathbb{P}\{\max_{1\leq i\leq a_T, 1\leq j\leq b_T} |F_{i,j,T}|/g_T > M(\varepsilon)\} \leq \varepsilon$. Note that this definition differs from the usual conventions in that the maximum entry rather than the 2-norm is considered. In case that the dimensions of F_T tend to infinity this may make a difference since norms are not necessarily equivalent in infinite dimensional spaces. We furthermore use $\langle a_t, b_t\rangle_i^{T-j} := \frac{1}{T}\sum_{t=i}^{T-j} a_t b_t'$, where we use for simplicity the same symbol for both the processes $(a_t)_{t\in\mathbb{Z}}$, $(b_t)_{t\in\mathbb{Z}}$ and the vectors a_t and b_t. For simplicity we use $\langle a_t, b_t\rangle := \langle a_t, b_t\rangle_{p+1}^T$, when used in the context of autoregressions of order p.

2 Definitions and Assumptions

In this paper we are interested in real valued multivariate unit root processes $(y_t)_{t\in\mathbb{Z}}$ with $y_t \in \mathbb{R}^s$. Let us define the difference operator at frequency ω as:

$$\Delta_\omega(L) := \begin{cases} 1 - e^{i\omega}L, & \omega \in \{0, \pi\} \\ (1 - e^{i\omega}L)(1 - e^{-i\omega}L), & \omega \in (0, \pi). \end{cases} \tag{1}$$

Here L denotes the backward-shift operator such that $L(y_t)_{t\in\mathbb{Z}} = (y_{t-1})_{t\in\mathbb{Z}}$. Somewhat sloppily we also use the notation $Ly_t = y_{t-1}$. Consequently for example $\Delta_\omega(L)y_t = y_t - 2\cos(\omega)y_{t-1} + y_{t-2}, t \in \mathbb{Z}$ for $\omega \in (0, \pi)$. In the definition of $\Delta_\omega(L)$ complex roots $e^{i\omega}, \omega \in (0, \pi)$ are taken in pairs of complex conjugate roots in order to ensure real valuedness of the filtered process $\Delta_\omega(L)(y_t)_{t\in\mathbb{Z}}$ for real valued $(y_t)_{t\in\mathbb{Z}}$. For stable transfer functions we use the notation $v_t = c(L)\varepsilon_t = \sum_{j=0}^\infty c_j \varepsilon_{t-j}$. We also formally use polynomials in the backward-shift operator applied to matrices such that $c(A) = \sum_{j=0}^p c_j A^j$ for a polynomial $c(L) = \sum_{j=0}^p c_j L^j$. Using this notation we define a unit root process as follows:

Definition 1 *The s-dimensional real process $(y_t)_{t\in\mathbb{Z}}$ has unit root structure*

$$\Omega := ((\omega_1, h_1), \ldots, (\omega_l, h_l))$$

with $0 \leq \omega_1 < \omega_2 < \ldots < \omega_l \leq \pi, h_k \in \mathbb{N}, k = 1, \ldots, l$, if with $D(L) := \Delta_{\omega_1}^{h_1}(L)\cdots\Delta_{\omega_l}^{h_l}(L)$ it holds that

$$D(L)(y_t - T_t) = v_t, \quad t \in \mathbb{Z} \tag{2}$$

for $v_t = \sum_{j=0}^{\infty} c_j \varepsilon_{t-j}, c_j \in \mathbb{R}^{s\times s}, j \geq 0$, *corresponding to the Wold representation of the stationary process* $(v_t)_{t\in\mathbb{Z}}$*, where for* $c(z) := \sum_{j=0}^{\infty} c_j z^j, z \in \mathbb{C}$ *with* $\sum_{j=0}^{\infty} \|c_j\| < \infty$ *it holds that* $c(e^{i\omega_k}) \neq 0$ *for* $k = 1, \ldots, l$*. Here* $(\varepsilon_t)_{t\in\mathbb{Z}}, \varepsilon_t \in \mathbb{R}^s$ *is assumed to be a zero mean weak white noise process with finite variance* $0 < \mathbb{E}\varepsilon_t\varepsilon_t' < \infty$*. Further* $(T_t)_{t\in\mathbb{Z}}$ *is a deterministic process.*

The s*-dimensional real process* $(y_t)_{t\in\mathbb{Z}}$ *has empty unit root structure* $\Omega_0 := \{\}$ *if there exists a deterministic process* $(T_t)_{t\in\mathbb{Z}}$ *such that* $(y_t - T_t)_{t\in\mathbb{Z}}$ *is weakly stationary.*

A process that has a non-empty unit root structure is called a unit root process*. If furthermore* $c(z)$ *is a rational function of* $z \in \mathbb{C}$ *then* $(y_t)_{t\in\mathbb{Z}}$ *is called a* rational unit root process*.*

See Bauer and Wagner (2004) for a detailed discussion of the arguments underlying this definition. We next define an MFI(1) process as follows:

Definition 2 *A real valued process with unit root structure* $((\omega_1, 1), \ldots, (\omega_l, 1))$ *and* $(T_t)_{t\in\mathbb{Z}}$ *solving* $\Pi_{i=1}^{l} \Delta_{\omega_i}(L) T_t = 0$ *is called multiple frequency I(1) process, or short MFI(1) process.*

Note as already indicated in the introduction that the definition of an MFI(1) process places restrictions on the deterministic process $(T_t)_{t\in\mathbb{Z}}$. E.g. in the I(1) case (when the only unit root in the above definition occurs at frequency zero) the definition guarantees that the first difference of the process is stationary. Thus, e.g. I(1) processes are a subset of processes with unit root structure $((0, 1))$. For the results in this paper some further assumptions are required on both the function $c(z)$ of Definition 1 and the process $(\varepsilon_t)_{t\in\mathbb{Z}}$.

Assumption 1 *The real valued process* $(y_t)_{t\in\mathbb{Z}}$ *is a solution to the difference equation*

$$D(L)y_t = \Delta_{\omega_1}(L) \cdots \Delta_{\omega_l}(L) y_t = v_t, \quad t \in \mathbb{Z} \tag{3}$$

where $v_t = \sum_{j=0}^{\infty} c_j \varepsilon_{t-1}$ *corresponds to the Wold decomposition and it holds, with* $c(z) = \sum_{j=0}^{\infty} c_j z^j$*, that* $\det c(z) \neq 0$ *for all* $|z| \leq 1$ *except possibly for* $z_k := e^{i\omega_k}$*,* $k = 1, \ldots, l$*. Here* $D(L)$ *corresponds to the unit root structure and is given as in Definition 1. Further* $\sum_{j=0}^{\infty} j^{3/2+H} \|c_j\| < \infty$*, with* $H := \sum_{k=1}^{l}(1 + \mathbb{I}(0 < \omega_k < \pi))$*, where* $\mathbb{I}$ *denotes the indicator function.*

Assumption 2 *The stochastic process* $(\varepsilon_t)_{t\in\mathbb{Z}}$ *is a strictly stationary ergodic martingale difference sequence with respect to the* σ*-algebra* $\mathcal{F}_t = \sigma\{\varepsilon_t, \varepsilon_{t-1}, \varepsilon_{t-2}, \ldots\}$. *Additionally the following assumptions hold:*

$$\begin{array}{lll} \mathbb{E}\{\varepsilon_t \mid \mathcal{F}_{t-1}\} = 0 & , & \mathbb{E}\{\varepsilon_t\varepsilon_t' \mid \mathcal{F}_{t-1}\} = \mathbb{E}\varepsilon_t\varepsilon_t' = \Sigma > 0, \\ \mathbb{E}\varepsilon_{t,j}^4 \log_+(|\varepsilon_{t,j}|) < \infty & , & j = 1, \ldots, s, \end{array} \tag{4}$$

where $\varepsilon_{t,j}$ *denotes the* j*-th coordinate of the vector* ε_t *and* $\log_+(x) = \log(\max(x, 1))$.

The assumptions on $(\varepsilon_t)_{t\in\mathbb{Z}}$ follow Hannan and Kavalieris (1986), see also the discussion in Section 7.4 of HD. They exclude conditionally heteroskedastic innovations. It appears possible to relax the assumptions in this direction, but these extensions are not in the scope of this paper.

The assumptions on the function $c(z)$ formulated in Assumption 1 are based on the assumptions formulated in Section 7.4 of HD for stationary processes. However, the allowed nonstationarities require stronger summability assumptions (see also Stock and Watson, 1988, Assumption A(ii), p. 787). These stronger summability assumptions guarantee that the stationary part of the process (see Theorem 1 for a definition) fulfills the summability requirements formulated in HD.

In the following Theorem 1 a convenient representation of the processes fulfilling Assumption 1 is derived. The result is similar in spirit to the discussion in Section 2 of Sims *et al.* (1990), who discuss unit root processes with unit root structure $((0, h))$ with $h \in \mathbb{N}$.

Theorem 1 *Let* $(y_t)_{t\in\mathbb{Z}}$ *be a process fulfilling Assumption 1. Denote with* $\tilde{c}_k$ *the rank (over* $\mathbb{C}$*) of the matrix* $c(e^{i\omega_k}) \in \mathbb{C}^{s\times s}$ *and let* $c_k := \tilde{c}_k(1 + \mathbb{I}(0 < \omega_k < \pi))$. *Further let*

$$J_k := \left\{ \begin{array}{rl} I_{c_k} & , \quad \omega_k = 0, \\ -I_{c_k} & , \quad \omega_k = \pi, \\ S_k \otimes I_{\tilde{c}_k} & , \quad \textit{else}, \end{array} \right. \quad \textit{with} \quad S_k := \left[\begin{array}{cc} \cos\omega_k & \sin\omega_k \\ -\sin\omega_k & \cos\omega_k \end{array} \right]. \tag{5}$$

Then there exist matrices $C_k \in \mathbb{R}^{s\times c_k}, K_k \in \mathbb{R}^{c_k\times s}, k = 1, \ldots, l$ *such that the state space systems* (J_k, K_k, C_k) *are minimal (see p. 47 of HD for a definition) and a transfer function* $c_\bullet(z) = \sum_{j=0}^\infty c_{j,\bullet} z^j, \sum_{j=0}^\infty j^{3/2}\|c_{j,\bullet}\| < \infty, \det c_\bullet(z) \neq 0, |z| < 1$ *such that with* $x_{t+1,k} = J_k x_{t,k} + K_k\varepsilon_t \in \mathbb{R}^{c_k}$, *there exists a process* $(y_{t,h})_{t\in\mathbb{Z}}$ *where* $D(L)(y_{t,h})_{t\in\mathbb{Z}} \equiv 0$ *such that* $y_t = \sum_{k=1}^l C_k x_{t,k} + \sum_{j=0}^\infty c_{j,\bullet}\varepsilon_{t-j} + y_{t,h} = \tilde{y}_t + y_{t,h}$, *where this equation defines the process* $(\tilde{y}_t)_{t\in\mathbb{Z}}$.

Proof: The proof centers around the representation for $c(z)$ given in Lemma 2 in Appendix A. In the proof we show that for appropriate choice of $c_\bullet(z)$ fulfilling the assumptions the corresponding process $(\tilde{y}_t)_{t\in\mathbb{Z}}$ defined above is a solution to the difference equation $D(L)\tilde{y}_t = v_t$. Once that is established $D(L)y_{t,h} = D(L)(y_t - \tilde{y}_t) = 0$ then proves the theorem. Therefore consider $\tilde{y}_t = \sum_{k=1}^{l} C_k x_{t,k} + \sum_{j=0}^{\infty} c_{j,\bullet}\varepsilon_{t-j}$. Note that for $0 < \omega_k < \pi$

$$\begin{aligned}(1 - 2\cos(\omega_k)L + L^2)x_{t,k} &= J_k x_{t-1,k} + K_k\varepsilon_{t-1} - 2\cos(\omega_k)(J_k x_{t-2,k} + K_k\varepsilon_{t-2}) + x_{t-2,k} \\ &= (J_k^2 - 2\cos(\omega_k)J_k + I_{c_k})x_{t-2,k} + K_k\varepsilon_{t-1} + (J_k - 2\cos(\omega_k)I_{c_k})K_k\varepsilon_{t-2} \\ &= K_k\varepsilon_{t-1} - J_k' K_k\varepsilon_{t-2}\end{aligned}$$

using $I_{c_k} - 2\cos(\omega_k)J_k + J_k^2 = 0$ and $-J_k' = J_k - 2\cos(\omega_k)I_{c_k}$. Then for $t \geq 1$

$$D(L)x_{t,k} = D_{\neg k}(L)\Delta_{\omega_k}(L)x_{t,k} = D_{\neg k}(L)(K_k\varepsilon_{t-1} - J_k' K_k\varepsilon_{t-2}\mathbb{I}(\omega_k \notin \{0,\pi\})$$

with $D_{\neg k}(L) = D(L)/\Delta_{\omega_k}(L)$. For $\omega_k \in \{0,\pi\}$ simpler evaluations give $x_{t,k} - \cos(\omega_k)x_{t-1,k} = K_k\varepsilon_{t-1}$. Therefore for $t \geq 1$

$$D(L)\tilde{y}_t = \sum_{j=1}^{l} C_k D_{\neg k}(L)\left[K_k\varepsilon_{t-1} - J_k' K_k\varepsilon_{t-2}\mathbb{I}(\omega_k \notin \{0,\pi\})\right] + D(L)c_\bullet(L)\varepsilon_t = c(L)\varepsilon_t$$

where the representation of $c(z)$ given in Lemma 2 is used to define $c_\bullet(z)$ and to verify its properties. Therefore $(\tilde{y}_t)_{t\in\mathbb{Z}}$ solves the difference equation $D(L)\tilde{y}_t = v_t$. □

This theorem is a key ingredient for the subsequent results. It provides a representation of the process as the sum of two components. The nonstationary part of $(\tilde{y}_t)_{t\in\mathbb{Z}}$ is a linear function of the building blocks $(x_{t,k})_{t\in\mathbb{Z}}$, which have unit root structure $((\omega_k, 1))$ and are not cointegrated due to the connection between the rank of $c(e^{i\omega_k})$ and the dimension of K_k. If $c(z)$ is rational the representation is related to the canonical form given in Bauer and Wagner (2004). In the I(1) case this corresponds to a Granger type representation.

Note that the representation given in Theorem 1 is not unique. This can be seen as follows, where we consider only complex unit roots, noting that the case of real unit roots is simpler: All solutions to the homogenous equation $D(L)y_{t,h} = 0$ are of the form $y_{t,h} = \sum_{k=1}^{l} D_{k,c}\cos(\omega_k t) + D_{k,s}\sin(\omega_k t)$ where $D_{k,s} = 0$ for $\omega_k \in \{0,\pi\}$. The processes $(d_{t,k,1})_{t\in\mathbb{Z}} = ([-\sin(\omega_k t), \cos(\omega_k t)]')_{t\in\mathbb{Z}}$ and $(d_{t,k,2})_{t\in\mathbb{Z}} = ([\cos(\omega_k t), \sin(\omega_k t)]')_{t\in\mathbb{Z}}$ are easily seen to span the set of all solutions to the homogeneous equation $d_{t,k} = S_k d_{t-1,k}$ for $\omega_k \notin \{0, \pi\}$. If for $C_k = [C_{k,c}, C_{k,s}]$ with $C_{k,c}, C_{k,s} \in \mathbb{R}^{s \times \tilde{c}_k}$ we have

$$\begin{bmatrix} D_{k,c} \\ D_{k,s} \end{bmatrix} = \begin{bmatrix} C_{k,c} & C_{k,s} \\ -C_{k,s} & C_{k,c} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix}$$

for $\alpha_i \in \mathbb{R}^{\tilde{c}_k \times 1}, i = 1, 2$, it follows that in the representation of $(y_t)_{t\in\mathbb{Z}}$ given in Theorem 1 there exist processes $(x_{t,k})_{t\in\mathbb{Z}}$ such that the corresponding $(y_{t,h})_{t\in\mathbb{Z}} \equiv 0$. In this case there is no need to model the deterministic components explicitly. Otherwise the model has to account for deterministic terms. These two cases are considered separately.

Assumption 3 *Let $(y_t)_{t\in\mathbb{Z}}$ be generated according to Assumption 1, then we distinguish two (nonexclusive) cases:*

(i) There exists a representation of $(y_t)_{t\in\mathbb{Z}}$ of the form $y_t = \tilde{y}_t + y_{t,h}$ as given in Theorem 1, such that $(y_{t,h})_{t\in\mathbb{Z}} \equiv 0$.

(ii) It holds that $y_t = \tilde{y}_t + T_t$, where $\tilde{y}_t$ is as in Theorem 1 and $(T_t)_{t\in\mathbb{Z}}$ is a deterministic process such that $D(L)(T_t)_{t\in\mathbb{Z}} \equiv 0$.

Note that the decomposition of $(y_t)_{t\in\mathbb{Z}}$ into $(\tilde{y}_t)_{t\in\mathbb{Z}}$ and $(T_t)_{t\in\mathbb{Z}}$ also is not unique due to non-identifiability with the processes $(x_{t,k})_{t\in\mathbb{Z}}$ as documented above. In particular $(T_t)_{t\in\mathbb{Z}}$ of the above assumption does not necessarily coincide with the process $(T_t)_{t\in\mathbb{Z}}$ as given in Definition 1.

Remark 1 *The restriction $D(L)(T_t)_{t\in\mathbb{Z}} \equiv 0$ is not essential for the results in this paper. Harmonic components of the form $([A\sin(\omega t), B\cos(\omega t)]')_{t\in\mathbb{Z}}$ with arbitrary frequency ω could be included. We refrain from discussing this possibility separately in detail.*

3 Autoregressive Approximations of Stationary Processes

We recall in this section the approximation results for stationary processes that build the basis for our extension to the MFI(1) case. The source of these results is Section 7.4 of HD, where however the Yule-Walker (YW) estimator of the autoregression is considered, whereas we consider the least squares (LS) estimator in this paper, see below. This necessitates to show that the relevant results also apply to the LS estimator (which are collected in Theorem 2).

In this section we consider autoregressive approximations of order p for $(v_t)_{t\in\mathbb{Z}}$ defined as (ignoring the mean and harmonic components for simplicity)

$$u_t(p) := v_t + \Phi_p^v(1)v_{t-1} + \ldots + \Phi_p^v(p)v_{t-p}.$$

Here the coefficient matrices $\Phi_p^v(j)$, $j = 1,\ldots,p$ are chosen such that $u_t(p)$ has minimum variance. Both the coefficient matrices $\Phi_p^v(j)$ and their YW estimators $\tilde{\Phi}_p^v(j)$ are defined from the Yule-Walker equations given below: Define the sample covariances as $G^v(j) := \langle v_t, v_{t-j}\rangle_{j+1}^T$ for $0 \le j < T$, $G^v(j) := G^v(-j)'$ for $-T < j < 0$ and $G^v(j) := 0$ else. We denote their population counterparts with $\Gamma^v(j) := \mathbb{E}v_t v_{t-j}'$. Then $\Phi_p^v(j)$ and $\tilde{\Phi}_p^v(j)$ are defined as the solutions to the respective YW equations (where $\Phi_p^v(0) = I_s, \tilde{\Phi}_p^v(0) = I_s$):

$$\begin{aligned}\sum_{j=0}^{p} \Phi_p^v(j)\Gamma^v(j-i) &= 0, \quad i = 1,\ldots,p,\\ \sum_{j=0}^{p} \tilde{\Phi}_p^v(j)G^v(j-i) &= 0, \quad i = 1,\ldots,p.\end{aligned}$$

The infinite order Yule-Walker equations and the corresponding autoregressive coefficient matrices are defined from (the existence of these solutions follows from the assumptions on the process imposed in this paper):

$$\sum_{j=0}^{\infty} \Phi^v(j)\Gamma^v(j-i) = 0, \quad i = 1,\ldots,\infty.$$

It appears unavoidable that notation becomes a bit heavy, thus let us indicate the underlying logic here. Throughout, superscripts refer to the variable under investigation and the subscripts indicate the autoregressive lag length, as already used for the coefficient matrices $\Phi_p^v(j)$ above. If no subscript is added, the quantities correspond to the infinite order autoregressions.

As indicated we focus on the LS estimator in this paper. Using the regressor vector $V_{t,p}^- := [v_{t-1}', \ldots, v_{t-p}']'$ for $t = p+1,\ldots,T$, the LS estimator, $\hat{\Theta}_p^v$, is defined by

$$\hat{\Theta}_p^v := -\left[\hat{\Phi}_p^v(1),\ldots,\hat{\Phi}_p^v(p)\right] := \langle v_t, V_{t,p}^-\rangle\langle V_{t,p}^-, V_{t,p}^-\rangle^{-1},$$

where this equation defines the LS estimators $\hat{\Phi}_p^v(j)$, $j = 1, \ldots, p$ of the autoregressive coefficient matrices. Define furthermore for $1 \leq p \leq H_T$ (with $\hat{\Sigma}_0^v := G^v(0)$):

$$\hat{\Sigma}_p^v := \langle v_t - \hat{\Theta}_p V_{t,p}^-, v_t - \hat{\Theta}_p V_{t,p}^- \rangle, \quad \Sigma_p^v := \sum_{j=0}^{p} \Phi_p^v(j)\Gamma^v(j)$$

and note the following identity for the covariance matrix of $(\varepsilon_t)_{t\in\mathbb{Z}}$ provided the infinite sum exists which will always be the case in our setting:

$$\Sigma = \mathbb{E}\varepsilon_t\varepsilon_t' = \sum_{j=0}^{\infty} \Phi^v(j)\Gamma^v(j).$$

Thus, $\hat{\Sigma}_p^v$ denotes the estimated one-step ahead prediction error. The lag lengths p are considered in the interval $0 \leq p \leq H_T$, where $H_T = o((T/\log T)^{1/2})$. Lag length selection over $0 \leq p \leq H_T$, when based on information criteria (see Akaike, 1975) is based on the quantities just defined and an 'appropriately' chosen penalty factor C_T. These elements are combined in the following criterion function:

$$IC^v(p; C_T) := \log\det\hat{\Sigma}_p^v + ps^2\frac{C_T}{T}, \quad 0 \leq p \leq H_T \tag{6}$$

where ps^2 is the number of parameters contained in $\hat{\Theta}_p^v$. Setting $C_T = 2$ results in AIC and $C_T = \log T$ is used in BIC. For given C_T the estimated order, $\hat{p}$ say, is given by the smallest minimizing argument of $IC^v(p; C_T)$, i.e.

$$\hat{p} := \min\left(\arg\min_{0\leq p\leq H_T} IC^v(p; C_T)\right). \tag{7}$$

Section 7.4 of HD contains many relevant results concerning the asymptotic properties of autoregressive approximations and information criteria. These results build the basis for the results of this paper. Assumption 1 on $(c(L)\varepsilon_t)_{t\in\mathbb{Z}}$ is closely related to the assumptions formulated in Section 7.4 of HD. In particular HD require that the transfer function $c(z) = \sum_{j=0}^{\infty} c_j z^j$ is such that $\sum_{j=0}^{\infty} j^{1/2}\|c_j\| < \infty$ and $\det c(z) \neq 0$ for all $|z| \leq 1$. However, for technical reasons in the MFI(1) case we need stronger summability assumptions on $c(z)$, see Lemma 3. In the important special case of MFI(1) VARMA processes these summability assumptions are clearly fulfilled. Theorem 2 below presents the results required for our paper for the LS estimator. Note again that the results in HD are for the YW estimator. The proof of the theorem is given in Appendix B.

Theorem 2 *Let $(v_t)_{t\in\mathbb{Z}}$ be generated according to $v_t = c(L)\varepsilon_t$, with $c(z) = \sum_{j=0}^{\infty} c_j z^j$, $c_0 = I_s$, where it holds that $\sum_{j=0}^{\infty} j^{1/2}\|c_j\| < \infty$, $\det c(z) \neq 0, |z| \leq 1$ and $(\varepsilon_t)_{t\in\mathbb{Z}}$ fulfills Assumption 2. Then the following statements hold:*

(i) For $1 \leq p \leq H_T$, with $H_T = o((T/\log T)^{1/2})$, it holds uniformly in p that

$$\max_{1\leq j\leq p} \|\hat{\Phi}_p^v(j) - \Phi_p^v(j)\| = O((\log T/T)^{1/2}).$$

(ii) For rational $c(z)$ the above bound can be sharpened to $O((\log\log T/T)^{1/2})$ for $1 \leq p \leq G_T$, with $G_T = (\log T)^a$ for any $a < \infty$.

(iii) If $(v_t)_{t\in\mathbb{Z}}$ is not generated by a finite order autoregression, i.e. if there exists no p_0 such that $\Phi^v(j) = 0$ for all $j > p_0$, then the following statements hold:

- *For $C_T/\log T \to \infty$ it holds that*

$$IC^v(p; C_T) = \log\det\dot{\Sigma} + \left\{\frac{ps^2}{T}(C_T - 1) + tr\left[\Sigma^{-1}(\Sigma_p^v - \Sigma)\right]\right\}\{1 + o(1)\},$$

  with $\dot{\Sigma} := T^{-1}\sum_{t=1}^{T}\varepsilon_t\varepsilon_t'$ and the approximation error is $o(1)$ uniformly in $0 \leq p \leq H_T$.

- *For $C_T \geq c > 1$ the same approximation holds with the $o(1)$ term replaced by $o_P(1)$.*

(iv) For rational $c(z)$ let $c(z) = a^{-1}(z)b(z)$ be a matrix fraction decomposition where $(a(z), b(z))$ are left coprime matrix polynomials $a(z) = \sum_{j=0}^{m} A_j z^j$, $A_0 = I_s$, $A_m \neq 0$, $b(z) = \sum_{j=0}^{n} B_j z^j$, $B_0 = I_s$, $B_n \neq 0, n > 0$ and $\det a(z) \neq 0$, $\det b(z) \neq 0$ for $|z| \leq 1$. Denote with $\rho_0 > 1$ the smallest modulus of the zeros of $\det b(z)$ and with $\hat{p}_{BIC}$ the smallest minimizing argument of $IC^v(p; \log T)$ for $0 \leq p \leq G_T$. Then it holds that

$$\lim_{T\to\infty} \frac{2\hat{p}_{BIC}\log\rho_0}{\log T} = 1 \quad a.s.$$

(v) Let $\tilde{P}_s \in \mathbb{R}^{r\times s}$, $r \leq s$ denote a selector matrix, i.e. a matrix composed of r rows of I_s. Then, if the autoregression of order $p-1$ is augmented by the regressor $\tilde{P}_s v_{t-p}$

results (i) to (iv) continue to hold, when the approximation to $IC^v(p;C_T)$ presented in (iii) is replaced by:

$$\begin{aligned}\widetilde{IC}^v(p;C_T) &\leq \log\det\dot{\Sigma} + \left\{\frac{ps^2}{T}(C_T-1) + tr\left[\Sigma^{-1}(\Sigma^v_{p-1}-\Sigma)\right]\right\}\{1+o(1)\},\\ \widetilde{IC}^v(p;C_T) &\geq \log\det\dot{\Sigma} + \left\{\frac{ps^2}{T}(C_T-1) + tr\left[\Sigma^{-1}(\Sigma^v_{p}-\Sigma)\right]\right\}\{1+o(1)\}\end{aligned}$$

for $C_T/\log T \to \infty$. Again for $C_T \geq c > 1$ the result holds with the $o(1)$ term replaced by $o_P(1)$. Here $\widetilde{IC}^v(p;C_T)$ denotes the information criterion from the regression of order $p-1$ augmented by $\tilde{P}_s v_{t-p}$.

(vi) All results formulated in (i) to (v) remain valid, if

$$v_t = c(L)\varepsilon_t + \sum_{k=1}^{l}(D_{k,c}\cos(\omega_k t) + D_{k,s}\sin(\omega_k t))$$

for $0 \leq \omega_k \leq \pi$, i.e. when a mean (if $\omega_1 = 0$) and harmonic components are present, when the autoregressions are applied to

$$\hat{v}_t := v_t - \langle v_t, d_t\rangle_1^T(\langle d_t, d_t\rangle_1^T)^{-1}d_t,$$

where $d_{t,k} := \begin{pmatrix}\cos(\omega_k t)\\ \sin(\omega_k t)\end{pmatrix}$ for $0 < \omega_k < \pi$ and $d_{t,k} := \cos(\omega_k t)$ for $\omega_k \in \{0,\pi\}$ and $d_t := [d'_{t,1},\ldots,d'_{t,l}]'$.

The theorem shows that the coefficients of autoregressive approximations converge even when the order is tending to infinity as a function of the sample size. Here it is of particular importance that the theorem derives error bounds that are uniform in the lag lengths. Uniform error bounds are required because order selection necessarily considers the criterion function $IC^v(p;C_T)$ for all values $0 \leq p \leq H_T$ simultaneously. Based upon the uniform convergence results for the autoregression coefficients the asymptotic properties of information criteria are derived, which are seen to depend upon characteristics of the true unknown system (in particular Σ^v_p, which in the VARMA case is closely related to ρ_0, see HD, p. 334). The result establishes a connection between the minimizing order of the information criterion and the deterministic function $\tilde{L}_T(p;C_T) := ps^2\frac{C_T-1}{T} + \mathrm{tr}\left[\Sigma^{-1}(\Sigma^v_p - \Sigma)\right]$. The approximation in loose terms implies that the order estimator $\hat{p}$ cannot be 'very far' from the optimizing value of the deterministic function (see also the discussion below Theorem 7.4.7

on p. 333–334 in HD). This implication heavily relies on the uniformity of the approximation. Here 'very far' refers to a large ratio of the value of the deterministic function to its minimal value. Under an additional assumption on the shape of the deterministic function (compare Corollary 1(ii)), results for the asymptotic behavior of $\hat{p}$ can be obtained (see Corollary 1 below). In particular in the stationary VARMA case it follows from (iv) that $\hat{p}_{BIC}$ increases essentially proportional to $\log T$ as does the minimizer of the deterministic function. The result in item (v) is required for the theorems in the following section, where it will be seen that the properties of autoregressive approximations in the MFI(1) case are related to the properties of autoregressive approximations of a related stationary process where only certain coordinates of the last lag are included in the regression. The final result in (vi) shows that the presence of a non-zero mean and harmonic components does not affect any of the stated asymptotic properties.

4 Autoregressive Approximations of MFI(1) Processes

In this section autoregressive approximations of MFI(1) processes $(y_t)_{t\in\mathbb{Z}}$ are considered. The discussion in the text focuses for simplicity throughout on the case of Assumption 3(i) without deterministic components (i.e. without mean and harmonic components), however, the theorems contain the results also for the case including these deterministic components, i.e. under Assumption 3(ii). Parallelling the notation in the previous section define

$$u_t(p) := y_t + \Phi_p^y(1)y_{t-1} + \ldots + \Phi_p^y(p)y_{t-p}.$$

The LS estimator of $\Phi_p^y(j), j = 1, \ldots, p$ is given by

$$\hat{\Theta}_p := -\left[\hat{\Phi}_p^y(1), \ldots, \hat{\Phi}_p^y(p)\right] := \langle y_t, Y_{t,p}^- \rangle \langle Y_{t,p}^-, Y_{t,p}^- \rangle^{-1}$$

with $Y_{t,p}^- := [y_{t-1}', \ldots, y_{t-p}']'$. Furthermore denote $\hat{\Sigma}_p^y = \langle y_t - \hat{\Theta}_p^y Y_{t,p}^-, y_t - \hat{\Theta}_p^y Y_{t,p}^- \rangle$ and, also as in the stationary case, for $0 \leq p \leq H_T$

$$IC^y(p; C_T) := \log\det \hat{\Sigma}_p^y + ps^2 \frac{C_T}{T},$$

where again C_T is a suitably chosen penalty function. An order estimator is again given by $\hat{p} := \min\left(\arg\min_{0\leq p\leq H_T} IC^y(p; C_T)\right)$.

The key tool for deriving the asymptotic properties of $\hat{\Theta}_p^y$ is a separation of the stationary and nonstationary directions in the regressor vector $Y_{t,p}^-$. Define the observability index $q \in \mathbb{N}$ as the minimal integer such that the matrix

$$\mathcal{O}_q := \begin{bmatrix} C_1 & \dots & C_l \\ C_1 J_1 & \dots & C_l J_l \\ \vdots & & \vdots \\ C_1 J_1^{q-1} & \dots & C_l J_l^{q-1} \end{bmatrix}$$

has full column rank. Due to minimality of the systems (J_k, K_k, C_k) for $k = 1, \dots, l$ this integer exists (cf. Theorem 2.3.3 on p. 48 of HD).

Lemma 1 *Let $(y_t)_{t\in\mathbb{Z}}$ be generated according to Assumption 1 and Assumption 3(i). Denote with $C := [C_1, \dots, C_l] \in \mathbb{R}^{s\times c}$, $K := [K_1', \dots, K_l']' \in \mathbb{R}^{c\times s}$, $J := diag(J_1, \dots, J_l) \in \mathbb{R}^{c\times c}$ and $x_t := [x_{t,1}', \dots, x_{t,l}']' \in \mathbb{R}^c$ where $c := \sum_{k=1}^l c_k$, with C_k, K_k, J_k and $x_{t,k}$ as in Theorem 1. Denote furthermore with $e_t := c_\bullet(L)\varepsilon_t$. Hence $y_t = Cx_t + e_t$.*

(i) If $q = 1$ define $\bar{C}' := [C^\dagger, C^\perp]$, with $C^\dagger := C(C'C)^{-1}$ and $C^\perp \in \mathbb{R}^{s\times(s-c)}$ is such that $(C^\perp)'C^\perp = I_{s-c}$, $C'C^\perp = 0$. Define

$$\mathcal{T}_p := Q_p\left(I_p \otimes \bar{C}\right), \textit{ where } \quad Q_p := \begin{bmatrix} I_c & 0 & 0 & & & & & \\ 0 & I_{s-c} & 0 & & & & & \\ I_c & 0 & -J & 0 & & & & \\ & 0 & 0 & I_{s-c} & & & & \\ & & I_c & 0 & -J & 0 & & \\ & & & \ddots & \ddots & \ddots & & \\ & & & & & & & I_{s-c} \end{bmatrix} \tag{8}$$

and

$$Z_{t,p}^- := \mathcal{T}_p Y_{t,p}^- = \begin{pmatrix} x_{t-1} + (C^\dagger)' e_{t-1} \\ (C^\perp)' e_{t-1} \\ K\varepsilon_{t-2} + (C^\dagger)' e_{t-1} - J(C^\dagger)' e_{t-2} \\ (C^\perp)' e_{t-2} \\ K\varepsilon_{t-3} + (C^\dagger)' e_{t-2} - J(C^\dagger)' e_{t-3} \\ \vdots \\ (C^\perp)' e_{t-p} \end{pmatrix}.$$

With the quantities just defined it holds that

$$y_t - CJ(C^\dagger)' y_{t-1} = \left[C, C^\perp\right] \begin{bmatrix} K\varepsilon_{t-1} + (C^\dagger)' e_t - J(C^\dagger)' e_{t-1} \\ (C^\perp)' e_t \end{bmatrix}.$$

Define the transfer function

$$\tilde{c}_\bullet(z) := \sum_{j=0}^{\infty} \tilde{c}_{j,\bullet} z^j = \left[C, C^\perp\right] \left[\begin{array}{c} Kz + (I - zJ)(C^\dagger)' c_\bullet(z) \\ (C^\perp)' c_\bullet(z) \end{array} \right]$$

and let $\tilde{e}_t := \tilde{c}_\bullet(L)\varepsilon_t$. *The transfer function* $\tilde{c}_\bullet(z)$ *has the following properties:* $\tilde{c}_\bullet(0) = I_s$, $\sum_{j=0}^{\infty} j^{3/2}\|\tilde{c}_{j,\bullet}\| < \infty$ *and hence* $\tilde{c}_\bullet(z)$ *has no poles on the closed unit disc. Furthermore* $\tilde{c}_\bullet(z)$ *has no zeros on the unit disc, i.e.* $\det \tilde{c}_\bullet(z) \neq 0$ *for all* $|z| < 1$.

(ii) If $q > 1$ *define* $\tilde{y}_t := [y_t', y_{t+1}', \ldots, y_{t+q-1}']' \in \mathbb{R}^{\tilde{s}}$, *with* $\tilde{s} := sq$. *Then for each* $i = 1, \ldots, q$, *the sub-sampled process* $(\tilde{y}_{tq+i})_{t\in\mathbb{Z}}$ *is generated according to Assumption 1 and by construction the observability index corresponding to this process is equal to 1. Thus, for the processes* $(\tilde{y}_{tq+i})_{t\in\mathbb{Z}}$ *part (i) of the lemma applies with* $\tilde{\varepsilon}_t^{(q)} = \tilde{c}_\bullet^{(q)}(L^q)\tilde{\varepsilon}_t$ *and it follows that there exists a matrix* $\tilde{\mathcal{T}}_{\tilde{p}} \in \mathbb{R}^{\tilde{p}\tilde{s}\times\tilde{p}\tilde{s}}$ *such that in* $Z_{t,\tilde{p}q}^- := \tilde{\mathcal{T}}_{\tilde{p}} Y_{t,\tilde{p}q}^-$ *the first* c *coordinates are unit root processes while the remaining components are stationary.*

The proof of the lemma is given in Appendix B. The lemma is stated only under Assumption 3(i), however, it is obvious that it also applies under Assumption 3 (ii) in which case $e_t := c_\bullet(L)\varepsilon_t + T_t$. The idea is, as stated above, to separate the stationary and the nonstationary directions in the regressor vector, which is achieved in $Z_{t,p}^-$. Only the first c components are unit root processes which are independent of the choice of the lag length p. Only the stationary part of the regressor vector $Z_{t,p}^-$ depends upon p. Therefore, for this part the theory reviewed for stationary processes in the previous section is an important input.

Note that in the important I(1) case it holds that $q = 1$ (due to minimality) and thus the simpler representation developed in (i) can be used and no sub-sampling arguments are required. As is usual in deriving the properties of autoregressive approximations an invertibility condition is required.

Assumption 4 *The true transfer function* $c(z)$ *is such that* $\det \tilde{c}_\bullet(z) \neq 0, |z| = 1$, *for* $\tilde{c}_\bullet(z)$ *as defined in Lemma 1. Note that in case (ii) of Lemma 1 this assumption has to hold for the correspondingly defined transfer function,* $\tilde{c}_\bullet^{(q)}(z^q)$ *say, of the sub-sampled processes.*

For $q = 1$ it follows that $y_t - CJ(C^\dagger)' y_{t-1} = \tilde{c}_\bullet(L)\varepsilon_t$ and hence under Assumption 4 we obtain $\tilde{c}_\bullet(L)^{-1}(I - CJ(C^\dagger)'L)y_t = \varepsilon_t$ showing that $(y_t)_{t\in\mathbb{Z}}$ is the solution to an infinite order autoregression. Letting $\Phi(z) := \tilde{c}_\bullet(z)^{-1}(I - CJ(C^\dagger)'z) = \sum_{j=0}^{\infty} \Phi^y(j) z^j$ we have $\sum_{j=0}^{\infty} j^{3/2}\|\Phi^y(j)\| < \infty$. For $q > 1$ as similar representation can be obtained.
The following bivariate example shows that Assumption 4 is not void. Let $\Delta_0(L)y_t = c(L)\varepsilon_t$, with

$$c(z) = \begin{bmatrix} 1 & 1.5 - z \\ 1 - z & 0.5z - 0.5z^2 \end{bmatrix},$$

which for simplicity is not normalized to $c(0) = I_2$. The determinant of $c(z)$ is equal to $\det c(z) = -1.5(1-z)^2$ and hence $z = 1$ is the only root. Furthermore, $c(1) = C_1 K_1$ is non-zero and equal to $[1, 0]'[1, 0.5]$. Now, using the representation of $c(z)$ as derived in Lemma 1 we find

$$c(z) = zC_1K_1 + (1-z)c_\bullet(z)$$

with

$$c_\bullet(z) = \begin{bmatrix} 1 & 1.5 \\ 1 & 0.5z \end{bmatrix}.$$

Thus, $\det c_\bullet(z) = 0.5z - 1.5$ and hence $\det c_\bullet(z)$ has its root outside the closed unit circle. However, if one considers $\tilde{c}_\bullet(z)$ for this example, given by

$$\tilde{c}_\bullet(z) = \begin{bmatrix} K_1 z + (1-z)(C^\dagger)' c_\bullet(z) \\ (C_1^\perp)' c_\bullet(z) \end{bmatrix} = \begin{bmatrix} 1 & 1.5 - z \\ 1 & 0.5z \end{bmatrix}.$$

evaluated at $z = 1$ one obtains

$$\tilde{c}_\bullet(1) = \begin{bmatrix} 1 & 0.5 \\ 1 & 0.5 \end{bmatrix},$$

from which one sees that $\det \tilde{c}_\bullet(1) = 0$. This example shows that indeed the assumption is not void. However, since all entries in K_1 are free parameters, Assumption 4 is not fulfilled only on a 'thin set', i.e. on the complement of an open and dense subset. Similar considerations also apply to the general case, but we abstain from discussing these issues here in detail.

It follows from the distinction of the two cases ($q = 1$ or $q > 1$) in the above Lemma 1 that the following theorems concerning autoregressive approximations have to be derived separately for these two cases. The first case is dealt with in Theorem 3 and the second is considered in Theorem 4.

Assume again for the moment that Assumption 3(i) holds (for simplicity only, since as already mentioned the results are also derived when Assumption 3(ii) holds) and consider the case $q = 1$. Note that for any choice of the autoregressive lag length p it holds that

$$\hat{\Theta}_p^y = \langle y_t, Y_{t,p}^- \rangle \langle Y_{t,p}^-, Y_{t,p}^- \rangle^{-1} = \langle y_t, Z_{t,p}^- \rangle \langle Z_{t,p}^-, Z_{t,p}^- \rangle^{-1} \mathcal{T}_p = \hat{\Theta}_p^z \mathcal{T}_p,$$

where this equation defines $\hat{\Theta}_p^z$. Now, since $y_t = CJ(C^\dagger)' y_{t-1} + \tilde{e}_t$ and $(C^\dagger)' y_{t-1}$ is equal to the first c components of $Z_{t,p}^-$ we obtain

$$\hat{\Theta}_p^z = \left[\begin{array}{cccc} CJ & 0 & \ldots & 0 \end{array} \right] + \langle \tilde{e}_t, Z_{t,p}^- \rangle \langle Z_{t,p}^-, Z_{t,p}^- \rangle^{-1}$$

and thus it is sufficient to establish the asymptotic behavior of the second term on the right hand side of the above equation. Now, let $Z_{t,p}^- := [z_t', (Z_{t,p,2}^-)']'$, where $z_t \in \mathbb{R}^c$ contains the nonstationary components and $Z_{t,p,2}^-$ contains the stationary components. The proof of the following Theorem 3 given in Appendix B shows that the asymptotic behavior of the estimator $\hat{\Theta}_p^y$ is governed by the asymptotic behavior of

$$\hat{\Theta}_p^{\tilde{e}} := \langle \tilde{e}_t, Z_{t,p,2}^- \rangle \langle Z_{t,p,2}^-, Z_{t,p,2}^- \rangle^{-1},$$

i.e. by the asymptotic distribution of an autoregression including only stationary quantities. It is this result that shows that the asymptotic behavior is in many aspects similar in the stationary and the MFI(1) case. Note here also that $Z_{t,p,2}^-$ is a linear function of the lags of $\tilde{e}_t$. Given that all quantities are stationary we can define

$$\Theta_p^{\tilde{e}} := \mathbb{E}\tilde{e}_t (Z_{t,p,2}^-)' (\mathbb{E} Z_{t,p,2}^- (Z_{t,p,2}^-)')^{-1}$$

and analogously $\Theta^{\tilde{e}}$ as the solution to the corresponding infinite order population YW equations. Finally, as in Section 3 define $\Sigma^{\tilde{e}}$ and $\Sigma_p^{\tilde{e}}$ as the population innovation variance of the process $(\tilde{e}_t)_{t \in \mathbb{Z}}$ and the best one-step ahead prediction error achieved using an autoregressive approximation of order p. Then, the next theorem states the properties of autoregressive approximations for MFI(1) processes with $q = 1$.

Theorem 3 *Let $(y_t)_{t \in \mathbb{Z}}$ be generated according to Assumptions 1, 2, 3(i) and 4 such that $q = 1$ and let $0 \le p \le H_T$.*

(i) Then it holds that

$$\max_{1\le p\le H_T} \left\| \hat{\Theta}_p^y - \left(\bar{C}^{-1} \left[\left(\begin{array}{c} J \\ 0 \end{array} \right), \Theta_p^{\tilde{e}} \right] \right) \mathcal{T}_p \right\|_1 = O_P((\log T/T)^{1/2}),$$

where $\|.\|_1$ *denotes the matrix 1-norm.*

(ii) For $C_T \ge c > 1$ *the following approximations hold*

$$\begin{array}{rcl} IC^y(p;C_T) & \le & \log\det\dot{\Sigma} + \left\{ \frac{ps^2}{T}(C_T-1) + tr\left[(\Sigma^{\tilde{e}})^{-1}(\Sigma^{\tilde{e}}_{p-1} - \Sigma^{\tilde{e}})\right] \right\} \{1+o_P(1)\}, \\ IC^y(p;C_T) & \ge & \log\det\dot{\Sigma} + \left\{ \frac{ps^2}{T}(C_T-1) + tr\left[(\Sigma^{\tilde{e}})^{-1}(\Sigma^{\tilde{e}}_{p} - \Sigma^{\tilde{e}})\right] \right\} \{1+o_P(1)\}. \end{array}$$

The error term here is $o_P(1)$ *uniformly in* $0 \le p \le H_T$.

(iii) All statements remain valid if Assumption 3(ii) holds with $T_t = Dd_t$ *and the autoregressive approximations are performed on* $\hat{y}_t$, *defined as* $\hat{y}_t := y_t - \langle y_t, d_t\rangle_1^T (\langle d_t, d_t\rangle_1^T)^{-1} d_t$, *with* d_t *as defined in Theorem 2.*

Here (i) is the analogue to Theorem 2(i), the only difference being that the result is stated in probability rather than a.s. This shows that the existence of (seasonal) integration does not alter the estimation accuracy of the autoregression coefficients. Result (ii) is essentially the analogue of Theorem 2(iii), where however due to the fact that in the considered regression components of $\tilde{e}_{t-p}$ are omitted (since only $(C^{\perp})'e_{t-p}$ is contained in the regressor vector $Z^{-}_{t,p,2}$) lower and upper bounds similar to the bounds derived in Theorem 2(v) are developed. As in the stationary case the result provides uniform bounds for the information criterion (in the range $0 \le p \le H_T$) given by the sum of a deterministic function and a noise term. These results imply that the asymptotic behavior of the autoregressive approximation essentially depends on the properties of the stationary process $(\tilde{e}_t)_{t\in\mathbb{Z}}$: Except for the first block all blocks of $\hat{\Theta}_p^z$ converge to blocks of the matrix $\Theta_p^{\tilde{e}}$ which correspond to an autoregressive approximation of the stationary process $(\tilde{e}_t)_{t\in\mathbb{Z}}$. The uniform bounds on the information criterion also provides a strong relation to the information criterion corresponding to autoregressive approximations of $(\tilde{e}_t)_{t\in\mathbb{Z}}$. Corollary 1 below uses this as the basis to show that the asymptotic properties of order estimators are similar in some aspects for the stationary and the MFI(1) case.

For $q > 1$ the sub-sampling argument outlined in Lemma 1 shows that similar results can be obtained by resorting to q time series of dimension $\tilde{s} = qs$, with time increment q that are combined for estimation. Focusing on the first block of this stacked process then leads to the results analogous to those obtained for $q = 1$ in Theorem 3 for $(y_t)_{t\in\mathbb{Z}}$ when $q > 1$.

Theorem 4 *Let $(y_t)_{t\in\mathbb{Z}}$ be generated according to Assumptions 1, 2, 3(i) and 4, assume $q > 1$ and let $0 \leq p = \tilde{p}q \leq H_T, \tilde{p} \in \mathbb{N} \cup \{0\}$. Let $\mathcal{C} := [C', J'C', \dots, (J^{q-1})'C']' \in \mathbb{R}^{\tilde{s}\times s}$. Further let $\bar{\mathcal{C}} := [\mathcal{C}^\dagger, \mathcal{C}^\perp]'$ where $(\mathcal{C}^\perp)'\mathcal{C}^\perp = I_{\tilde{s}-c}, \mathcal{C}'\mathcal{C}^\perp = 0$ and $\mathcal{C}^\dagger := \mathcal{C}(\mathcal{C}'\mathcal{C})^{-1}$ and use $\tilde{\mathcal{T}}_{\tilde{p}}$ as defined in Lemma 1.*

(i) Defining $\tilde{I}_s = \left[I_s, 0^{s\times(q-1)s}\right]$ it holds that

$$\max_{1\leq\tilde{p}\leq H_T} \left\| \hat{\Theta}^y_{\tilde{p}q} - \tilde{I}_s \left(\bar{\mathcal{C}}^{-1} \left[\begin{pmatrix} J \\ 0 \end{pmatrix}, \Theta^{\tilde{e}}_{\tilde{p}q} \right] \right) \tilde{\mathcal{T}}_{\tilde{p}} \right\|_1 = O_P((\log T/T)^{1/2}).$$

(ii) Further letting now $\tilde{p} := \lfloor p/q \rfloor q$ for $C_T \geq c > 1$ the following approximations hold (where again the $o_P(1)$ term holds uniformly in $0 \leq p \leq H_T$)

$$\begin{aligned} IC^y(p; C_T) &\leq \log\det\dot{\Sigma} + \left\{ \frac{(\tilde{p}+1)qs^2}{T}(C_T - 1) + tr\left[\Sigma^{-1}\tilde{I}_s(\Sigma^{\tilde{e}}_{\tilde{p}-1} - \Sigma^{\tilde{e}})\tilde{I}'_s\right] \right\} \{1 + o_P(1)\}, \\ IC^y(p; C_T) &\geq \log\det\dot{\Sigma} + \left\{ \frac{\tilde{p}qs^2}{T}(C_T - 1) + tr\left[\Sigma^{-1}\tilde{I}_s(\Sigma^{\tilde{e}}_{\tilde{p}} - \Sigma^{\tilde{e}})\tilde{I}'_s\right] \right\} \{1 + o_P(1)\}. \end{aligned}$$

(iii) All statements remain valid if Assumption 3(ii) holds with $T_t = Dd_t$ and the autoregressive approximations are performed on $\hat{y}_t$, defined as $\hat{y}_t := y_t - \langle y_t, d_t\rangle_1^T(\langle d_t, d_t\rangle_1^T)^{-1}d_t$, with d_t as defined in Theorem 2.

Compared to the results in case that $q = 1$, the results obtained when $q > 1$ are weaker. The approximation results in (i) are only stated for p being an integer multiple of q, although it seems to be possible to extend the uniform bound on the estimation error to $p \in \mathbb{N}$. The bounds on the information criterion are also related to the closest integer multiple of q. Nevertheless, as $\hat{p}$ tends to infinity this difference might be considered minor.

We close this section by using the results derived above in Theorems 3(ii) and 4(ii) to study the asymptotic properties of information criterion based order estimation. In the approximation to the information criterion (discussing here the case corresponding to Theorem 4), the deterministic function

$$\tilde{L}_T(\tilde{p}; C_T) := \text{tr}\left[\Sigma^{-1}\tilde{I}_s(\Sigma^{\tilde{e}}_{\tilde{p}} - \Sigma^{\tilde{e}})\tilde{I}_s'\right] + \frac{\tilde{p}qs^2}{T}(C_T - 1)$$

has a key role. If we assume that $C_T/T \to 0$, then it follows that the minimizing argument of this function, $l_T(C_T)$ say, tends to infinity unless there exists an index p_0, such that $\Sigma^{\tilde{e}}_p = \Sigma^{\tilde{e}}$ for $p \geq p_0$, which is the case if the process is an autoregression of order p_0. The discussion on p. 333–334 of HD links $l_T(C_T)$ and $\hat{p}(C_T)$ minimizing the information criterion $IC^y(p; C_T)$. The main building block is the uniform convergence of the information criterion to the deterministic function $\tilde{L}_T(p; C_T)$. The lower and upper bounds on the information criteria as established in (ii) above are sufficient for the arguments in HD to hold, as will be shown in Corollary 1 below.

We also consider the important special case of VARMA processes, where the underlying transfer function $c(z)$ is a rational function. Recall from Theorem 2(iv) in Section 2 that for stationary VARMA processes the choice of $C_T = \log T$ (i.e. using `BIC`) leads to the result that $\lim_{T\to\infty} \frac{2\hat{p}_{BIC} \log \rho_0}{\log T} = 1$ almost surely. Here we denote again with $\hat{p}_{BIC}$ the smallest minimizing argument of the information criterion `BIC` and by ρ_0 the smallest modulus of the zeros of the moving average polynomial. This result is extended to the MFI(1) case, however, only in probability and not a.s. in item (iii) of Corollary 1 below. The above discussion concerning lag length selection based on information criteria is formalized in the following corollary, whose proof is given in Appendix B.

Corollary 1 *Let $(y_t)_{t\in\mathbb{Z}}$ be generated according to Assumptions 1, 2, 3(i) and 4. Assume that for all $p \in \mathbb{N} \cup \{0\}$ it holds that $\Sigma^{\tilde{e}}_p > \Sigma^{\tilde{e}}$, i.e. $(\tilde{e}_t)_{t\in\mathbb{Z}}$ has no finite order VAR representation. Denote with $\hat{p}(C_T)$ the smallest minimizing argument of $IC^y(p; C_T)$ over the set of integers $0 \leq p \leq H_T, H_T = o((T/\log T)^{1/2})$ and assume that $C_T \geq c > 1$ and $C_T/T \to \infty$. Then the following results hold:*

(i) $\mathbb{P}\{\hat{p}(C_T) < M\} \to 0$ for any constant $M < \infty$.

(ii) Assume that there exists a twice differentiable function $\tilde{\theta}(p)$ with second derivative $\tilde{\theta}''(p)$ such that $\lim_{p\to\infty} tr\left[\Sigma^{-1}\tilde{I}_s(\Sigma^{\tilde{e}}_p - \Sigma^{\tilde{e}})\tilde{I}_s'\right]/\tilde{\theta}(p) = 1$ and $\liminf_{p\to\infty} |p^2\tilde{\theta}''(p)/\tilde{\theta}(p)| >$

0. *Then* $\hat{p}(C_T)/(ql_T(C_T)) \to 1$ *in probability, where* q *denotes again the observability index and* $l_T(C_T)$ *is as defined above the formulation of the corollary.*

(iii) If $(y_t)_{t\in\mathbb{Z}}$ *is an MFI(1) VARMA process, then* $2\hat{p}\log\rho_0/\log T \to 1$ *in probability, where* $\rho_0 = \min\{|z| : z \in \mathbb{C}, \det \tilde{c}_\bullet^{(q)}(z^q) = 0\}$*, with* $\tilde{c}_\bullet^{(1)}(z) = \tilde{c}_\bullet(z)$*.*

(iv) All statements remain valid if Assumption 3(ii) holds with $T_t = Dd_t$ *and the autoregressive approximations are performed on* $\hat{y}_t$*, defined as* $\hat{y}_t = y_t - \langle y_t, d_t\rangle_1^T(\langle d_t, d_t\rangle_1^T)^{-1}d_t$*, with* d_t *as defined in Theorem 2,*

Note finally that as in the stationary case also almost sure results can be obtained for a sharper bound on the admitted lags given by $G_T = (\log T)^a$ for some $a < \infty$ and stronger assumptions on the errors. In particular for i.i.d. errors $(\varepsilon_t)_{t\in\mathbb{Z}}$ with the same moment restrictions as formulated in Assumption 2 it is possible to derive the a.s. counterparts of (i) of Theorems 3 and 4 for $1 \leq p \leq G_T$ under the additional assumption that all unit root frequencies are rational multiples of π. Also the smaller bounds on the estimation error provided in Theorem 2(ii) for VARMA processes are straightforward to generalize to the MFI(1) setting. However, the practical relevance of these results might be doubted and hence no details are given.

5 Summary and Conclusions

In this paper we have studied the asymptotic properties of autoregressive approximations of multiple frequency I(1) processes. These are defined in this paper as processes with unit roots of integration orders all equal to one and with rather general assumptions on the underlying transfer function (and certain restrictions on the deterministic components). The only assumptions on the transfer function are that the coefficients converge sufficiently fast (see Assumption 1) and that an appropriate invertibility condition (see Assumption 4, which is standard in autoregressive approximations) holds. These assumptions imply that we do not restrict ourselves to VARMA processes (where the transfer functions are restricted to be rational), but exclude long memory processes (e.g. fractionally integrated processes). Also the assumptions on the noise process are rather standard in this literature, and essentially allow for martingale difference sequence type errors with a moment assumption that is slightly stronger than finite fourth moments. The innovations are restricted to be conditionally homoskedastic.

The main insight from our results is that the properties of autoregressive approximations in the MFI(1) case are essentially driven by the properties of a related stationary process, $(\tilde{c}_{\bullet}^{(q)}(L^q)\tilde{\varepsilon}_t)_{t\in\mathbb{Z}}$ in the notation used throughout. This observation is important, since the approximation properties of autoregressions are well understood for stationary processes (compare Section 7.4 of HD). Thus, based on the above insight we obtain uniform convergence of the autoregressive coefficients when the lag lengths are tending to infinity at a rate not faster than $o((T/\log T)^{1/2})$. The obtained bound on the estimation error, which is of order $O_P((\log T/T)^{1/2})$, appears to be close to minimal, being slightly larger than $T^{-1/2}$.

The convergence results are used in a second step to study the asymptotic properties of order estimators based on information criteria. It is shown, establishing again a similarity to the stationary case, that the autoregressive approximation order obtained by minimizing information criteria typically behaves as a deterministic function of the sample size and certain characteristics of the data generating process. One particularly important result obtained in this respect is that for MFI(1) VARMA processes order estimation according to BIC leads to divergence (in probability) of the order proportionally to $\log T$. This result is a generalization of the almost sure result stated for stationary processes in Theorem 6.6.3 in HD. This result closes an important gap in the existing literature, since previously available results (e.g. Lemma 4.2 of Ng and Perron, 1995) do not provide sharp enough bounds on the error terms, which imply that such results can only be used in conjunction with overly large penalty terms. Thus, even for the fairly well studied I(1) case the corresponding results in this paper are new.

This paper does not analyze estimators of the autoregressive approximations that are based on estimated orders in detail (e.g. no limiting distributions are provided in this paper, see e.g. Kuersteiner, 2005 in this respect for stationary processes), since we only derive uniform error bounds. Of particular importance in this respect seems to be the extension of the estimation theory for seasonally integrated processes from the finite order autoregressive case dealt with in Johansen and Schaumburg (1999) to the case of infinite order autoregressions. This includes both extending the asymptotic theory for tests for the cointegrating ranks to the infinite autoregression case (analogous to the results in Saikkonen and Luukkonen, 1997) as well as providing asymptotic distributions for the estimated coefficients. This is left for future research, for which important prerequisites have been established in this paper.

References

Akaike, H. (1975). Markovian representation of stochastic processes by canonical variables. *SIAM Journal of Control* **13**(1), 162–172.

Bauer, D. and M. Wagner (2002). Estimating cointegrated systems using subspace algorithms. *Journal of Econometrics* **111**, 47–84.

Bauer, D. and M. Wagner (2004). A canonical form for unit root processes in the state space framework. Technical report. TU Wien, Austria.

Chan, N. H. and C. Z. Wei (1988). Limiting distributions of least squares estimates of unstable autoregressive processes. *Annals of Statistics* **16**, 367–401.

Hall, P. and C. Heyde (1980). *Martingale Limit Theory and its Application*. Academic Press.

Hannan, E. J. and M. Deistler (1988). *The Statistical Theory of Linear Systems*. John Wiley. New York.

Hannan, E. J. and L. Kavalieris (1986). Regression, autoregression models. *Journal of Time Series Analysis* **7**, 27–49.

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

Johansen, S. and E. Schaumburg (1999). Maximum likelihood estimation of seasonally integrated series. *Journal of Econometrics* **88**, 301–339.

Kuersteiner, G.M. (2005). Automatic inference for infinite order vector autoregressions. *Econometric Theory* **21**, 85–115.

Lai, T. L. and C. Z. Wei (1982). Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *Annals of Statistics* **10**(1), 154–166.

Lai, T. L. and C. Z. Wei (1983). Asymptotic properties of general autoregressive models and strong consistency of least squares estimates of their parameters. *Journal of Multivariate Analysis* **13**, 1–23.

Lewis, R. and G. Reinsel (1985). Prediction of multivariate time series by autoregressive model fitting. *Journal of Multivariate Analysis* **16**, 393–411.

Lütkepohl, H. and P. Saikkonen (1999). Order selection in testing for the cointegrating rank of a VAR process, 168–199. In Engle, R.F. and H. White (Eds.) *Cointegration, Causality and Forecasting.* Oxford University Press. Oxford, New York.

Ng, S. and P. Perron (1995). Unit root tests in ARMA models with data-dependent methods for the selection of the truncation lag. *Journal of the American Statistical Association* **90**, 268–281.

Pötscher, B. (1989). Model selection under nonstationarity: Autoregressive models and linear stochastic regression models. *Annals of Statistics* **17**, 1257–1274.

Saikkonen, P. (1992). Estimation and testing of cointegrated systems by an autoregressive approximation. *Econometric Theory* **8**, 1–27.

Saikkonen, P. (1993). Estimation of cointegration vectors with linear restrictions. *Econometric Theory* **9**, 19–35.

Saikkonen, P. and H. Lütkepohl (1996). Infinite order cointegrated vector autoregressive processes: estimation and inference. *Econometric Theory* **12**, 814–844.

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

Sims, C. A., J. H. Stock and M. W. Watson (1990). Inference in linear time series models with some unit roots. *Econometrica* **58**, 113–144.

Stock, J.H. and M.W. Watson (1988). Testing for common trends. *Journal of the American Statistical Association* **83**, 1097–1107.

A Preliminaries

In this first appendix several preliminary lemmata are collected. We start with Lemma 2, which discusses a specific factorization of analytic functions useful for Theorem 1.

Lemma 2 *Let $c(z) = \sum_{j=0}^{\infty} c_j z^j, c_j \in \mathbb{R}^{s\times s}, j \geq 0$, be analytic on $|z| \leq 1$. Assume that there exists an integer $L \geq H$ for $H := \sum_{k=1}^{l}(1 + \mathbb{I}(0 < \omega_k < \pi))$ such that $\sum_{j=0}^{\infty} j^{1/2+L}\|c_j\| < \infty$. Denote with $0 \leq \omega_1 < \cdots < \omega_l \leq \pi$ a set of frequencies and denote again with $D(z) := \Delta_{\omega_1}(z)\dots\Delta_{\omega_l}(z)$. Further define $D_{\neg k}(z) := \prod_{j\neq k}\Delta_{\omega_j}(z)$. Denote with $\tilde{c}_k$ the rank of $c(e^{-i\omega_k}) = \tilde{C}_k\tilde{K}_k \in \mathbb{C}^{s\times s}$ with $\tilde{C}_k \in \mathbb{C}^{s\times\tilde{c}_k}$, $\tilde{K}_k \in \mathbb{C}^{\tilde{c}_k\times s}$ and with $c_k := \tilde{c}_k(1+\mathbb{I}(0 < \omega_k < \pi))$. Further define J_k and S_k as in (5). Then there exist matrices $C_k \in \mathbb{R}^{s\times c_k}, K_k \in \mathbb{R}^{c_k\times s}, k = 1,\dots,l$ and a function $c_\bullet(z) = \sum_{j=0}^{\infty} c_{j,\bullet}z^j, c_{j,\bullet} \in \mathbb{R}^{s\times s}$, such that:*

(i) $\sum_{j=0}^{\infty} j^{1/2+L-H}\|c_{j,\bullet}\| < \infty$. Thus, $c_\bullet(z)$ is analytic on the closed unit disc.

(ii) The function $c(z)$ can be decomposed as

$$c(z) = \sum_{k=1}^{l} zD_{\neg k}(z)C_k(I - zJ_k'\mathbb{I}(0 < \omega_k < \pi))K_k + D(z)c_\bullet(z). \tag{9}$$

(iii) Representation (9) is unique up to the decomposition of the products $C_k[K_k, -J_k'K_k]$.

Proof: For algebraic convenience the proof uses complex quantities and the real representation in the formulation of the lemma is derived from the complex results at the end of the proof.

Thus, let $0 \leq \omega_1 < \ldots < \omega_H < 2\pi$ be the set of frequencies where we now consider complex conjugate frequencies separately. We denote the unit roots corresponding to the frequencies with $z_k := e^{i\omega_k}$. The fact that unit roots appear in pairs of complex conjugate roots follows immediately from the fact that the coefficients c_j of $c(z)$ are real valued. Denote with $\tilde{D}_{\neg k} := D(z)/(1 - z\overline{z_k})$.

The proof is inductive in the unit roots. Thus, let

$$c^{(r)}(z) = c(z) - \sum_{k=1}^{r} \frac{z\tilde{D}_{\neg k}(z)}{\overline{z_k}\tilde{D}_{\neg k}(\overline{z_k})}\tilde{C}_k\tilde{K}_k$$

where $c^{(r)}(z) = (1 - zz_1)\dots(1 - zz_r)c_\bullet^{(r)}(z)$ is such that $c_\bullet^{(r)}(z) = \sum_{j=0}^{\infty} c_{\bullet,j}^{(r)} z^j$, with $\sum_{j=0}^{\infty} j^{1/2+L-r}\|c_{j,\bullet}^{(r)}\| < \infty$. Now consider $\hat{c}^{(r+1)}(z) = c_\bullet^{(r)}(z) - \frac{z\tilde{D}_{\neg r+1}(z)}{\overline{z_{r+1}}\tilde{D}_{\neg r+1}(\overline{z_{r+1}})}\tilde{C}_{r+1}\tilde{K}_{r+1}/[(1 - zz_1)\dots(1-zz_r)]$. By inserting it follows immediately that $\hat{c}^{(r+1)}(\overline{z_{r+1}}) = 0$, and we can thus write $\hat{c}^{(r+1)}(z) = (1-zz_{r+1})c_\bullet^{r+1}(z)$. Also, since $\hat{c}^{(r+1)}(z)$ and $c_\bullet^{(r)}(z)$ differ only by a polynomial they have the same summability properties. We can write $\hat{c}^{(r+1)}(z) = \sum_{j=0}^{\infty} \hat{c}_j^{(r+1)} z^j = (1 - zz_{r+1})\sum_{j=0}^{\infty} c_{j,\bullet}^{(r+1)} z^j$ and using a formal power series expansion we obtain: $c_{0,\bullet}^{(r+1)} = I_s$ and $c_{j,\bullet}^{(r+1)} = \hat{c}_j^{(r+1)} + z_{r+1}c_{j-1,\bullet}^{(r+1)}$, which implies

$$c_{j,\bullet}^{(r+1)} = \sum_{i=0}^{j} \hat{c}_{j-i}^{(r+1)} z_{r+1}^i = \sum_{i=0}^{j} \hat{c}_i^{(r+1)} z_{r+1}^{j-i} = z_{r+1}^j \sum_{i=0}^{j} \hat{c}_i^{(r+1)}\overline{z_{r+1}}^i = -z_{r+1}^j \sum_{i=j+1}^{\infty} \hat{c}_i^{(r+1)}\overline{z_{r+1}}^i$$

using $\hat{c}^{(r+1)}(\overline{z_{r+1}}) = \sum_{j=0}^{\infty} \hat{c}_j^{(r+1)}\overline{z_{r+1}}^j = 0$. Therefore

$$\begin{aligned}
\sum_{j=0}^{\infty} j^{1/2+L-r-1}\|c_{j,\bullet}^{(r+1)}\| &= \sum_{j=0}^{\infty} j^{1/2+L-r-1}\left\|-z_{r+1}^j \sum_{i=j+1}^{\infty} \hat{c}_i^{(r+1)}\overline{z_{r+1}}^i\right\| \\
&\leq \sum_{j=0}^{\infty}\sum_{i=j+1}^{\infty} j^{1/2+L-r-1}\|\hat{c}_i^{(r+1)}\| \leq \sum_{i=1}^{\infty}\sum_{j=0}^{i-1} j^{1/2+L-r-1}\|\hat{c}_i^{(r+1)}\| \\
&\leq \sum_{i=1}^{\infty}(i-1)i^{1/2+L-r-1}\|\hat{c}_i^{(r+1)}\| < \infty,
\end{aligned}$$

using the induction hypothesis. Setting $r = 0$ and $c^{(0)}(z) = c(z)$ starts the induction and the above arguments show the induction step. For $r = H$ the following representation is obtained

$$c(z) = D(z)c_\bullet(z) + z\sum_{k=1}^{H} \frac{\tilde{D}_{\neg k}(z)}{\overline{z_k}\tilde{D}_{\neg k}(\overline{z_k})}\tilde{C}_k\tilde{K}_k.$$

Evaluating the above representation at z_k, $k = 1, \dots, H$ shows uniqueness up to the decomposition $\tilde{C}_k\tilde{K}_k$.

What is left to show is how to obtain the real valued decomposition from the above decomposition formulated for complex quantities. In order to do so note first that terms corresponding to complex conjugate roots occur in complex conjugate pairs, i.e. for each index $k \leq l$ corresponding to a complex root, there exists an index $j > l$ such that $\overline{z_k} = z_j$ and it holds that

$$z\frac{\tilde{D}_{\neg k}(z)}{\overline{z_k}\tilde{D}_{\neg k}(\overline{z_k})}\tilde{C}_k\tilde{K}_k = \overline{\overline{z}\frac{\tilde{D}_{\neg j}(\overline{z})}{\overline{z_j}\tilde{D}_{\neg j}(\overline{z_j})}\tilde{C}_j\tilde{K}_j}$$

since $\tilde{C}_k\tilde{K}_k = c(\overline{z_k}) = \overline{c(\overline{z_j})} = \overline{\tilde{C}_j\tilde{K}_j}$. Noting that $\tilde{D}_{\neg j}(z) = D_{\neg j}(z)(1 - zz_k)$ we obtain

$$\frac{\tilde{D}_{\neg k}(z)}{\overline{z_k}\tilde{D}_{\neg k}(\overline{z_k})}\tilde{C}_k\tilde{K}_k + \frac{\tilde{D}_{\neg j}(z)}{\overline{z_j}\tilde{D}_{\neg j}(\overline{z_j})}\tilde{C}_j\tilde{K}_j = D_{\neg k}(z)\left[\frac{\tilde{C}_k\tilde{K}_k(1-zz_j)}{(1-z_j^2)\overline{z_k}D_{\neg k}(\overline{z_k})} + \frac{\tilde{C}_j\tilde{K}_j(1-zz_k)}{(1-z_k^2)\overline{z_j}D_{\neg j}(\overline{z_j})}\right].$$

To obtain the real valued expression given in the formulation of the lemma define for complex roots:

$$C_k := \left[\mathcal{R}\{z_k\tilde{C}_k/[(1-z_j^2)D_{\neg k}(\overline{z_k})]\}, \quad \mathcal{I}\{z_k\tilde{C}_k/[(1-z_j^2)D_{\neg k}(\overline{z_k})]\}\right], K_k := \begin{bmatrix} 2\mathcal{R}(\tilde{K}_k) \\ -2\mathcal{I}(\tilde{K}_k) \end{bmatrix}$$

where $\mathcal{R}$ and $\mathcal{I}$ denote the real and the imaginary part of a complex quantity. For real roots define $C_k = \tilde{C}_k/(z_k D_{\neg k}(z_k)), K_k = \tilde{K}_k$ noting that for real roots $\tilde{D}_{\neg k}(z) = D_{\neg k}(z)$ holds. Note finally that due to the fact that the coefficients $c_j(z)$ and $D(z)$ are real, also the coefficients $c_{j,\bullet}(z)$ are real, which completes the proof of the lemma. □

Lemma 3 *For $k = 1, \ldots, l$ let $x_{t+1,k} = J_k x_{t,k} + K_k\varepsilon_t$, where $x_{1,k} = 0$. Here J_k corresponds to $z_k := e^{i\omega_k}$ and is defined in equations (5) and $(\varepsilon_t)_{t\in\mathbb{Z}}$ fulfills Assumption 2. Assume furthermore that $[K_k, J_k K_k]$ has full row rank. Let $v_t := c(L)\varepsilon_t$ for $c(z) = \sum_{j=0}^{\infty} c_j z^j$ with $\sum_{j=0}^{\infty} j^{3/2}\|c_j\| < \infty$, hence $c(z)$ is analytic on the closed unit disc.*

We denote the stacked process as $x_t := \left[x'_{t,1}, \ldots, x'_{t,l}\right]'$ and define $G^v(j) := \langle v_t, v_{t-j}\rangle_{j+1}^T$ for $j = 1, \ldots, T-1$, $G^v(j) := G^v(-j)'$ for $j = -T+1, \ldots, -1$, $G^v(j) := 0$ for $|j| \geq T$ and $\Gamma^v(j) := \mathbb{E}v_t v'_{t-j}$. Introduce furthermore the following notation: $H_T = o((T/\log T)^{1/2})$ and $G_T = (\log T)^a$ for $0 < a < \infty$.

(i) Then we obtain:

$$\max_{0\leq j\leq H_T} \|G^v(j) - \Gamma^v(j)\| = O((\log T/T)^{1/2}), \tag{10}$$

$$T^{-1}\langle x_t, x_t\rangle_1^T \xrightarrow{d} W, \tag{11}$$

$$\max_{0\leq j\leq H_T} \|\langle x_t, v_{t-j}\rangle_{j+1}^T\| = O_P(1). \tag{12}$$

Here it holds that $W > 0$ a.s. and thus it follows that $\left[T^{-1}\langle x_t, x_t\rangle_1^T\right]^{-1} = O_p(1)$, *where* $\xrightarrow{d}$ *denotes convergence in distribution.*

(ii) If $c(z)$ is a rational function, then $\max_{0\leq j\leq G_T}\|G^v(j)-\Gamma^v(j)\| = O((T^{-1}\log\log T)^{1/2})$.

(iii) If the processes $(x_t)_{t\in\mathbb{Z}}$ and $(v_t)_{t\in\mathbb{Z}}$ are corrected for mean and harmonic components, the above results remain true for the processes $(\hat{x}_t)_{t\in\mathbb{Z}}$ defined analogously to $(\hat{v}_t)_{t\in\mathbb{Z}}$ in Theorem 2. The definition of W in (11) has to be changed appropriately in this case.

Proof:

Proof of (i): Equation (10) follows immediately from Theorem 7.4.3 (p. 326) in HD. The assumptions concerning summability and the supremum required in that theorem are guaranteed in our framework since we require summability with a factor $j^{3/2}$ and also our assumptions on the noise $(\varepsilon_t)_{t\in\mathbb{Z}}$ are sufficient.

The second result (11) follows from Theorem 2.2 on p. 372 of Chan and Wei (1988) and the continuous mapping theorem. Chan and Wei (1988) only consider univariate processes, however, the Cramer-Wold device allows for a generalization to the multivariate case. Using the notation of Chan and Wei (1988), the required components $t_1,\dots,t_{2l}$ of the random vector $X_n(u,v,t_1,\dots,t_{2l})$ are essentially equal to $\sqrt{2}\sum_{s=1}^{t_i}\sin(\theta_k s)\varepsilon_s$ and $\sqrt{2}\sum_{s=1}^{t_i-1}\cos(\theta_k s)\varepsilon_s$. Now,

$$\begin{aligned} x_{t,k} &= \sum_{s=1}^{t-1} J_k^{s-1}K_k\varepsilon_{t-s} = J_k^{t-1}\sum_{s=1}^{t-1} J_k^{s-t}K_k\varepsilon_{t-s} = J_k^{t-1}\sum_{s=1}^{t-1} J_k^{-s}K_k\varepsilon_s \\ &= J_k^{t-1}\sum_{s=1}^{t-1}\left(\begin{bmatrix}\cos(\omega_k s) & -\sin(\omega_k s)\\ \sin(\omega_k s) & \cos(\omega_k s)\end{bmatrix}\otimes I_{\tilde{c}_k}\right)\begin{pmatrix} u_s^1 \\ u_s^2\end{pmatrix}, \end{aligned}$$

where $[(u_s^1)',(u_s^2)']' = K_k\varepsilon_s$. Thus, for any $x\neq 0$, $x'J_k^{1-t}x_{t,k}$ is composed of the terms collected in X_n of Chan and Wei (1988). Therefore the Cramer-Wold device combined with Theorem 2.2 of Chan and Wei (1988) shows convergence of $J_k^{1-t}x_{t,k}$, when scaled by $T^{-1/2}$, to a multivariate Brownian motion.

For establishing non-singularity of the limiting distribution it is sufficient to look at each unit root separately (cf. Theorem 3.4.1 of Chan and Wei, 1988, p. 392, which establishes asymptotic uncorrelatedness of the components corresponding to different unit roots). In case of real unit roots non-singularity follows immediately from full rank of K_k in that case. For complex unit roots the arguments are more involved and the proof proceeds indirectly. Chan and Wei (1988) show that the sine and cosine terms (for any given frequency ω_k) are asymptotically uncorrelated, irrespective of the properties of the noise process. This implies that the diagonal block of the limiting variance of the Brownian motion that corresponds to a given unit root (i.e. that corresponds to $J_k^{1-t}x_{t,k}$) is singular, if and only if there exists a non-zero vector $x' = [x_1', x_2']'$ such that the variances of both $x_1'u_s^1 + x_2'u_s^2$ (corresponding to the cosine terms) and of $x_1'u_s^2 - x_2'u_s^1$ (corresponding to the sine terms) are zero. This is equivalent to $x'[K_k, J_k'K_k] = 0$. The latter matrix has full rank by assumption and thus the contradiction is shown.

Now the continuous mapping theorem can be applied to show that

$$T^{-1}\langle x_t, x_t\rangle_1^T = T^{-1}\sum_{t=1}^{T} J^{t-1}(J^{1-t}x_t/\sqrt{T})(J^{1-t}x_t/\sqrt{T})'(J^{t-1})' \xrightarrow{d} W.$$

The a.s. non-singularity of W follows from the non-singularity of the limiting covariance matrices of $J^{1-t}x_t/\sqrt{T}$. Since W is a continuous function of a Brownian motion, it has a density with respect to the Lebesgue measure (i.e. it is absolutely continuous with respect to the Lebesgue measure). Therefore, for each $\eta > 0$ there exists an $\varepsilon > 0$, such that $\mathbb{P}\{\lambda_{min}(W) > \varepsilon\} = \mathbb{P}\{\|W^{-1}\|_2 > \varepsilon^{-1}\} < \eta$, where $\lambda_{min}(W)$ denotes a minimal eigenvalue of W. Due to the convergence in distribution it holds that $\mathbb{P}\{\lambda_{min}(T^{-1}\langle x_t, x_t\rangle_1^T) > \varepsilon\} \to \mathbb{P}\{\lambda_{min}(W) > \varepsilon\}$ showing that $[T^{-1}\langle x_t, x_t\rangle_1^T]^{-1} = O_P(1)$.

The bounds formulated in (12) are derived for each k separately. Thus, fix k for the moment and assume that $0 < \omega_k < \pi$, since for real unit roots the result follows analogously and is thus not derived separately. Applying Lemma 2 with $l = 1$ to $c(z)$ and using $z^2 = \Delta_{\omega_k}(z) - 1 + 2\cos(\omega_k)z$ we obtain $c(z) = \alpha_1 + \alpha_2 z + \Delta_{\omega_k}(z)c_\bullet(z)$, where $c_\bullet(z) = \sum_{j=0}^{\infty} c_{j,\bullet}z^j$ is such that $\sum j^{1/2}\|c_{j,\bullet}\| < \infty$. Using this decomposition we obtain $v_t = c(L)\varepsilon_t = \alpha_1\varepsilon_t + \alpha_2\varepsilon_{t-1} + \Delta_{\omega_k}(L)v_t^*$, with $v_t^* := c_\bullet(L)\varepsilon_t$. Hence

$$\begin{aligned}\langle x_{t,k}, v_{t-j}\rangle_{j+1}^T &= J_k^j\langle x_{t-j,k}, v_{t-j}\rangle_{j+1}^T + \sum_{i=0}^{j-1} J_k^i K_k\langle \varepsilon_{t-i-1}, v_{t-j}\rangle_{j+1}^T \\ &= J_k^j\langle x_{t-j,k}, v_{t-j}\rangle_{j+1}^T + O(j(\log T/T)^{1/2}) + O(1).\end{aligned}$$

Here the first equality stems immediately from the definition of $x_{t,k}$. The second equality follows from $\langle \varepsilon_{t-i-1}, v_{t-j}\rangle_{j+1}^T = O((\log T/T)^{1/2})$ for $i < j-1$ and $\langle \varepsilon_{t-j}, v_{t-j}\rangle_{j+1}^T = O(1)$.

This last result follows from the uniform convergence of the estimated covariance sequence, i.e. by applying (10) to the stacked process $([\varepsilon_t', v_t']')_{t\in\mathbb{Z}}$. Here it has to be noted that the difference between $\langle \varepsilon_{t-j}, v_{t-j}\rangle_{j+1}^T$ and $\langle \varepsilon_t, v_t\rangle_1^T$ is of order $O((\log T/T)^{1/2})$. Now, the assumption that $0 \leq j \leq H_T$ implies that the two O(.)-terms above are uniformly $O(1)$ for $0 \leq j \leq H_T$. This shows that the essential term that has to be investigated further is $T^{-1}\sum_{t=1}^{T-j} x_{t,k}v_t'$. This term can be developed as follows:

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T-j} x_{t,k}v_t' &= \frac{1}{T}\sum_{t=1}^{T-j} x_{t,k}(\alpha_1\varepsilon_t + \alpha_2\varepsilon_{t-1} + \Delta_{\omega_k}(L)v_t^*)' \\ &= \frac{1}{T}\sum_{t=1}^{T-j} x_{t,k}\varepsilon_t'\alpha_1' + \frac{1}{T}\sum_{t=1}^{T-j} x_{t,k}\varepsilon_{t-1}'\alpha_2' + \frac{1}{T}\sum_{t=1}^{T-j} x_{t,k}(v_t^* - 2\cos(\omega_k)v_{t-1}^* + v_{t-2}^*)'.\end{aligned}$$

Chan and Wei (1988) show in their Theorem 2.4 that random variables similar to the first two terms above converge for $j = 0$, i.e. when summation takes place from 1 to T (see also Theorem 6 of Johansen and Schaumburg, 1999). Thus, concerning the first two terms above it remains to characterize the behavior for $1 \leq j \leq H_T$. Note again that c_k denotes the dimension of $x_{t,k}$ and consider the difference between the expressions for $j = 0$ and $j \neq 0$, which is for the first term equal to $T^{-1}\sum_{t=T-j+1}^{T} x_{t,k}\varepsilon_t'\alpha_1'$. We obtain that

$$\begin{aligned}\mathbb{E}\max_{1\leq j\leq H_T}\|T^{-1}\sum_{t=T-j+1}^{T}\text{vec}(x_{t,k}\varepsilon_t')\|_1 &\leq \frac{\sqrt{sc_k}}{T}\max_{1\leq j\leq H_T}\sum_{t=T-j+1}^{T}(\mathbb{E}\|x_{t,k}\varepsilon_t'\|_{Fr}^2)^{1/2} \\ &\leq \frac{\sqrt{sc_k}}{T}\max_{1\leq j\leq H_T}\sum_{t=T-j+1}^{T}(\mathbb{E}\|x_{t,k}\|_2^2\mathbb{E}\|\varepsilon_t\|_2^2)^{1/2} \\ &\leq \frac{c}{T}\max_{1\leq j\leq H_T}\sum_{t=T-j+1}^{T} t^{1/2} \leq \frac{c}{T}\sum_{t=T-H_T+1}^{T} t^{1/2} \leq \frac{cH_T}{\sqrt{T}} \to 0,\end{aligned}$$

where we have used the inequality $\mathbb{E}\|x_{t,k}\varepsilon_t'\|_{Fr}^2 \leq \mathbb{E}\|x_{t,k}\|_2^2\mathbb{E}\|\varepsilon_t\|_2^2$, which follows from $\mathbb{E}\{\varepsilon_t\varepsilon_t'|\mathcal{F}_{t-1}\} = \Sigma$ using $\mathbb{E}\|x_{t,k}\varepsilon_t'\|_{Fr}^2 = \mathbb{E}(x_{t,k}'x_{t,k}\varepsilon_t'\varepsilon_t) = \mathbb{E}[x_{t,k}'x_{t,k}\text{tr}(\varepsilon_t\varepsilon_t')]$. Therefore we have established uniform convergence in $0 \leq j \leq H_T$ of $T^{-1}\sum_{t=1}^{T-j} x_{t,k}\varepsilon_t'$ to a random variable. Similar arguments apply to the second term above and thus only the third term has to be investigated further. The third term is equal to

$$
\begin{aligned}
&\tfrac{1}{T}\textstyle\sum_{t=1}^{T-j} x_{t,k}(v_t^* - 2\cos(\omega_k)v_{t-1}^* + v_{t-2}^*)' = \tfrac{1}{T}\sum_{t=1}^{T-j} x_{t,k}(v_t^*)' - 2\cos(\omega_k)x_{t,k}(v_{t-1}^*)' + x_{t,k}(v_{t-2}^*)' \\
&= \tfrac{1}{T}\textstyle\sum_{t=1}^{T-j} x_{t,k}(v_t^*)' - 2\cos(\omega_k)\tfrac{1}{T}\sum_{t=1}^{T-j} x_{t,k}(v_{t-1}^*)' + \tfrac{1}{T}\sum_{t=1}^{T-j} x_{t,k}(v_{t-2}^*)' \\
&= \tfrac{1}{T}\textstyle\sum_{t=1}^{T-j} x_{t,k}(v_t^*)' - 2\cos(\omega_k)\tfrac{1}{T}\sum_{t=1}^{T-j-1} x_{t+1,k}(v_t^*)' + \tfrac{1}{T}\sum_{t=1}^{T-j-2} x_{t+2,k}(v_t^*)' + o(1) \\
&= \tfrac{1}{T}\textstyle\sum_{t=1}^{T-j-2}(x_{t,k} - 2\cos(\omega_k)x_{t+1,k} + x_{t+2,k})(v_t^*)' \\
&\qquad + \tfrac{1}{T}\left[x_{T-j-1,k}(v_{T-j-1}^*)' + x_{T-j,k}(v_{T-j}^*)' - 2\cos(\omega_k)x_{T-j,k}(v_{T-j-1}^*)'\right] + o(1) \\
&= \tfrac{1}{T}\textstyle\sum_{t=1}^{T-j-2}(K_k\varepsilon_{t+1} - J_k' K_k\varepsilon_t)(v_t^*)' + o(1).
\end{aligned}
$$

Here the first $o(1)$ term comes from the omission of three initial terms and the second $o(1)$ term holds uniformly in $0 \leq j \leq H_T$, as can be shown as follows: Due to the law of the iterated logarithm (see Theorem 4.9 on p. 125 of Hall and Heyde, 1980) $T^{-1}x_T = O((T^{-1}\log\log T)^{1/2})$ and $v_T^* = o(T^{1/4})$ due to ergodicity and the existence of the fourth moments of ε_t and thus of v_t^*. This together implies that $T^{-1}x_{T-j,k}(v_{T-j}*)' = o(T^{-1}(T-j)^{3/4}\sqrt{\log\log(T-j)}) = o(1)$. In the sum in the final line above only stationary quantities appear. Therefore, using (10) this term is $O(1)$ uniformly in $0 \leq j \leq H_T$, which concludes the proof of (12).

Proof of (ii): The sharper result for the case of rational transfer functions is given in Theorem 5.3.2 (p. 167) in HD.

Proof of (iii): Recall the definition of the variable $\hat{x}_t = x_t - \langle x_t, d_t\rangle_1^T(\langle d_t, d_t\rangle_1^T)^{-1}d_t$, with d_t defined in Theorem 2. Note first that it follows from the law of the iterated logarithm (see e.g. Theorem 4.7 on p. 117 of Hall and Heyde, 1980) that $\langle v_t, d_t\rangle_1^T = O((T^{-1}\log\log T)^{1/2})$. This fact allows to derive equation (10) also for $\hat{v}_t$ and furthermore this observation also allows to derive the stronger bound (which is of exactly this order) in item (ii) of the lemma.

Let us next turn to establishing equation (11) for $\hat{x}_t$. This result follows from the observation that for all $k = 1, \dots, l$ it holds using the continuous mapping theorem and the results achieved in the proof of (i) that $T^{-1/2}\langle J^{1-t}x_t, d_{t,k}\rangle_1^T = O_P(1)$. Nonsingularity of the limiting variable W follows from nonsingularity of W in (ii) and the properties of Brownian motions. For details see Johansen and Schaumburg (1999).

We are thus left to establish (12) for $\hat{x}_t$ and $\hat{v}_{t-j}$. In order to do so note that

$$\langle \hat{x}_t, \hat{v}_{t-j}\rangle_{j+1}^T = \langle x_t, v_{t-j}\rangle_{j+1}^T + \langle x_t, \hat{v}_{t-j} - v_{t-j}\rangle_{j+1}^T + \langle \hat{x}_t - x_t, \hat{v}_{t-j}\rangle_{j+1}^T.$$

The first term above is dealt with in item (i). Next note $v_{t-j} - \hat{v}_{t-j} = \langle v_t, d_t\rangle_1^T(\langle d_t, d_t\rangle_1^T)^{-1}d_{t-j}$. As just mentioned above $\langle v_t, d_t\rangle_1^T = O((T^{-1}\log\log T)^{1/2})$ and it furthermore holds that

$T^{1/2}\langle v_t, d_t\rangle_1^T$ converges in distribution (to a normally distributed random variable). It is easy to show, analogously to $\max \|\langle x_t, v_{t-j}\rangle_{j+1}^T\| = O_P(1)$, that also $\max T^{-1/2}\|\langle x_t, d_{t-j}\rangle_{j+1}^T\| = O_P(1)$. This implies that the second term above fulfills the required constraint on the order and we are left with the third term, which can be rewritten as $-\langle x_t, d_t\rangle_1^T(\langle d_t, d_t\rangle_1^T)^{-1}\langle d_t, \hat{v}_{t-j}\rangle_{j+1}^T$. From above we know that $T^{-1/2}\langle x_t, d_t\rangle_1^T = O_P(1)$ and also $(\langle d_t, d_t\rangle_1^T)^{-1} = O(1)$. Using $\langle d_t, \hat{v}_t\rangle_1^T = 0$ and $d_t = J^j d_{t-j}$ we obtain $T^{1/2}\langle d_t, \hat{v}_{t-j}\rangle_{j+1}^T = -T^{-1/2}J^j\sum_{t=T-j+1}^T d_t\hat{v}_t'$. Now

$$\left\| \max_{1\leq j\leq H_T} T^{-1/2}J^j \sum_{t=T-j+1}^{T} d_t\hat{v}_t' \right\| \leq T^{-1/2}\sum_{t=T-H_T+1}^{T} l\|\hat{v}_t\|$$

since $\|J^j\| \leq 1$ and $\|d_t\| \leq l$ by definition. Since $\mathbb{E}\|\hat{v}_t\| = O(1)$ as is easy to verify $H_T/T^{1/2} \to 0$ shows that this term is uniformly in $j = o((T/\log T)^{1/2})$ of order $o_P(1)$. Therefore we have established $\max_{0\leq j\leq H_T}\|\langle \hat{x}_t, \hat{v}_{t-j}\rangle_{j+1}^T\| = O_P(1)$. □

Remark 2 *Although in the formulation of the lemma we assume $x_{1,k} = 0$, it is straightforward to verify that all results hold unchanged if $x_{1,k}$ is instead given by any random variable.*

We present one more preliminary lemma without proof, the well known matrix inversion lemma, for convenience of reference.

Lemma 4 *For any nonsingular symmetric matrix $X \in \mathbb{R}^{m\times m}$ and any partitioning into blocks A, B, C it holds that*

$$X^{-1} = \begin{bmatrix} A & B \\ B' & C \end{bmatrix}^{-1} = \begin{bmatrix} 0 & 0 \\ 0 & C^{-1} \end{bmatrix} + \begin{bmatrix} I \\ -C^{-1}B \end{bmatrix}\left(A - BC^{-1}B'\right)^{-1}\left[I, -B'C^{-1}\right].$$

Furthermore X is invertible if and only if C and $A - BC^{-1}B'$ are. In this case also the following inequality holds true:

$$\|\left(A - BC^{-1}B'\right)^{-1}\|_2 \leq \|X^{-1}\|_2 + \|C^{-1}\|_2. \tag{13}$$

B Proofs of the Theorems

In this appendix the proofs of the theorems in the main text are given.

B.1 Proof of Theorem 2

The proof of the theorem is based on the theory presented in Chapters 6 and 7 of HD. The difference is that HD consider the YW estimator, whereas we consider the LS estimator. The general strategy of the proof is thus to establish that the results apply also to the LS estimator by showing that the differences that occur are asymptotically sufficiently small. As a side remark note here that HD use the symbol ‘ $\hat{.}$ ’ for the YW estimator, whereas we use it for the LS estimator. In this paper the YW estimators carry the symbol ‘$\tilde{.}$’. Note for completeness that this symbol is also used in other contexts, where, however, no confusion should arise.

Proof of (i), (ii): In Theorem 7.4.5 (p. 331) of HD it is shown that $\max_{1\leq j\leq p}\|\tilde{\Phi}_p(j)-\Phi_p(j)\| = O((\log T/T)^{1/2})$ uniformly in $p = o((T/\log T)^{1/2})$. The tighter bound for the case of rational $c(z)$ is derived in Theorem 6.6.1 (p. 259) of HD. Thus, to establish the results also for the LS estimator it has to be shown that the difference between $\tilde{\Phi}_p(j)$ and $\hat{\Phi}_p(j)$ is ‘small enough’ asymptotically. For example for the first bound this means that we have to show that $\max_{1\leq p\leq H_T}\max_{1\leq j\leq p}\|\tilde{\Phi}^v_p(j)-\hat{\Phi}^v_p(j)\| = O((\log T/T)^{1/2})$ and the correspondingly tighter bound for the rational case. In order to show this we consider the difference between the YW and the LS estimator. In the LS estimator $\hat{\Theta}^v_p$, quantities of the form $\langle v_{t-i}, v_{t-j}\rangle = T^{-1}\sum_{t=p+1}^T v_{t-i}v'_{t-j}$ for $i,j=1,\dots,p$ appear, whereas the YW estimator uses $G^v(j-i) = T^{-1}\sum_{t=1+j-i}^T v_t v'_{t-j+i}$ for $j\geq i$ and the corresponding similar expression for $j<i$. Thus, the difference between these two terms is given (discussing here only the case $j\geq i$; with the case $i>j$ following analogously) by:

$$\langle v_{t-i}, v_{t-j}\rangle - G^v(j-i) = -\frac{1}{T}\sum_{t=1+j-i}^{p-i} v_t v'_{t-j+i} - \frac{1}{T}\sum_{t=T-i+1}^{T} v_t v'_{t-j+i}.$$

We know from equation (10) in Lemma 3 that $T^{-1}\sum_{t=1+r}^T v_t v'_{t-r} - \Gamma^v(r) = O((\log T/T)^{1/2})$ uniformly in $0\leq r\leq H_T$. This directly implies $\langle v_{t-i}, v_{t-j}\rangle - G^v(j-i) = O((\log T/T)^{1/2})$ uniformly in $0\leq i,j\leq H_T$ and $1\leq p\leq H_T$. Next note that $\|\langle V^-_{t,p}, V^-_{t,p}\rangle^{-1}\|_\infty$ and $\|\langle v_t, V^-_{t,p}\rangle\|_\infty$ are uniformly bounded in $1\leq p\leq H_T$, which follows from the bound derived above, equation (10) and $\max_{1\leq p\leq H_T}\|(\mathbb{E}V^-_{t,p}(V^-_{t,p})')^{-1}\|_\infty<\infty$ and $\max_{1\leq p\leq H_T}\|\mathbb{E}v_t(V^-_{t,p})'\|_\infty<\infty$, see Theorem 6.6.11 (p. 267–268) in HD. This shows (i) by calculating the difference between the YW and the LS estimators.

In case of rational $c(z)$, the same type of argument as above but using Lemma 3(ii) instead of (i) leads to the tighter bound $\langle v_{t-i}, v_{t-j}\rangle - G^v(j-i) = O((\log\log T/T)^{1/2})$ for $p \leq G_T = (\log T)^a$ for $a < \infty$. This proves the bounds on the estimation error for $\hat{\Phi}_p^v(j)$ given in the theorem.

Proof of (iii): The approximation results to $IC^v(p; C_T)$ as stated in Theorem 7.4.7 (p. 332) of HD, which is based upon Hannan and Kavalieris (1986), are formulated for the YW estimator. Again a close inspection of the proofs of the underlying theorems forms the basis for the adaption of the results to the LS estimator.

Some main ingredients required for Theorem 7.4.7 are derived in Theorem 7.4.6 (p. 331) of HD (also dealing with the YW estimator). Inspection of the proof of this theorem shows that the key element is equation (7.4.31) on p. 340. It is sufficient to verify that this relationship concerning the properties of autoregressive approximations also holds for the LS estimator. In other words, if this equation is verified for the LS estimator, then Theorems 7.4.6 and 7.4.7 of HD stated for the YW estimator also hold for the LS estimator.

Therefore, denote as in HD $\tilde{g}(j,k) := T^{-1}\sum_{t=1}^T v_{t-j}v_{t-k}'$ and $\hat{u}_k := T^{-1}\sum_{t=1}^T \varepsilon_t v_{t-k}'$. Then we obtain

$$\tilde{g}(j,k) = \langle v_{t-j}, v_{t-k}\rangle + T^{-1}\sum_{t=1}^{p} v_{t-j}v_{t-k}' = \langle v_{t-j}, v_{t-k}\rangle + o(\frac{p}{T}j^{1/4}k^{1/4})$$

uniformly in $j,k \leq p$, which follows from the assumption of finite fourth moments of $(v_t)_{t\in\mathbb{Z}}$. Further $p^{-1}\sum_{t=1}^p(\sum_{j=1}^p(\hat{\Phi}_p^v(j) - \Phi^v(j))v_{t-j})v_{t-k}' = O(1)$ is easy to verify from the convergence of $\hat{\Phi}_p^v(j)$, the summability of $\Phi^v(j)$ and the uniform boundedness of $p^{-1}\sum_{t=-p}^p v_t v_t'$ (which follows from ergodicity of $(v_t)_{t\in\mathbb{Z}}$). This implies due to the assumptions concerning the upper bounds on the number of lags, the uniform error bound on the autoregressive coefficients and $\varepsilon_t = \sum_{j=0}^\infty \Phi^v(j)v_{t-j}$ that:

$$\begin{aligned}
\sum_{j=1}^p \left(\hat{\Phi}_p^v(j) - \Phi^v(j)\right)\tilde{g}(j,k) &= \sum_{j=1}^p \left(\hat{\Phi}_p^v(j) - \Phi^v(j)\right)\langle v_{t-j}, v_{t-k}\rangle + o(T^{-1/2}) \\
&= -\langle \varepsilon_t, v_{t-k}\rangle + \sum_{j=p+1}^\infty \Phi^v(j)\frac{1}{T}\sum_{t=p+1}^T v_{t-j}v_{t-k}' + o(T^{-1/2}) \\
&= -\hat{u}_k + \sum_{j=p+1}^\infty \Phi^v(j)\left[\tilde{g}(j,k) + o(pj^{1/2}T^{-1})\right] + o(T^{-1/2})
\end{aligned}$$

$$= \quad -\hat{u}_k + \sum_{j=p+1}^{\infty} \Phi^v(j)\tilde{g}(j,k) + o(T^{-1/2})$$

due to $\sum_{j=1}^{\infty} j^{1/2}\|\Phi^v(j)\| < \infty$ and $p/T^{1/2} \to 0$ by assumption. This establishes HD's equation (7.4.31) also for the LS estimator. Thus, their Theorem 7.4.6 continues to hold without changes also for the LS estimator. Since the proof of Theorem 7.4.7 in HD does not use any properties of the estimator exceeding those established in Theorem 7.4.6 it follows that also this theorem holds for the LS estimator. Only certain assumptions on the noise $(\varepsilon_t)_{t\in\mathbb{Z}}$ (see the formulation of Theorem 7.4.7 for details), which hold in our setting (cf. Assumption 2), are required.

Proof of (iv): The result is contained in Theorem 6.6.3 (p. 261) of HD for the YW estimator. Inspection of the proof shows that two quantities have to be changed to adapt the theorem to the LS estimator. The first is the definition of F on p. 274 of HD, which has to be modified appropriately when using the LS instead of the YW estimator. The second is the replacement of G_h in the proof by $\langle V_{t,p}^-, V_{t,p}^-\rangle$, where our $V_{t,p}^-$ corresponds to HD's $y(t,h)$. All arguments in the proof remain valid with these modifications.

Proof of (v): This item investigates the effect of including components of v_{t-p} as regressors in the autoregression of order $p-1$, which is equivalent to the exclusion of certain components of v_{t-p} in the autoregression of order p. This evident observation is exactly what is reflected in the results. Denote with $\tilde{V}_{t,p}^-$ the regressor vector $V_{t,p-1}^-$ augmented by $\tilde{P}_s v_{t-p}$. Note that in this proof '$\tilde{.}$' is used to denote quantities relating to the augmented regression and not to the YW estimators. Using the block-matrix inversion formula and (13) from Lemma 4 (with the blocks corresponding to $V_{t,p-1}^-$ and $\tilde{P}_s v_{t-p}$) it is straightforward to show that $\|\langle \tilde{V}_{t,p}^-, \tilde{V}_{t,p}^-\rangle^{-1}\| < \infty$ and $\|\langle \tilde{V}_{t,p}^-, \tilde{V}_{t,p}^-\rangle^{-1}\|_\infty < \infty$ a.s. for T large enough, uniformly in $1 \le p \le H_T$ This can be used to show the approximation properties of the autoregression including $\tilde{P}_s v_{t-p}$ as follows:

$$\begin{aligned}
\tilde{\Theta}_p^v &:= \langle v_t, \tilde{V}_{t,p}^-\rangle\langle \tilde{V}_{t,p}^-, \tilde{V}_{t,p}^-\rangle^{-1} \\
&= \mathbb{E}v_t(\tilde{V}_{t,p}^-)'(\mathbb{E}\tilde{V}_{t,p}^-(\tilde{V}_{t,p}^-)')^{-1} + \left[\langle v_t, \tilde{V}_{t,p}^-\rangle - \mathbb{E}v_t(\tilde{V}_{t,p}^-)'\right]\langle \tilde{V}_{t,p}^-, \tilde{V}_{t,p}^-\rangle^{-1} + \\
&\quad \mathbb{E}v_t(\tilde{V}_{t,p}^-)'(\mathbb{E}\tilde{V}_{t,p}^-(\tilde{V}_{t,p}^-)')^{-1}\left[\mathbb{E}\tilde{V}_{t,p}^-(\tilde{V}_{t,p}^-)' - \langle \tilde{V}_{t,p}^-, \tilde{V}_{t,p}^-\rangle\right]\langle \tilde{V}_{t,p}^-, \tilde{V}_{t,p}^-\rangle^{-1}.
\end{aligned}$$

Now applying the derived uniform bounds on the estimation errors in $\langle v_{t-j}, v_{t-k}\rangle - \mathbb{E}v_{t-j}v_{t-k}'$ shows the result. With the appropriate bounds on the lag lengths, both the result for the general and the sharper result for the rational case follow.

The next point discussed is the effect of the inclusion of $\tilde{P}_s v_{t-p}$ on the approximation formula derived for $\widetilde{IC}^v(p; C_T)$. By construction it holds that

$$\hat{\Sigma}^v_{p-1} = \langle v_t - \hat{\Theta}^v_{p-1} V^-_{t,p-1}, v_t - \hat{\Theta}^v_{p-1} V^-_{t,p-1} \rangle^T_p \geq \tilde{\Sigma}^v_p = \langle v_t - \tilde{\Theta}^v_p \tilde{V}^-_{t,p}, v_t - \tilde{\Theta}^v_p \tilde{V}^-_{t,p} \rangle^T_{p+1} \geq \hat{\Sigma}^v_p.$$

Adding the penalty term $ps^2 C_T/T$ does not change the inequalities. Then the approximation result under (iii) shows the claim.

Proof of (vi): We have shown in (iv) of Lemma 3 that the inclusion of the deterministic components does not change the convergence properties of the estimated autocovariance sequence. This implies that all the statements of the theorem related to the properties of the autoregressive approximations remain valid unchanged.

Concerning the evaluation of $IC^v(p; C_T)$ it is stated in HD on p. 330 that the inclusion of the deterministic components (i.e. mean and harmonic components) does not change the result. From this it also follows immediately that the asymptotic properties of $\hat{p}_{BIC}$ are not influenced, since that result stems entirely from the approximation derived for $IC^v(p; C_T)$ and the decrease in Σ^v_p as a function of p, which also does not depend upon the considered deterministic components. □

B.2 Proof of Lemma 1

Proof of (i): The starting point is the representation derived in Theorem 1. The properties of $Z^-_{t,p}$ are straightforward to verify using $(C^\perp)'C = 0$ and $(C^\dagger)'y_t - J(C^\dagger)'y_{t-1} = x_t + (C^\dagger)'e_t - Jx_{t-1} - J(C^\dagger)'e_{t-1}$. These relationships also immediately establish the expression given for $y_t - CJ(C^\dagger)'y_{t-1}$ and thus also the definition of $\tilde{c}_\bullet(z)$. Furthermore, $\tilde{c}_\bullet(0) = I_s$ follows immediately from $c_\bullet(0) = I_s$. The summability properties of $\tilde{c}_\bullet(z)$ follow directly from the analogous properties of $c_\bullet(z)$. Since $c_\bullet(z)$ has no poles inside the unit circle, neither has $\tilde{c}_\bullet(z)$, since the latter is a polynomial transformation of the former (see the definition in the formulation of the lemma). Concerning the roots of the determinant of $\tilde{c}_\bullet(z)$ note that from the representation of $\tilde{c}_\bullet(z)$ given in the theorem the following relation is obtained for $|z| < 1$:

$$\tilde{c}_\bullet(z) = \bar{C}^{-1} \begin{bmatrix} (I - zJ) & 0 \\ 0 & I \end{bmatrix} D(z)^{-1} \bar{C} c(z)$$

Now, since by assumption $\det c(z) \neq 0$ for all $|z| < 1$ and $D(z) \neq 0, |z| < 1$ it follows that $\det \tilde{c}_\bullet(z) \neq 0, |z| < 1$.

Proof of (ii): By recursive inserting it is straightforward to show that

$$\tilde{y}_{tq+i} = \begin{pmatrix} C \\ CJ \\ \vdots \\ CJ^{q-1} \end{pmatrix} x_{tq+i} + \underbrace{\begin{bmatrix} I_s & 0 & \cdots & 0 \\ CK & I_s & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ CJ^{q-2}K & \cdots & CK & I_s \end{bmatrix}}_{\mathcal{E}_q} \begin{pmatrix} \varepsilon_{tq+i} \\ \varepsilon_{tq+1+i} \\ \vdots \\ \varepsilon_{tq+q-1+i} \end{pmatrix} + \sum_{j=1}^{\infty} \bar{c}_j^{(q)} \bar{\varepsilon}_{(t-j)q+i}$$

with $\bar{\varepsilon}_t := [\varepsilon_t', \varepsilon_{t+1}', \ldots, \varepsilon_{t+q-1}']'$ and where the coefficients in $\sum_{j=1}^{\infty} \bar{c}_j^{(q)} \bar{\varepsilon}_{(t-j)q+i}$ can be obtained by cumbersome but straightforward computations. It is clear that $(\bar{\varepsilon}_{tq+i})_{t\in\mathbb{Z}}$ is a martingale difference sequence with respect to the filtration $\mathcal{F}_{jq+i}, j \in \mathbb{Z}$. To obtain the innovations representation (i.e. a representation with leading coefficient equal to the identity matrix) a renormalization has to be performed (given that in the above representation the leading coefficient is equal to $\mathcal{E}_q$). Since $\mathcal{E}_q$ is non-singular, this is achieved by setting $\tilde{\varepsilon}_t := \mathcal{E}_q \bar{\varepsilon}_t$ and correspondingly this also defines $\tilde{c}^{(q)}(z) = \sum_{j=0}^{\infty} \tilde{c}_j^{(q)} z^j = \sum_{j=0}^{\infty} \bar{c}_j^{(q)} \mathcal{E}_q^{-1} z^j$. Summability of the coefficients of $\tilde{c}^{(q)}(z)$ follows from summability of $c_\bullet(z)$. Since $\mathcal{E}_q$ is block lower triangular with diagonal blocks equal to the identity matrix it follows that the first block of $\tilde{\varepsilon}_t$ equals ε_t.

Note also that $D(z)\tilde{y}_t = \tilde{v}_t$, with $(\tilde{v}_t)_{t\in\mathbb{Z}}$ defined analogously to $(\tilde{y}_t)_{t\in\mathbb{Z}}$. Thus, the sub-sampling argument leads to q processes $(\tilde{y}_{tq+i})_{t\in\mathbb{Z}}$, $i = 1, \ldots, q$ that all have the same unit root structure as $(y_t)_{t\in\mathbb{Z}}$ and for which part (i) of the lemma can be applied, since by construction for this process $q = 1$. □

B.3 Proof of Theorem 3

The discussion above the theorem in the main text shows that the regression of $\tilde{e}_t$ on $Z_{t,p}^- := [z_t', (Z_{t,p,2}^-)']'$ has to be analyzed. Here z_t is independent of the choice of p and collects the nonstationary components.

Proof of (i): Partitioning the coefficient matrix according to the partitioning of the regressor vector we obtain:

$$[\hat{\beta}_1, \hat{\beta}_{2,p}] \;\; := \;\; \langle \tilde{e}_t, Z_{t,p}^- \rangle \langle Z_{t,p}^-, Z_{t,p}^- \rangle^{-1} \tag{14}$$

$$= \langle \tilde{e}_t, z_t^\Pi \rangle \langle z_t^\Pi, z_t^\Pi \rangle^{-1} \left[I_c, -\langle z_t, Z_{t,p,2}^- \rangle \langle Z_{t,p,2}^-, Z_{t,p,2}^- \rangle^{-1} \right] + \left[0^{s \times c}, \langle \tilde{e}_t, Z_{t,p,2}^- \rangle \langle Z_{t,p,2}^-, Z_{t,p,2}^- \rangle^{-1} \right],$$

with $\hat{\beta}_1 := \langle \tilde{e}_t, z_t^\Pi \rangle \langle z_t^\Pi, z_t^\Pi \rangle^{-1}$ and $z_t^\Pi := z_t - \langle z_t, Z_{t,p,2}^- \rangle \langle Z_{t,p,2}^-, Z_{t,p,2}^- \rangle^{-1} Z_{t,p,2}$. Thus, z_t^Π denotes the residuals of a regression of z_t onto $Z_{t,p,2}^-$ for $t = p+1, \ldots, T$. The above evaluation follows from the matrix inversion Lemma 4 using $A = \langle z_t, z_t \rangle$, $B = \langle z_t, Z_{t,p,2} \rangle$ and $C = \langle Z_{t,p,2}, Z_{t,p,2} \rangle$. The second term above, $\hat{\Theta}_p^{\tilde{e}} = \langle \tilde{e}_t, Z_{t,p,2}^- \rangle \langle Z_{t,p,2}^-, Z_{t,p,2}^- \rangle^{-1}$, contains only stationary quantities. In particular $Z_{t,p,2}$ contains $\tilde{e}_{t-j}, j = 1, \ldots, p-1$ and a part of $\tilde{e}_{t-p}$ as blocks. Thus, the asymptotic behavior of this term is covered by Theorem 2, from which we obtain $\hat{\Theta}_p^{\tilde{e}} - \Theta_p^{\tilde{e}} = O((T^{-1} \log T)^{1/2})$. Therefore, in order to establish (i), it is sufficient to show that the other terms above are of at most this order (in probability).

Let us start with the term $\left[I_c, -\langle z_t, Z_{t,p,2}^- \rangle \langle Z_{t,p,2}^-, Z_{t,p,2}^- \rangle^{-1} \right]$. Note first that $z_t = x_t + (C^\dagger)' e_t$ and again that $Z_{t,p,2}^-$ contains only stationary variables. Therefore equation (12) of Lemma 3 shows that $\langle x_t, Z_{t,p,2}^- \rangle$ is $O_p(1)$ uniformly in p. Furthermore, Theorem 6.6.11 (p. 267) of HD and Assumption 4 imply that $\| \langle Z_{t,p,2}^-, Z_{t,p,2}^- \rangle^{-1} \|_\infty < M$ a.s. for T large enough. Equation (10) implies that $\langle (C^\dagger)' e_t, Z_{t,p,2} \rangle = O_p(1)$ and hence $\langle z_t, Z_{t,p,2}^- \rangle \langle Z_{t,p,2}^-, Z_{t,p,2}^- \rangle^{-1}$ $= O_P(1)$ uniformly in p.

Consider $\hat{\beta}_1 = \langle \tilde{e}_t, z_t^\Pi \rangle \langle z_t^\Pi, z_t^\Pi \rangle^{-1}$ next. We start with the first term, i.e. with $\langle \tilde{e}_t, z_t^\Pi \rangle = \langle \tilde{e}_t, z_t \rangle - \langle \tilde{e}_t, Z_{t,p,2}^- \rangle \langle Z_{t,p,2}^-, Z_{t,p,2}^- \rangle^{-1} \langle Z_{t,p,2}^-, z_t \rangle$. We know $\langle \tilde{e}_t, Z_{t,p,2}^- \rangle \langle Z_{t,p,2}^-, Z_{t,p,2}^- \rangle^{-1} = O_p(1)$ already from above. Using again (12) of Lemma 3 it follows that both $\langle \tilde{e}_t, x_t \rangle$ and $\langle Z_{t,p,2}^-, x_t \rangle$ are $O_p(1)$ uniformly in p showing that $\langle \tilde{e}_t, z_t^\Pi \rangle = O_P(1)$ uniformly in $1 \leq p \leq H_T$ (where we here use $z_t = x_t + (C^\dagger)' e_t$).

Thus, the term $\langle z_t^\Pi, z_t^\Pi \rangle$ is left to be analyzed. In order to do so consider

$$T^{-1} \langle z_t^\Pi, z_t^\Pi \rangle = T^{-1} \langle z_t, z_t \rangle - T^{-1} \langle z_t, Z_{t,p,2}^- \rangle \langle Z_{t,p,2}^-, Z_{t,p,2} \rangle^{-1} \langle Z_{t,p,2}, z_t \rangle.$$

The first term above converges in distribution to a random variable W with positive definite covariance matrix, compare Lemma 3(i). With respect to the second term uniform boundedness of $\langle z_t, \tilde{e}_{t-j} \rangle$ together with the established properties of $\langle Z_{t,p,2}^-, Z_{t,p,2}^- \rangle$ immediately implies that it is of order $O_p(pT^{-1})$, which is due to our restriction that $1 \leq p \leq H_T$

in fact $o_p(1)$. Therefore we obtain

$$\begin{aligned}
&\mathbb{P}\left\{\|\langle \tilde{e}_t, z_t^\Pi\rangle (T^{-1}\langle z_t^\Pi, z_t^\Pi\rangle)^{-1}\| > M\right\} \leq \mathbb{P}\left\{\|\langle \tilde{e}_t, z_t^\Pi\rangle\| \|(T^{-1}\langle z_t^\Pi, z_t^\Pi\rangle)^{-1}\| > M\right\} \\
&\leq \mathbb{P}\left\{\|\langle \tilde{e}_t, z_t^\Pi\rangle\| > \sqrt{M}\right\} + \mathbb{P}\left\{\|(T^{-1}\langle z_t^\Pi, z_t^\Pi\rangle)^{-1}\| > \sqrt{M}\right\} \\
&\leq \eta/2 + \mathbb{P}\left\{\lambda_{min}(T^{-1}\langle z_t^\Pi, z_t^\Pi\rangle) < 1/\sqrt{M}\right\} \leq \eta.
\end{aligned}$$

In the above expression the first probability can be made arbitrarily small by choosing M large enough, since $\langle \tilde{e}_t, z_t^\Pi\rangle = O_p(1)$ and the second probability can be made arbitrarily small since $T^{-1}\langle z_t^\Pi, z_t^\Pi\rangle = T^{-1}\langle z_t, z_t\rangle + o_P(1) \stackrel{d}{\to} W$, where the random variable W has non-singular covariance matrix. Here $\lambda_{\min}(X)$ denotes the smallest eigenvalue of the matrix X. Thus, we have established that $\hat{\beta}_1 = O_p(T^{-1})$. This concludes the proof of (i).

Proof of (ii): We now derive the bounds to $IC_p^y(p; C_T)$, which requires to assess the approximation error in $\hat{\Sigma}_p^y$. The strategy is to show that the difference between $\hat{\Sigma}_p^y$ and $\tilde{\Sigma}_p^{\tilde{e}}$ is small enough. Here again $\tilde{\Sigma}_p^{\tilde{e}}$ denotes the error covariance matrix from the autoregression where the components $(C^\perp)' e_{t-p}$ are added (see the expression for $Z_{t,p}^-$ provided in Lemma 1). Then (ii) follows since for $\tilde{\Sigma}_p^{\tilde{e}}$ the result presented in Theorem 2(v) applies and thus the bounds to the information criterion are established also for MFI(1) processes and the case $q = 1$ if $\hat{\Sigma}_p^y - \tilde{\Sigma}_p^{\tilde{e}} = O_P(T^{-1})$ uniformly in p can be established. To this end consider

$$\hat{\Sigma}_p^y = \langle \tilde{e}_t - \hat{\beta}_1 z_t - \hat{\beta}_{2,p} Z_{t,p,2}^-, \tilde{e}_t - \hat{\beta}_1 z_t - \hat{\beta}_{2,p} Z_{t,p,2}^-\rangle \quad \text{and} \quad \tilde{\Sigma}_p^{\tilde{e}} = \langle \tilde{e}_t - \hat{\Theta}_p^{\tilde{e}} Z_{t,p,2}^-, \tilde{e}_t - \hat{\Theta}_p^{\tilde{e}} Z_{t,p,2}^-\rangle$$

where

$$\tilde{e}_t - \hat{\Theta}_p^{\tilde{e}} Z_{t,p,2}^- - \tilde{e}_t + \hat{\beta}_1 z_t + \hat{\beta}_2 Z_{t,p,2}^- = \hat{\beta}_1 z_t + (\hat{\beta}_2 - \hat{\Theta}_p^{\tilde{e}}) Z_{t,p,2}^- = \hat{\beta}_1 z_t^\Pi$$

follows from (14). Hence

$$\hat{\Sigma}_p^y - \tilde{\Sigma}_p^{\tilde{e}} = \langle \tilde{e}_t - \hat{\Theta}_p^{\tilde{e}} Z_{t,p,2}^-, \hat{\beta}_1 z_t^\Pi\rangle + \langle \hat{\beta}_1 z_t^\Pi, \tilde{e}_t - \hat{\Theta}_p^{\tilde{e}} Z_{t,p,2}^-\rangle + \langle \hat{\beta}_1 z_t^\Pi, \hat{\beta}_1 z_t^\Pi\rangle.$$

Recall that $\hat{\beta}_1 = O_P(T^{-1}), \langle z_t^\Pi, z_t^\Pi\rangle = O_P(T), \langle z_t^\Pi, \tilde{e}_t\rangle = O_P(1)$ and $\langle z_t^\Pi, Z_{t,p,2}^-\rangle = 0$ have been shown above. Using these results implies that $\hat{\Sigma}_p^y - \tilde{\Sigma}_p^{\tilde{e}} = O_P(T^{-1})$ uniformly in p, which shows the results.

Proof of (iii): The proofs above are all based on error bounds derived in Lemma 3 and the results summarized for stationary processes in Theorem 2. In both the lemma and the theorem the respective bounds are also derived for the case including the mean and harmonic components. This implies that the results also hold under Assumption 3(ii).

B.4 Proof of Theorem 4

Proof of (i): Consider Lemma 1(ii): There $\tilde{y}_t := [y_t', y_{t+1}', y_{t+2}', \dots, y_{t+q-1}']'$ is defined, where $(\tilde{y}_{tq+i})_{t\in\mathbb{Z}}$ fulfills the assumptions of Theorem 3. Let $p = \tilde{p}q$ for some integer $\tilde{p}$ and consider the sub-sampled processes $(\tilde{y}_{tq+i})_{t\in\mathbb{Z}}, i = 1, \dots, q$. For each value of i the corresponding autoregressive estimators are given as

$$\hat{\Theta}_p^{y,(i)} := [I_s, 0^{s\times s(q-1)}]\langle \tilde{y}_{tq+i}, Z_{tq+i,p}^-\rangle_{\tilde{p}}^{\tilde{T}-1}(\langle Z_{tq+i,p}^-, Z_{tq+i,p}^-\rangle_{\tilde{p}}^{\tilde{T}-1})^{-1}\mathcal{T}_{\tilde{p}}, i = 1, \dots, q.$$

Note that here summation is over $t = \tilde{p}, \dots, \tilde{T}-1$, where $\tilde{T} := \lfloor T/q \rfloor$ and $\lfloor x \rfloor$ denotes the integer part of x. In the arguments below it is always assumed for notational simplicity that $T = \tilde{T}q$ in order to simplify the notation. The expressions for other values of T differ from these only in the addition of finitely many values in the summation. This does not change any of the error bounds provided below. Note here that the effective sample size for estimating $\hat{\Theta}_p^{y,(i)}$ is reduced to $\tilde{T} = \lfloor T/q \rfloor$ due to the sub-sampling.

For each of these estimators the error bound $(\hat{\Theta}_p^{y,(i)} - \Theta_p^y)\tilde{\mathcal{T}}_{\tilde{p}}^{-1} = O_P((\tilde{T}^{-1}\log(\tilde{T}))^{1/2}) = O_P((\log T/T)^{1/2})$ follows according to the proof of Theorem 3(i). It is straightforward to see that

$$\hat{\Theta}_p^y = \langle y_t, Y_{t,p}^-\rangle_{p+1}^T(\langle Y_{t,p}^-, Y_{t,p}^-\rangle_{p+1}^T)^{-1} = \sum_{i=1}^q \hat{\Theta}_p^{y,(i)}\tilde{\mathcal{T}}_{\tilde{p}}^{-1}\langle Z_{tq+i,p}^-, Z_{tq+i,p}^-\rangle_{\tilde{p}}^{\tilde{T}-1}(\langle Z_{t,p}^-, Z_{t,p}^-\rangle_{p+1}^T)^{-1}\tilde{\mathcal{T}}_{\tilde{p}}$$

Therefore under the assumption of nonsingularity of $\langle Z_{t,p}^-, Z_{t,p}^-\rangle_{p+1}^T$ from $\sum_{i=1}^q \langle Z_{tq+i,p}^-, Z_{tq+i,p}^-\rangle_{\tilde{p}}^{\tilde{T}-1} = \langle Z_{t,p}^-, Z_{t,p}^-\rangle_{p+1}^T$ one obtains

$$\hat{\Theta}_p^y - \Theta_p^y = \sum_{i=1}^q (\hat{\Theta}_p^{y,(i)} - \Theta_p^y)\tilde{\mathcal{T}}_{\tilde{p}}^{-1}\langle Z_{tq+i,p}^-, Z_{tq+i,p}^-\rangle_{\tilde{p}}^{\tilde{T}-1}(\langle Z_{t,p}^-, Z_{t,p}^-\rangle_{p+1}^T)^{-1}\tilde{\mathcal{T}}_{\tilde{p}}.$$

In the proof of Theorem 3 it has actually been shown that the first c components of $\hat{\Theta}_p^{y,(i)} - \Theta_p^y$ are of order $O_P(T^{-1})$. Thus, letting $D_T = \text{diag}(T^{-1/2}I_c, I)$ one obtains:

$$D_T\langle Z_{tq+i,p}^-, Z_{tq+i,p}^-\rangle_{\tilde{p}}^{\tilde{T}-1}D_T(D_T\langle Z_{t,p}^-, Z_{t,p}^-\rangle_{p+1}^T D_T)^{-1} \xrightarrow{d} \begin{bmatrix} Z_{11}^{(i)} & 0 \\ 0 & \frac{1}{q}I \end{bmatrix}$$

for $i = 1, \dots, q$ where $\langle z_{tq+i}, z_{tq+i}\rangle_{\tilde{p}}^{\tilde{T}-1}(\langle z_t, z_t\rangle_{p+1}^T)^{-1} \xrightarrow{d} Z_{11}^{(i)}$. In the above matrix the off diagonal elements are of order $O_P(T^{-1/2})$ uniformly in $1 \le p \le H_T$. All evaluations use Lemma 3(i) and are straightforward. Since there are only finitely many terms it follows that $\hat{\Theta}_p^y - \Theta_p^y = O_P(\sqrt{(\log T)/T})$ for $p = \tilde{p}q$ as claimed in the theorem.

Proof of (ii): Again the sub-sampling argument is used to define the processes $(\tilde{e}_{tq+i})_{t\in\mathbb{Z}}$ according to $\tilde{e}_t := \tilde{c}_\bullet^{(q)}(L^q)\tilde{\varepsilon}_t$ where $\tilde{\varepsilon}_t := \mathcal{E}_f[\varepsilon_t', \ldots, \varepsilon_{t+q-1}']'$. Since $\tilde{c}_\bullet^{(q)}(z^q)$ is invertible due to Assumption 4 it follows that there exists a transfer function $\tilde{\Phi}^{\tilde{e}}(L^q)$ such that $\tilde{\Phi}^{\tilde{e}}(L^q)\tilde{e}_t = \tilde{\varepsilon}_t$. The first block equation (note that $\mathcal{E}_f$ is block lower triangular) here states that ε_t can be obtained by filtering $\tilde{e}_t$ which implies that $\varepsilon_t = \tilde{I}_s\tilde{\Phi}^{\tilde{e}}(L^q)\tilde{e}_t$. Therefore for $p = \tilde{p}q, \tilde{p} \in \mathbb{N} \cup \{0\}$ consider

$$\hat{\Sigma}_p^y = \langle y_t - \hat{\Theta}_p^y Y_{t,p}^-, y_t - \hat{\Theta}_p^y Y_{t,p}^- \rangle_{p+1}^T = \langle y_t - \hat{\Theta}_p^z Z_{t,p}^-, y_t - \hat{\Theta}_p^z Z_{t,p}^- \rangle_{p+1}^T.$$

As in the the proof of Theorem 3 again using a sub-sampling argument it can be shown that

$$\hat{\Sigma}_p^y = \langle \tilde{I}_s[\tilde{e}_t - \hat{\Theta}_{\tilde{p}}^{\tilde{e}} Z_{t,p,2}^-], \tilde{I}_s[\tilde{e}_t - \hat{\Theta}_{\tilde{p}}^{\tilde{e}} Z_{t,p,2}^-]\rangle_{p+1}^T + O_P(T^{-1})$$

where $Z_{t,p}^- = [z_t', (Z_{t,p,2}^-)']'$. Here $z_t \in \mathbb{R}^c$ denotes again the nonstationary components of $Y_{t,p}^-$. Note that the blocks of $Z_{t,p,2}^-$ are the lags $\tilde{e}_{t-jq}, j = 1, \ldots, \tilde{p}-1$ and a sub-vector of $\tilde{e}_{t-\tilde{p}q}$. In- or excluding this sub-vector we obtain lower and upper bounds respectively for $\hat{\Sigma}_p^y$ (see the proof of Theorem 2(v) for details). Therefore it is sufficient to derive the bounds only for $p = \tilde{p}q, \tilde{p} \in \mathbb{N} \cup \{0\}$. In the sequel we discuss lag length selection for the process $(\tilde{e}_t)_{t\in\mathbb{Z}}$ based on autoregressive approximations of lag length $h = \tilde{h}q$.

Analogously to the proof of Theorem 3 the proof is based on mimicking the proof of Theorems 7.4.6 and 7.4.7 of HD. There are two differences to the theory presented there: First, the order selection is not performed on the whole processes $(\tilde{e}_{tq+i})_{t\in\mathbb{Z}}$ but only on a sub-vector obtained by pre-multiplying with $\tilde{I}_s$. Second, the sub-sampled processes $(\tilde{e}_{tq+i})_{t\in\mathbb{Z}}$ use q as the time increment whereas in the selection 1 is used as time increment. Therefore the proofs of Theorem 7.4.6 and 7.4.7 of HD need to be reconsidered for the present setting.

As in the proof of Theorem 2(iii) the result follows from verifying (7.4.31) of HD. We obtain from summing the results for the sub-sampled processes (which follow directly from the proof of Theorem 2) over $i = 1, \ldots, q$ that for $k = 1, \ldots, h$

$$\sum_{j=1}^{h} \tilde{I}_s\left(\hat{\Phi}_h^{(q)}(j) - \Phi^{(q)}(j)\right)\tilde{g}(j,k) = -\hat{u}_k + \sum_{j=h+1}^{\infty} \tilde{I}_s\Phi^{(q)}(j)\tilde{g}(j,k) + o(T^{-1/2}),$$

where $\hat{u}_k = T^{-1}\sum_{t=qk+1}^{T}\varepsilon_t\tilde{e}'_{t-kq}$ and $\hat{\Phi}_h^{(q)}(j)$ denotes the least squares estimates in the regression for fixed h. Let further $\Phi^{(q)}(j)$ denote the true coefficients. This corresponds to equation (7.4.31) on p. 340 of HD with the $o(\log T/T)^{1/2})$ replaced by $o(T^{-1/2})$, which is discussed below the equation on p. 340 of HD. The arguments leading to the final line on p. 340 of HD then are based on population moments and the error bounds on the estimation of the covariance sequence (both of which hold in our setting as is straightforward to verify). The autoregressive approximation of $\tilde{e}_t$ underlying the estimation shows that $(\phi_h^{(2)})'\{\Gamma_{22}-\Gamma_{21}\Gamma_{11}^{-1}\Gamma_{12}\}\phi_h^{(2)} = \tilde{I}_s\Sigma_{\tilde{h}}^{\tilde{e}}\tilde{I}_s'-\Sigma$. Therefore in order to establish (7.4.32) on p. 341 of HD it is sufficient to show that $\tilde{G}_{22}-\tilde{G}_{21}\tilde{G}_{11}^{-1}\tilde{G}_{12}$ can be replaced by its expectation introducing an error of magnitude $o(h/T)$. For $\tilde{G}_{22}$ this again follows by sub-sampling and decomposing the sum over all t involved in the formation of $\tilde{G}_{22}$ into q sums over $tq+i$ where for each of these q sums the arguments below (7.4.32) can be used to obtain the required result. Similar arguments show the claim for the remaining terms.
The next step in the proof of Theorem 7.4.6 on p. 331 of HD is to show that

$$\hat{u}'\Gamma_{11}^{-1}\hat{u} = T^{-2}\sum_{j=1}^{h}\left(\sum_{t=1}^{T}\varepsilon_t\varepsilon_{t-j}\right)^2(1+o(1))$$

(here the scalar case is shown), which essentially involves replacing (in the notation of HD) $\Gamma_{11}^{-1/2}y(t,h)$ with $\varepsilon(t,h)$. In our setup this amounts to replacing $\Gamma_{11}^{-1/2}[\tilde{e}'_{t-q},\tilde{e}'_{t-2q},\ldots,\tilde{e}'_{t-h}]'$ with $[\varepsilon'_{t-1},\varepsilon'_{t-2},\ldots,\varepsilon'_{t-h}]'$. Showing that this replacement is valid can be shown using the same arguments as in HD, since the proof only involves error bounds on the estimated covariance sequences and the convergence of the coefficient matrices in $\tilde{\varepsilon}_t = \tilde{\Phi}(L^q)\tilde{e}_t$ which follow from the assumptions on $\tilde{c}_\bullet^{(q)}(z^q)$. The rest of the proof of Theorem 7.4.6 of HD uses only properties of ε_t. Then Theorem 7.4.6 of HD shows the required approximation for $h=\tilde{h}q$ for any integer $\tilde{h}$.

Now given any value of $p\in\mathbb{N}$ we use as in the proof of Theorem 2(v) with $\tilde{p} := \lfloor p/q\rfloor$

$$\langle y_t-\hat{\Theta}^y_{(\tilde{p}+1)q}Y^-_{t,(\tilde{p}+1)q}, y_t-\hat{\Theta}^y_{(\tilde{p}+1)q}Y^-_{t,(\tilde{p}+1)q}\rangle_{(\tilde{p}+1)q+1}^T \le \hat{\Sigma}_p^y \le \langle y_t-\hat{\Theta}^y_{\tilde{p}q}Y^-_{t,\tilde{p}q}, y_t-\hat{\Theta}^y_{\tilde{p}q}Y^-_{t,\tilde{p}q}\rangle_{\tilde{p}q+1}^T.$$

Then using the result for $\tilde{p}q$ and $(\tilde{p}+1)q$ shows the claim.

Proof of (iii): The changes necessary to prove (iii) are obvious and hence omitted.

B.5 Proof of Corollary 1

Proof of (i): This result follows from $\Sigma_p^{\tilde{e}} > \Sigma^{\tilde{e}}$, $\Sigma_p^{\tilde{e}} \to \Sigma^{\tilde{e}}$ and the fact that the penalty term tends to zero by assumption.

Proof of (ii): The proof is based on the arguments of HD, p. 333–334: Let $\tilde{p}(C_T) := \lfloor \hat{p}(C_T)/q \rfloor$. Then a mean value expansion is used to derive

$$\left(\frac{\tilde{p}(C_T)}{l_T(C_T)} - 1\right)^2 = 2\frac{\tilde{L}_T(\tilde{p}(C_T)) - \tilde{L}_T(l_T(C_T))}{l_T(C_T)^2\tilde{\theta}''(\bar{l}_T)} = 2\left(\frac{\tilde{L}_T(\tilde{p}(C_T))}{\tilde{L}_T(l_T(C_T))} - 1\right)\frac{\tilde{L}_T(l_T(C_T))}{\tilde{\theta}(l_T(C_T))}\frac{\tilde{\theta}(l_T(C_T))}{l_T(C_T)^2\tilde{\theta}''(\bar{l}_T)},$$

where $\bar{l}_T$ is an intermediate value. Since the latter two terms are bounded as in HD it is sufficient to show that $\tilde{L}_T(\tilde{p}(C_T))/\tilde{L}_T(l_T(C_T)) \to 1$. However, the following inequalities hold uniformly in p:

$$\begin{aligned}\tilde{L}_T(l_T(C_T); C_T) &\leq \tilde{L}_T(\tilde{p}(C_T)) \leq IC^y(\hat{p}(C_T); C_T)(1 + o_P(1)) \\ &\leq IC^y(q(l_T(C_T) - 1); C_T)(1 + o_P(1)) \leq \tilde{L}_T(l_T(C_T); C_T)(1 + o_P(1)).\end{aligned}$$

Here the first inequality follows from optimality of $l_T(C_T)$ with respect to $\tilde{L}_T$, the second from the lower bound of Theorem 3 (ii) (or Theorem 4(ii) resp.), the third from optimality of $\hat{p}(C_T)$ with respect to $IC^y(p, C_T)$ and the last again from Theorem 3 (ii) (or Theorem 4(ii) resp.). Here the uniformity of the $o_P(1)$ term in p is essential.

Proof of (iii): This result is an immediate consequence from the discussion in the lower half of p. 334 of HD.

Proof of (iv): Since all results in this paper are robust with respect to correcting for the mean and harmonic components prior to LS estimation it is evident that (iv) holds.