Autoregressive Approximations of Multiple Frequency I(1) Processes

Dietmar Bauer
arsenal research
Giefinggasse 2
A-1210 Wien, Austria

Martin Wagner
Institute for Advanced Studies
Stumpergasse 56
A-1060 Wien, Austria
Proposed running head:

**Autoregressive Approximations of MFI(1) Processes**

Corresponding author:

Martin Wagner

Institute for Advanced Studies
Stumpergasse 56
A-1060 Wien, Austria
Tel: ++43+1 59991 150
Fax: ++43+1 59991 163
Email: Martin.Wagner@ihs.ac.at
Abstract

We investigate autoregressive approximations of multiple frequency I(1), MFI(1), processes. 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 MFI(1) VARMA processes. The main results link the approximation properties of autoregressions for the nonstationary MFI(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, orders of convergence of 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. In particular for MFI(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
1 Introduction

This paper considers unit root processes that admit an infinite order autoregressive representation where the autoregressive coefficients satisfy mild summability constraints. More precisely the class of multiple frequency I(1) vector processes is analyzed. Following Bauer and Wagner (2005) a unit root 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 allow for nonrational transfer functions whose sequences of power series coefficients fulfill certain summability restrictions. On the other hand long memory processes (e.g. fractionally integrated processes) are not contained in the discussion.

Finite order vector autoregressive models 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, 1983), Chan and Wei (1988), Johansen (1995) or Johansen and Schaumburg (1999). Also model selection issues are well understood in this context, see e.g. Pötscher (1989).

For stationary processes finite order vector autoregressions have been used as approximate models for more general processes by letting the order tend to infinity as a function of the sample size and certain characteristics of the underlying process. 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, orders of convergence 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 order 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,
1993) and Saikkonen and Lütkepoh (1996). Here the first two papers derive the asymptotic properties of the estimators of the cointegrating space and the third one derives the asymptotic distribution of 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 the Johansen testing procedure for the cointegrating rank remains valid with unchanged asymptotic distributions if the autoregressive order diverges to infinity. The asymptotic distribution of the autoregressive coefficients, however, depends on the properties of the sequence of selected lag lengths.

For the seasonal or more general the multiple frequency integration case similar results on the properties of infinite order autoregressive approximations are not yet available.

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 where 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 Lütkepoh 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 $C_T = 2 \text{(AIC)}$ or $C_T = \log T \text{(BIC)}$ is used. Information criteria take decisions based on comparing the decrease in the log determinant of the estimated noise variance to a term penalizing the increase in the number of parameters measuring model complexity. Thus in order to show that the information criteria computed for the original and the differenced data lead to the same decision it is necessary to show that the differences in the estimated innovation variance are dominated - in a specific sense made clear below - by the penalty term. Now if the
difference is of order $o_P(T^{-1/2})$ then it can only be followed that the asymptotic behavior of the two approaches is identical if the penalty term dominates this order. This is only the case if $C_T/T^{1/2} \to \infty$, which is not fulfilled by any of the suggested criteria. Also in Lemma 5.1 of Lütkepohl and Saikkonen (1999) 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 $C_T$.

An alternative to using information criteria for lag length selection consists in sequential general-to-specific testing procedures such as the one proposed in Ng and Perron (1995). Such procedures, however, have several problems. First, the computation of the exact conditional distributions of the individual tests performed in the sequential testing procedure is practically impossible. Second, by construction such procedures typically lead to chosen lag lengths that are close to the arbitrarily chosen upper bound and are hence relatively large, e.g. Ng and Perron (1995) propose the upper bound to be proportional to $T^{1/4}$. However, there is no theoretical argument guiding the choice. Also choosing overly large lag lengths inflates the estimation errors of the estimated autoregressive coefficients. After all this is the motivation for performing model selection.

This paper extends the available theory in two ways: First, we show that the estimation error in autoregressive approximations is 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, from the stationary case 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 the special case of VARMA processes it follows that the BIC order estimator increases proportionally to $\log T$ to infinity in probability. The results in this paper are e.g. key inputs in generalizing the Johansen and Schaumburg (1999) tests for seasonal cointegration from the finite order autoregressive to the infinite order autoregressive case.

The paper is organized as follows: Section 2 presents some basic definitions, assumptions
and the considered class of processes. 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 the paper. Two appendices follow the main text. In Appendix A several useful lemmas 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 \( F_T \in \mathbb{R}^{g_T \times b_T} \) if \( \lim_{T \to \infty} \max_{1 \leq i \leq a_T, 1 \leq j \leq b_T} |F_{ij,T}|/g_T = 0 \) a.s., where \( F_{ij,T} \) denotes the \((i,j)\)-th entry of \( F_T \). \( F_T = O(g_T) \) means \( \limsup_{T \to \infty} \max_{1 \leq i \leq a_T, 1 \leq j \leq b_T} |F_{ij,T}|/g_T < M < \infty \) a.s. for some constant \( M \). Analogously \( F_T = o_P(g_T) \) means that \( \max_{1 \leq i \leq a_T, 1 \leq j \leq b_T} |F_{ij,T}|/g_T \) 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 \( \mathbb{P}\{\max_{1 \leq i \leq a_T, 1 \leq j \leq b_T} |F_{ij,T}|/g_T > M(\varepsilon)\} \leq \varepsilon \). 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 write \( \langle a_t, b_t \rangle_{T-j}^{T} := \frac{1}{T} \sum_{t-i}^{T-j} a_t b'_t \), 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 \( b_t \). Furthermore we write for brevity \( \langle a_t, b_t \rangle := \langle a_t, b_t \rangle_{p+1}^{T} \), when used in the context of autoregressions of order \( p \). The integer part of \( x \) is denoted by \([x]\), \( z \) denotes a complex variable and \( I \) denotes the indicator function.

## 2 Definitions and Assumptions

In this paper we consider 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 \( \omega \) as:

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

(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 \( L y_t = y_{t-1} \). 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 to ensure real valuedness of the differenced 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 \( v_t = c(L)\varepsilon_t = \sum_{j=0}^{\infty} c_j \varepsilon_{t-j} \). Using this notation we define a unit root process as follows:
Definition 1  The s-dimensional real process \((y_t)_{t \in \mathbb{Z}}\) where \(\mathbb{E}\|y_t\| < \infty\) 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

\[
D(L)(y_t - \mathbb{E}y_t) = v_t, \quad t \in \mathbb{Z}
\]

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 \(0 < \mathbb{E}v_t v_t' < \infty\).

The s-dimensional real process \((y_t)_{t \in \mathbb{Z}}\) has empty unit root structure \(\Omega_0 := \{\}\) if \((y_t - \mathbb{E}y_t)_{t \in \mathbb{Z}}\) 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 (2005) 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 \(((\omega_1, 1), \ldots, (\omega_l, 1))\) with \((\mathbb{E}y_t)_{t \in \mathbb{Z}}\) such that \(D(L)\mathbb{E}y_t = 0\) is called multiple frequency I(1) process, or short MFI(1) process.

Note, as already indicated in the introduction, that the definition of MFI(1) processes places restrictions on the expectation \(\mathbb{E}y_t\). 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 MFI process \((y_t)_{t \in \mathbb{Z}}\) is a solution to the difference equation

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

where \(v_t = \sum_{j=0}^{\infty} c_j \varepsilon_{t-j}\) corresponds to the Wold decomposition of the stationary process \((v_t)_{t \in \mathbb{Z}}\) 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. Denote by \(H := \sum_{k=1}^{l}(1 + \mathbb{I}(0 < \omega_k < \pi))\) the degree of \(D(z)\). Using this notation the summability assumption \(\sum_{j=0}^{\infty} j^{3/2 + H}\|c_j\| < \infty\) holds.
Assumption 2 The stochastic process \((\varepsilon_t)_{t \in \mathbb{Z}}\) is a strictly stationary ergodic martingale difference sequence with respect to the \(\sigma\)-algebra \(\mathcal{F}_t = \sigma\{\varepsilon_t, \varepsilon_{t-1}, \varepsilon_{t-2}, \ldots\}\) fulfilling
\[
\mathbb{E}\{\varepsilon_t | \mathcal{F}_{t-1}\} = 0, \quad \mathbb{E}\{\varepsilon_t \varepsilon'_t | \mathcal{F}_{t-1}\} = \mathbb{E} \varepsilon_t \varepsilon'_t = \Sigma > 0, \quad \mathbb{E} \varepsilon_{t,j}^4 \log_+ (|\varepsilon_{t,j}|) < \infty, \quad j = 1, \ldots, s,
\]
where \(\varepsilon_{t,j}\) denotes the \(j\)-th coordinate of the vector \(\varepsilon_t\) 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 \(((0, h))\) for \(h \in \mathbb{N}\).

**Theorem 1** Let \((y_t)_{t \in \mathbb{Z}}\) be a process fulfilling Assumption 1. Then there exists a representation as
\[
y_t = \sum_{k=1}^l C_k x_{t,k} + \sum_{j=0}^\infty c_{j,\bullet} \varepsilon_{t-j} + y_{t,h} = \tilde{y}_t + y_{t,h},\]
where \(\varepsilon_{t,j}\) denotes the \(j\)-th coordinate of the vector \(\varepsilon_t\) and \(\log_+ (x) = \log(\max(x,1))\).

Further
- \(\tilde{c}_k\) denotes the rank of the matrices \(c(e^{i\omega_k}), c(e^{-i\omega_k}) \in \mathbb{C}^{s \times s}\),
- \(x_{t+1,k} = J_k x_{t,k} + K_k \varepsilon_t \in \mathbb{R}^{c_k}, t \in \mathbb{Z}\) where \(c_k := \tilde{c}_k\) if \(\omega_k \in \{0, \pi\}\) and \(c_k := 2\tilde{c}_k\) else.

and the matrices \(C_k \in \mathbb{R}^{s \times c_k}, K_k \in \mathbb{R}^{c_k \times s}, k = 1, \ldots, l\) are such that the state space systems \((J_k, K_k, C_k)\) are minimal (see p. 47 of HD for a definition);
- the transfer function \(c_\bullet(z) = \sum_{j=0}^\infty c_{j,\bullet} z^j\) fulfills \(\sum_{j=0}^\infty j^{3/2} ||c_{j,\bullet}|| < \infty, \det c_\bullet(z) \neq 0, |z| < 1\).
for the process \((y_{t,h})_{t \in \mathbb{Z}}\) it holds that \(D(L)(y_{t,h})_{t \in \mathbb{Z}} \equiv 0\).

**Proof:** The proof centers around the representation for \(c(z)\) given in Lemma 2 in Appendix A. 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 this is established \(D(L)y_{t,h} = D(L)(y_t - \tilde{y}_t) = 0\) then proves the theorem.

Therefore consider \(\tilde{y}_t = \sum_{k=1}^t C_k x_{t,k} + \sum_{j=0}^{\infty} c_j \cdot \varepsilon_{t-j}\). Note that for \(0 < \omega_k < \pi\)

\[
(1 - 2 \cos(\omega_k)L + L^2)x_{t,k} = J_k x_{t-1,k} + K_k \varepsilon_{t-1} - 2 \cos(\omega_k)(J_k x_{t-2,k} + K_k \varepsilon_{t-2}) + x_{t-2,k}
\]

\[
= (J_k^2 - 2 \cos(\omega_k)J_k + I_{c_k})x_{t-2,k} + K_k \varepsilon_{t-1} + (J_k - 2 \cos(\omega_k)I_{c_k})K_k \varepsilon_{t-2}
\]

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 \(t \in \mathbb{Z}\)

\[
D(L)x_{t,k} = D_{-k}(L)\Delta_{\omega_k}(L)x_{t,k} = D_{-k}(L)(K_k \varepsilon_{t-1} - J_k'K_k \varepsilon_{t-2} | \omega_k \notin \{0, \pi\})
\]

with \(D_{-k}(z) = D(z)/\Delta_{\omega_k}(z)\). 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 \(t \in \mathbb{Z}\)

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

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 \(D(L)\tilde{y}_t = v_t\). □

Theorem 1 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 (2005). In the I(1) case it 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 homogeneous equation \(D(L)y_{t,h} = 0\) are of the form \(y_{t,h} = \sum_{k=1}^t D_{k,c} \cos(\omega_k t) + D_{k,s} \sin(\omega_k t)\) 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 $\omega_k \notin \{0, \pi\}$. Consequently the homogeneous solutions to the equations $x_{t+1,k} = J_k x_{t,k} + K_k \varepsilon_t$ contain such processes $d_{t,k}$ as their components. Now if – for all $k = 1, \ldots, l$ – $C_k = [C_{k,c}, C_{k,s}]$ 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 $\sum_{k=1}^l D_{k,c} \cos(\omega_k t) + D_{k,s} \sin(\omega_k t)$ can be either be a summand in $y_{t,h}$ or be attributed to a particular solution for the equations defining $(x_{t,k})_{t \in \mathbb{Z}}$. Thus in this case 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 $(y_{t,h})_{t \in \mathbb{Z}} \equiv 0$. Consequently in this special case there is no need to model the deterministic components explicitly. This is formulated for later reference in the following assumption:

**Assumption 3** $(y_{t})_{t \in \mathbb{Z}}$ is generated according to Assumption 1 where 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 $(y_{t,h})_{t \in \mathbb{Z}} \equiv 0$.

If such a representation does not exist the autoregressive model has to account for deterministic terms by adding additional regressors corresponding to processes $(d_{t,k,i})_{t \in \mathbb{Z}}, k = 1, \ldots, l, i = 1, 2$. For the results of this paper it is not critical that only the necessary regressors are added. Hence we will only consider the cases where the deterministics are neglected or where the vector $d_t := [d_{t,1}, \ldots, d_{t,l}]'$ is added to the estimated autoregression equation. Here $d_{t,k} := \begin{bmatrix}
\cos(\omega_k t) \\
\sin(\omega_k t)
\end{bmatrix}$ for $0 < \omega_k < \pi$ and $d_{t,k} := \cos(\omega_k t)$ for $\omega_k \in \{0, \pi\}$.

**Remark 1** Note further that the restriction $D(L)(\mathbb{E}y_{t})_{t \in \mathbb{Z}} \equiv 0$ 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 $\omega$ could be included, if they are also accounted for in the estimation. For sake of brevity 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 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 (collected in Theorem 2) also apply to the LS estimator.

11
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 \(\hat{\Phi}_p^v(j)\) and their YW estimators \(\hat{\Phi}_{p}^v(j)\) are defined from the Yule-Walker equations given below: Define the sample covariances as

\[
    G^v(j) := \langle v_t - j, v_t \rangle_{j+1}^T\text{ for } 0 \leq j < T, G^v(j) := G^v(-j)'\text{ for } -T < j < 0 \text{ and } G^v(j) := 0 \text{ else.}
\]

We denote their population counterparts with \(\Gamma^v(j) := \mathbb{E}v_t - j v_t'\). Then \(\Phi_p^v(j)\) and \(\hat{\Phi}_p^v(j)\) are defined as the solutions to the respective YW equations (where \(\Phi_p^v(0) = \hat{\Phi}_p^v(0) = I_s\)):

\[
    \sum_{j=0}^{p} \Phi_p^v(j) \Gamma^v(j - i) = 0, \quad i = 1, \ldots, p,
\]

\[
    \sum_{j=0}^{p} \hat{\Phi}_p^v(j) G^v(j - i) = 0, \quad i = 1, \ldots, p.
\]

The infinite order Yule-Walker equations and the corresponding autoregressive coefficient matrices are defined from:

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

where the existence of these solutions follows from the assumptions imposed in this paper, see below. 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 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 mentioned we focus on the LS estimator in this paper. Using the regressor vector \(V_{t,p} := [v'_t, \ldots, v'_{t-p}]'\) for \(t = p + 1, \ldots, T\), the LS estimator, \(\hat{\Theta}_p^v\), is defined by

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

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 \(\Sigma^v_0 := G^v(0)\) and suitable upper bound \(H_T\), see below):

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

12
and note the following identity for the covariance matrix of \((\varepsilon_t)_{t \in \mathbb{Z}}\):

\[
\Sigma = \mathbb{E} \varepsilon_t \varepsilon_t' = \sum_{j=0}^{\infty} \Phi^v(j) \Gamma^v(j),
\]

where the infinite sum exists due to Assumption 1. Thus, \(\hat{\Sigma}_p^v\) denotes the estimated variance of the residuals of the autoregression of order \(p\). 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 using 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:

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

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.

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

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 \((v_t)_{t \in \mathbb{Z}}\) 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. Note that in the important special case of MFI(1) VARMA processes these summability assumptions are clearly fulfilled. Theorem 2 below, whose proof is contained in Appendix B, presents the results required for our paper for the LS estimator. Note again that the results in HD are for the YW estimator.

**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

\[
\max_{1 \leq j \leq p} \| \Phi^v_p(j) - \hat{\Phi}^v_p(j) \| = O((\log T/T)^{1/2}).
\]
(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 some $a < \infty$.

(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 $C_T/\log T \to \infty$, $C_T/T \to 0$ it holds that
  \[ IC^v(p; C_T) = \log \det \hat{\Sigma} + \left\{ \frac{p^2 s^2}{T}(C_T - 1) + tr[\Sigma^{-1}(\Sigma^v - \Sigma)] \right\} \{ 1 + o(1) \}, \]
  with $\hat{\Sigma} := T^{-1} \sum_{t=1}^{T} \varepsilon_t \varepsilon_t'$ and the approximation error is $o(1)$ uniformly in $0 \leq p \leq H_T$.

- For $C_T \geq c > 1$, $C_T/T \to 0$ 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 $p_0 > 1$ 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
  \[ \lim_{T \to \infty} \frac{2\hat{p}_{BIC}\log p_0}{\log T} = 1 \text{ a.s.} \]

(v) Assume as in (iv) that $(v_t)_{t \in \mathbb{Z}}$ is not generated by a finite order autoregression and let $\hat{P}_s \in \mathbb{R}^{r \times s}$, $r \leq s$ 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 $\hat{P}_s v_{t-p}$ results (i) to (iv) continue to hold, when the approximation to $IC^v(p; C_T)$ presented in (iii) is replaced by:

\[
\tilde{IC}^v(p; C_T) \leq \log \det \hat{\Sigma} + \left\{ \frac{p^2 s^2}{T}(C_T - 1) + tr[\Sigma^{-1}(\Sigma^v - \Sigma)] \right\} \{ 1 + o(1) \},
\]
\[
\tilde{IC}^v(p; C_T) \geq \log \det \hat{\Sigma} + \left\{ \frac{p^2 s^2}{T}(C_T - 1) + tr[\Sigma^{-1}(\Sigma^v - \Sigma)] \right\} \{ 1 + o(1) \}\]

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 $\tilde{IC}^v(p; C_T)$ denotes the information criterion from the regression of order $p-1$ augmented by $\hat{P}_s v_{t-p}$.
(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

\[ \hat{v}_t := v_t - \langle v_t, d_t \rangle_T (\langle d_t, d_t \rangle_T)^{-1} d_t, \]

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 \( d_t := [d^T_{t,1}, \ldots, d^T_{t,l}]^T \).

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 \( IC^w(p; C_T) \) for all values \( 0 \leq p \leq H_T \) simultaneously. Based upon the uniform convergence results for the autoregressive coefficients the asymptotic properties of information criteria are derived, which are seen to depend upon characteristics of the true unknown system (in particular upon \( \Sigma^w_p \), which in the VARMA case is closely related to \( \rho_0 \), see HD, p. 334). The result establishes a connection between the information criterion and the deterministic function \( \tilde{L}_T(p; C_T) := ps^2 \frac{C_T}{T} + \text{tr} \left[ \Sigma^{-1}(\Sigma^w_p - \Sigma) \right] \). 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). Here ‘very far’ refers to a large ratio of the value of the deterministic function compared to its minimal value. Under an additional assumption on the shape of the deterministic function (compare Corollary 1(iii)), 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 if properly accounted for.
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 without deterministic components (i.e. without mean and harmonic components), however, the theorems contain the results in the general case without Assumption 3. Paralleling the notation of the previous section define

\[ u_t(p) := y_t + \Phi^y_p(1)y_{t-1} + \ldots + \Phi^y_p(p)y_{t-p}. \]

The LS estimator of \(\Phi^y_p(j), j = 1, \ldots, p\) is given by

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

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

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

where again \(C_T\) is a suitably chosen penalty function. The 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

\[ \mathcal{O}_q := \begin{bmatrix} C_1 & \ldots & C_l \\ C_1J_1 & \ldots & C_lJ_l \\ \vdots & \vdots & \vdots \\ C_1J_1^{q-1} & \ldots & C_lJ_l^{q-1} \end{bmatrix} \]

has full column rank. Due to minimality of the systems \((J_k, K_k, C_k)\), see Theorem 1, for \(k = 1, \ldots, l\) it holds that \(q < \infty\) (cf. Theorem 2.3.3 on p. 48 of HD).

**Lemma 1** Let \((y_t)_{t \in \mathbb{Z}}\) be generated according to Assumptions 1 and 3. Denote with \(C := [C_1, \ldots, C_l] \in \mathbb{R}^{s \times c}\), \(K := [K'_1, \ldots, K'_l] \in \mathbb{R}^{c \times s}\), \(J := \text{diag}(J_1, \ldots, J_l) \in \mathbb{R}^{c \times c}\) and \(x_t := [x_{t,1}, \ldots, x_{t,l}]' \in \mathbb{R}^c\) 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 \(e_t := c_*(L)e_t\). Hence \(y_t = Cx_t + e_t\).
If \( q = 1 \) define \( \tilde{C}^i := [C^i, C^\perp] \), with \( C^i := C(C^\perp C)^{-1} \) and \( C^\perp \in \mathbb{R}^{s \times (s-c)} \) is such that
\[
(C^\perp)^t C^\perp = I_{s-c}, \ C^\perp C^\perp = 0.
\]
Define
\[
T_p := Q_p (I_p \otimes \tilde{C}), \quad \text{where} \quad Q_p := \begin{bmatrix}
I_c & 0 & 0 \\
0 & I_{s-c} & 0 \\
I_c & 0 & -J \\
0 & 0 & I_{s-c} \\
0 & 0 & -J \\
\ddots & \ddots & \ddots \\
I_{s-c} & 0 & -J \\
0 & 0 & 0 \\
\end{bmatrix}
\]
and
\[
Z_{t,p}^- := T_p Y_{t,p}^-. \quad \text{It holds that}
\]
\[
y_t - CJ(C^\perp)^t y_{t-1} = [C, C^\perp] \begin{bmatrix}
x_{t-1} + (C^\perp)^t e_{t-1} \\
K e_{t-2} + (C^\perp)^t e_{t-1} - J(C^\perp)^t e_{t-2} \\
K e_{t-3} + (C^\perp)^t e_{t-2} - J(C^\perp)^t e_{t-3} \\
\vdots \\
(C^\perp)^t e_{t-p}
\end{bmatrix}.
\]
Define the transfer function
\[
\tilde{c}_\bullet(z) := \sum_{j=0}^{\infty} \tilde{c}_j z^j = [C, C^\perp] \begin{bmatrix}
K z + (I - zJ)(C^\perp)^t c_\bullet(z) \\
(C^\perp)^t c_\bullet(z)
\end{bmatrix}
\]
and let \( \tilde{e}_t := \tilde{c}_\bullet(L) \tilde{e}_t \). The transfer function \( \tilde{c}_\bullet(z) \) has the following properties: \( \tilde{c}_\bullet(0) = I_s, \sum_{j=0}^{\infty} j^{3/2} \| \tilde{c}_j \| < \infty \) and hence \( \tilde{c}_\bullet(z) \) has no zeros on the 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) In case \( q > 1 \) define \( \tilde{y}_t := [y_{t+q-1}' , \ldots , y_{t+1}', y_t]' \in \mathbb{R}^{\tilde{s}} \), with \( \tilde{s} := sq \). Then the sub-sampled stacked processes \( \tilde{y}_{t+i} \) for \( i = 1, \ldots, q \) full Assumptions 1 and 3 and have by construction observability index equal to 1. Consequently for these processes part (i) of the lemma applies with \( \tilde{z}_{t,q}^{(q)} = \tilde{c}_\bullet(L^q) \tilde{e}_t \). This implies that there exists a transformation matrix \( \tilde{T}_p \in \mathbb{R}^{\tilde{s} \times \tilde{s}} \) such that in \( Z_{t,p}^- := \tilde{T}_p Y_{t,p}^- \) 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, however, it is obvious that it also applies without Assumption 3 in which case \( e_t := c_\bullet(L) \tilde{e}_t + y_{t,h} \). Note that in \( Z_{t,p}^- \) the first \( c \) components are unit root processes which
moreover 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 $q = 1$ (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** Under Assumption 1 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 $(y_t)_{t \in \mathbb{N}}$ is the solution to an infinite order autoregression. Letting $\Phi^y(z) := \tilde{c}_\bullet(z)^{-1}(I - CJ(C^\dagger)'z) = \sum_{j=0}^{\infty} \Phi^y(j)z^j$ we have $\sum_{j=0}^{\infty} j^{3/2}||\Phi^y(j)|| < \infty$. For $q > 1$ a 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 convenience is not normalized to $c(0) = I_2$. Note that $\det c(z) = -1.5(1 - z)^2$ and hence $z = 1$ is the only root. Furthermore, $c(1) = C_1K_1$ is non-zero and equal to $[1, 0]'[1, 0.5]$.

Now, using the representation of $c(z)$ 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

$$\tilde{c}_\bullet(z) = \begin{bmatrix} K_1z + (1 - z)(C^\dagger)'c_\bullet(z) \\ (C^\perp)'c_\bullet(z) \end{bmatrix} = \begin{bmatrix} 1 & 1.5 - z \\ 1 & 0.5z \end{bmatrix},$$

evaluated at $z = 1$ one obtains

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

18
from which one sees that \( \text{det} \hat{\epsilon}_\bullet(1) = 0 \). This example shows that indeed the additional Assumption 4 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. Exactly this type of problem on ‘thin sets’ occurs also in the general case, but we abstain from a more detailed discussion of the associated issues here and instead rely upon Assumption 4.

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 case \( q = 1 \) is dealt with in Theorem 3 and the case \( q > 1 \) is considered in Theorem 4.

Thus consider the case \( q = 1 \). Note that for any choice of the autoregressive lag length \( p \) it holds that

\[
\hat{\Theta}_p^y = \langle y_t, Y_{t,p}^- \rangle \langle Y_{t,p}^-, Y_{t,p}^- \rangle^{-1} = \langle y_t, Z_{t,p}^- \rangle \langle Z_{t,p}^-, Z_{t,p}^- \rangle^{-1} = \hat{\Theta}_p^z T_p,
\]

where this equation defines \( \hat{\Theta}_p^z \). 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. Note that \( z_t \) does not depend on \( p \). Since \( y_t = CJ(C^\dagger)_y y_{t-1} + \bar{C}^{-1} \tilde{e}_t, z_t = (C^\dagger)'y_{t-1} \) and \( CJ = \bar{C}^{-1} \begin{pmatrix} J \\ 0 \end{pmatrix} \) we obtain

\[
\hat{\Theta}_p^z = \langle CJ z_t, Z_{t,p}^- \rangle \langle Z_{t,p}^-, Z_{t,p}^- \rangle^{-1} + \langle \bar{C}^{-1} \tilde{e}_t, Z_{t,p}^- \rangle \langle Z_{t,p}^-, Z_{t,p}^- \rangle^{-1} \]
\[
= \bar{C}^{-1} \begin{pmatrix} J \\ 0 \end{pmatrix} \begin{pmatrix} 0 & \cdots & 0 \end{pmatrix} + \bar{C}^{-1} \langle \tilde{e}_t, Z_{t,p}^- \rangle \langle Z_{t,p}^-, Z_{t,p}^- \rangle^{-1}
\]

and thus it is sufficient to establish the asymptotic behavior of the second term on the right hand side of the above equation. Next using the block matrix inversion (see Lemma 4) one obtains

\[
\langle \tilde{e}_t, Z_{t,p}^- \rangle \langle Z_{t,p}^-, Z_{t,p}^- \rangle^{-1} = \hat{\Theta}_p^\tilde{e} + \langle \tilde{e}_t, Z_{t,p}^\Pi \rangle \langle z_t^\Pi, z_t^\Pi \rangle^{-1} \left[ I, -\langle z_t, Z_{t,p,2}^- \rangle \langle Z_{t,p,2}^-, Z_{t,p,2}^- \rangle^{-1} \right]
\]

where

\[
\hat{\Theta}_p^\tilde{e} := \langle \tilde{e}_t, Z_{t,p,2}^- \rangle \langle Z_{t,p,2}^-, Z_{t,p,2}^- \rangle^{-1}, \quad z_t^\Pi = z_t - \langle z_t, Z_{t,p,2}^- \rangle \langle Z_{t,p,2}^-, Z_{t,p,2}^- \rangle^{-1} Z_{t,p,2}^-.
\]

Since it can be shown (see the proof of the upcoming Theorem 3) that the second summand is of order \( O_P(T^{-1}) \) uniformly in \( 0 \leq p \leq H_T \), the asymptotic behavior of \( \hat{\Theta}_p^y \) is governed by the asymptotic behavior of \( \hat{\Theta}_p^\tilde{e} \), 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 for 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 both $\tilde{e}_t$ and $Z_{t,p,2}^-$ are jointly stationary we can define

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

and analogously $\Theta^\tilde{e}$ as the solution to the corresponding infinite order population YW equations. Finally, as in Section 3 define $\Sigma_{\tilde{e}}^p$ and $\Sigma_{\tilde{e}}^p$ as the population innovation variance of the process $(\tilde{e}_t)_{t \in \mathbb{Z}}$ and the minimal residual variance achievable with an autoregressive approximation of order $p$. 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 and 4 with $q = 1$ and let $0 \leq p \leq H_T$.

(i) Then it holds that

$$\max_{1 \leq p \leq H_T} \|\hat{\Theta}_y^p - \left(\begin{array}{c} C^{-1} \left(\begin{array}{c} J \\ 0 \end{array}\right), \Theta_{\tilde{e}}^p \end{array}\right) \right\|_1 = O_P((\log T/T)^{1/2}),$$

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

(ii) If $(y_t)_{t \in \mathbb{Z}}$ is not generated by a finite autoregression then for $C_T \geq c > 1, C_T/T \to 0$ the following approximations hold

$$IC_y^p(p; C_T) \leq \log \det \Sigma + \left\{ \frac{ps^2}{T}(C_T - 1) + tr[(\Sigma^\tilde{e})^{-1}(\Sigma_{p-1}^\tilde{e} - \Sigma^\tilde{e})] \right\} \{1 + o_P(1)\}$$

$$IC_y^p(p; C_T) \geq \log \det \Sigma + \left\{ \frac{ps^2}{T}(C_T - 1) + tr[(\Sigma^\tilde{e})^{-1}(\Sigma_{p}^\tilde{e} - \Sigma^\tilde{e})] \right\} \{1 + o_P(1)\}.$$  

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

(iii) All statements remain valid if Assumption 3 does not hold and the autoregressive approximations are performed on $\tilde{y}_t$, defined as $\tilde{y}_t := y_t - \langle y_t, d_t \rangle \left(\begin{array}{c} (dt, dt)^T \end{array}\right)^{-1} d_t$, with $d_t$ as defined in Theorem 2.

Here (i) is the analogue to Theorem 2(i), the only difference is that the result is stated in probability rather than a.s. This shows that the existence of (seasonal or multiple frequency) integration does not alter the estimation accuracy of the autoregressive coefficients substantially compared to the stationary case. Result (ii) is essentially the analogue of Theorem 2(iii), where however due to the fact that in the considered regression components of $\dot{e}_{t-p}$ are omitted (since only $(C^\perp)^'e_{t-p}$ is contained in the regressor vector $Z_{t,p,2}^-$, see the representation...
given in Lemma 1) lower and upper bounds similar to the bounds derived in Theorem 2(v) are developed. As in the stationary case it is this result that provides uniform bounds for the information criterion (in the range \(0 \leq p \leq H_T\)) given by the sum of a deterministic function and a noise term.

These results imply that the asymptotic behavior of the autoregressive approximation depends essentially on the properties of the stationary process \((\tilde{e}_t)_{t \in \mathbb{Z}}\): Except for the first block all blocks of \(\hat{\Theta}^z_p\) converge to blocks of the matrix \(\Theta^z_p\) 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 provide a strong relation to the information criterion corresponding to autoregressive approximations of \((\tilde{e}_t)_{t \in \mathbb{Z}}\), which is detailed in Corollary 1 below.

For \(q > 1\) the sub-sampling argument outlined in Lemma 1 shows that similar results can be obtained by resorting to \(q\) stacked and sub-sampled time series of dimension \(s = qs\):

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

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

\[
\max_{1 \leq p \leq H_T} \left\| \frac{\hat{\Theta}^y_{pq}}{T} - \tilde{I}_s \left( \tilde{C}^{-1} \begin{pmatrix} J & 0 \\ 0 & 0 \end{pmatrix}, \Theta^z_{pq} \right) \tilde{T}_p \right\|_1 = O_P((\log T/T)^{1/2}).
\]

(ii) If \((y_t)_{t \in \mathbb{Z}}\) is not generated by a finite order autoregression, then for \(C_T \geq c > 1, C_T/T \to 0\) the following approximations hold, denoting with \(\tilde{p} := \lfloor p/q \rfloor\) (noting that \(\Sigma^z_p\) denotes the residual variance corresponding to autoregressive approximations of \(\tilde{e}^{(q)}\))

\[
\begin{align*}
IC^y(p; C_T) &\leq \log \det \Sigma + \left\{ \frac{\tilde{p}qs^2}{T}(C_T - 1) + \text{tr} \left[ \Sigma^{-1} \tilde{I}_s (\Sigma^z_{p-1} - \Sigma^z_0) \tilde{T}_p \right] \right\} \{1 + o_p(1)\}, \\
IC^y(p; C_T) &\geq \log \det \Sigma + \left\{ \frac{\tilde{p}qs^2}{T}(C_T - 1) + \text{tr} \left[ \Sigma^{-1} \tilde{I}_s (\Sigma^z_{p} - \Sigma^z_0) \tilde{T}_p \right] \right\} \{1 + o_p(1)\}.
\end{align*}
\]

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

(iii) All statements remain valid without Assumption 3 if the autoregressive approximations are performed on \(\tilde{y}_t\), defined as \(\tilde{y}_t := y_t - \langle y_t, d_t \rangle_{1}^{T}((\langle d_t, d_t \rangle_{1}^{T})^{-1} - 1)d_t\), with \(d_t\) as defined in Theorem 2.

Compared to the results for \(q = 1\), 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 all \( p \in \mathbb{N} \). The bounds on the information criterion are also related to the closest integer multiple of \( q \). Nevertheless, as \( \hat{p} \) tends to infinity with increasing sample size availability of the approximation results only for multiples of \( q \) might be considered a minor issue.

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 of the information criterion (discussing here the case corresponding to Theorem 4), the deterministic function

\[
\tilde{L}_T(\hat{p}; C_T) := \frac{\hat{p}qs^2}{T}(C_T - 1) + \text{tr} \left[ \Sigma^{-1} \tilde{I}_s (\Sigma^\tilde{e}_p - \Sigma^\tilde{e}) \tilde{I}_s^\prime \right]
\]  

(10)

has a key role. Since 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 and only if the process \((\tilde{e}_t(q))_{t \in \mathbb{Z}}\) 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 \( \lim_{T \to \infty} \frac{2BIC \log \rho_0}{\log T} = 1 \) almost surely. Here we denote again with \( \hat{p}_{BIC} \) the smallest minimizing argument of the information criterion BIC and by \( \rho_0 \) the smallest modulus of the zeros of the determinant 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 proof of Corollary 1 is given in Appendix B.

**Corollary 1** Let \((y_t)_{t \in \mathbb{Z}}\) be generated according to Assumptions 1, 2, 3 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 \( C_T/T \to 0 \). Then the following results hold:

\[
i(i) \quad P\{\hat{p}(C_T) < M\} \to 0 \text{ for any constant } M < \infty.
\]
(ii) Assume that there exists a twice differentiable function \( \tilde{\theta}(p) \) with second derivative \( \tilde{\theta}''(p) \) such that \( \lim_{p \to \infty} \frac{\text{tr}[\Sigma^{-1}s(\Sigma p - \Sigma \hat{\Sigma})s]}{\hat{\theta}(p)} = 1 \) and \( \liminf_{p \to \infty} |p^2\hat{\theta}''(p)/\hat{\theta}(p)| > 0 \). 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.

(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 \( \rho_0 = \min\{|z| : z \in \mathbb{C}, \det \tilde{\epsilon}^{(q)}(z^q) = 0\}\), with \( \tilde{\epsilon}^{(1)}(z) = \tilde{\epsilon}(z) \).

(iv) All statements remain valid without Assumption 3 if the autoregressive approximations are performed on \( \hat{y}_t \), defined as \( \hat{y}_t = y_t - \langle y_t, d_t \rangle_T (\langle d_t, d_t \rangle_T)^{-1} d_t \), with \( d_t \) as defined in Theorem 2.

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). In particular the 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, \((\tilde{\epsilon}^{(q)}(L^q)x_t)_{t \in \mathbb{Z}}\) 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 behaves essentially 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 difference of the residual variance estimates $\hat{\Sigma}_p - \tilde{\Sigma}_p$ 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.

In this paper we do not analyze the asymptotic distributions of the estimated autoregressive coefficients based on estimated orders but only derive uniform error bounds (see e.g. Kuersteiner, 2005 in this respect for stationary processes). Of particular importance 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 autoregressive coefficients. This is left for future research, for which important prerequisites have been established in this paper.

Acknowledgements

Part of this work has been carried out whilst the first author was affiliated with the Vienna University of Technology and held a post-doc position at the Cowles Foundation, Yale University, financed by the Max Kade Foundation and the second author was affiliated with the University of Bern and was visiting the Economics Departments of Princeton University and of the European University Institute in Florence. Support from these institutions is gratefully acknowledged. Helpful comments from participants of the workshops ‘The Cointegrated VAR Approach’ in Copenhagen and ‘Persistence in Economic and Financial Time Series’
in Frankfurt, of participants of the ‘Econometric Society European Meeting’ in Vienna, the ‘Unit Root and Cointegration Testing’ conference in Faro and the ‘Second Italian Conference on Econometrics and Empirical Economics’ in Rimini, and of seminar participants at the Institute for Advanced Studies in Vienna, the University of Oslo and Concordia University Montreal are gratefully acknowledged. The usual disclaimer applies.

References


A Preliminaries

In this first appendix several preliminary lemmas are collected. We start with a lemma that 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 \( |z| \leq 1 \). Denote with \( 0 \leq \omega_1 < \cdots < \omega_l \leq \pi \) a set of frequencies and denote with \( D(z) := \Delta_{\omega_1}(z) \cdots \Delta_{\omega_l}(z) \). Assume that there exists an integer \( L \geq H \) for \( H := \sum_{k=1}^{l} (1 + \mathbb{I}(0 < \omega_k < \pi)) \) denoting the degree of \( D(z) \) such that \( \sum_{j=0}^{\infty} j^{1/2+L} \|c_j\| < \infty \). Further define \( c_k := \tilde{c}_k(1 + \mathbb{I}(0 < \omega_k < \pi)) \). Further define \( J_k \) and \( S_k \) as in (5). Then there exist matrices \( C_k \in \mathbb{R}^{s \times c_k}, K_k \in \mathbb{R}^{c_k \times s} \) 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

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

(iii) Representation \( (11) \) 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 \( 0 \leq \omega_1 < \cdots < \omega_H < 2\pi \) 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 \( \tilde{D}_{-k} := D(z)/(1 - z\overline{z_k}) \).

The proof is inductive in the unit roots. Setting \( r = 0 \) and \( c^{(0)}(z) = c(z) \) starts the induction. Thus in order to prove the induction step assume that the following representation has been obtained already:

\[
c^{(r)}(z) = c(z) - \sum_{k=1}^{r} \frac{z\tilde{D}_{-k}(z)}{\overline{z_k}\tilde{D}_{-k}(\overline{z_k})} \tilde{C}_k \tilde{K}_k
\]
with \( c^{(r)}(z) = (1 - zz_1) \ldots (1 - zz_r) c^{(r)}(z) \) such that \( c^{(r)}(z) = \sum_{j=0}^{\infty} c_j^{(r)} z^j \), with \( \sum_{j=0}^{\infty} j^{1/2+L-r} \| c_j^{(r)} \| < \infty \). Now consider \( \hat{c}^{(r+1)}(z) = c^{(r)}(z) - \frac{z \hat{D}_{r+1}(z)}{\hat{z} \hat{D}_{r+1}(\hat{z})} \hat{C}_{r+1} \hat{K}_{r+1} / [(1 - zz_1) \ldots (1 - zz_r)]. \) By inserting it follows immediately that \( \hat{c}^{(r+1)}(\hat{z}) = 0 \). We can thus write \( \hat{c}^{(r+1)}(z) = (1 - zz_{r+1}) c^{(r+1)}_{r+1}(z) \). Also, since \( c^{(r+1)}(z) \) and \( c_j^{(r)}(z) \) differ only by a polynomial they have the same summability properties. We can write \( c^{(r+1)}(z) = \sum_{j=0}^{\infty} c_j^{(r+1)} z^j = (1 - zz_{r+1}) \sum_{j=0}^{\infty} c_j^{(r+1)} z^j \) and using a formal power series expansion we obtain: \( c_0^{(r+1)} = I_s \) and \( c_j^{(r+1)} = c_j^{(r+1)} + z z_{r+1} c_j^{(r+1)}_{r+1} \), which implies

\[
\hat{c}_{j,•}^{(r+1)} = \sum_{i=0}^{j} \hat{c}_{j-i}^{(r+1)} z_{r+1}^{-i} = \sum_{i=0}^{j} \hat{c}_i^{(r+1)} z_{r+1}^{-i} = \sum_{i=0}^{j} \hat{c}_i^{(r+1)} z_{r+1}^{-i} = -z_{r+1} \sum_{i=j+1}^{\infty} \hat{c}_i^{(r+1)} z_{r+1}^{-i}
\]

using \( \hat{c}^{(r+1)}(\hat{z}) = \sum_{i=0}^{\infty} \hat{c}_i^{(r+1)} \hat{z}^{-i} = 0 \). Therefore

\[
\sum_{j=0}^{\infty} j^{1/2+L-r-1} \| c_j^{(r+1)} \| = \sum_{j=0}^{\infty} j^{1/2+L-r-1} \left\| \hat{c}_i^{(r+1)} z_{r+1}^{-i} \right\| \\
\leq \sum_{j=0}^{\infty} j^{1/2+L-r-1} \| \hat{c}_i^{(r+1)} \| \leq \sum_{i=1}^{\infty} \sum_{j=0}^{i-1} j^{1/2+L-r-1} \| \hat{c}_i^{(r+1)} \| \\
\leq \sum_{i=1}^{\infty} (i-1) i^{1/2+L-r-1} \| \hat{c}_i^{(r+1)} \| < \infty,
\]

using the induction hypothesis.

For \( r = H \) the following representation is obtained

\[
c(z) = z \sum_{k=1}^{H} \frac{\hat{D}_{-k}(z)}{\hat{z} \hat{D}_{-k}(\hat{z})} \hat{C}_{k} \hat{K}_{k} + D(z) c_{•}(z).
\]

Evaluating the above representation at \( z_k, k = 1, \ldots, H \) shows uniqueness up to the decomposition \( c(e^{-i\omega_k}) = \hat{C}_{k} \hat{K}_{k} \).

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 \( z_k = z_j \) and it holds that

\[
z \frac{\hat{D}_{-k}(z)}{\hat{z} \hat{D}_{-k}(\hat{z})} \hat{C}_{k} \hat{K}_{k} = \frac{\hat{D}_{-j}(\bar{z})}{\hat{z} \hat{D}_{-j}(\bar{z})} \hat{C}_{j} \hat{K}_{j}
\]

since \( \hat{C}_{k} \hat{K}_{k} = c(z_k) = c(\bar{z}_j) = \hat{C}_{j} \hat{K}_{j} \). Noting that \( \hat{D}_{-j}(z) = D_{-j}(z)(1 - zz_j) \) we obtain

\[
\frac{\hat{D}_{-k}(z)}{z_k \hat{D}_{-k}(\hat{z})} \hat{C}_{k} \hat{K}_{k} + \frac{\hat{D}_{-j}(z)}{\bar{z}_j \hat{D}_{-j}(\bar{z})} \hat{C}_{j} \hat{K}_{j} = D_{-k}(z) \left[ \frac{\hat{C}_{k} \hat{K}_{k}(1 - zz_j)}{(1 - z_j^2) z_k \hat{D}_{-k}(\bar{z})} + \frac{\hat{C}_{j} \hat{K}_{j}(1 - zz_k)}{(1 - z_k^2) \bar{z}_j \hat{D}_{-j}(\bar{z})} \right].
\]
To obtain the real valued expression given in the formulation of the lemma define for complex roots:

\[ C_k := \left[ \mathcal{R}\{z_k\bar{C}_k/[(1 - z_j^2)D_{-k}(\pi_k)]\}, \quad \mathcal{I}\{z_k\bar{C}_k/[(1 - z_j^2)D_{-k}(\pi_k)]\} \right], \quad K_k := \left[ \frac{2\mathcal{R}(\bar{\hat{K}}_k)}{2\mathcal{I}(\bar{\hat{K}}_k)} \right] \]

where \( \mathcal{R} \) and \( \mathcal{I} \) denote the real and the imaginary part of a complex quantity. For real roots define \( C_k = \bar{C}_k/(z_kD_{-k}(z_k)), K_k = \bar{\hat{K}}_k \) noting that for real roots \( \hat{D}_{-k}(z) = D_{-k}(z) \) 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 \( z_k := e^{\text{i}x_k} \) and is defined in equations (5) and corresponds to a minimal representation as in Theorem 1 and \((\varepsilon_t)_{t \in \mathbb{Z}}\) fulfills Assumption 2. Let \( v_t := c(L)\varepsilon_t \) for \( c(z) = \sum_{j=0}^{\infty} c_j z^j \) with \( \sum_{j=0}^{\infty} |c_j| < \infty \), hence \( c(z) \) is analytic on the closed unit disc.

We denote the stacked process as \( x_t := [x_1^T, \ldots, x_l^T]^T \) and define \( G^v(j) := \langle v_{t-j}, v_t \rangle_{T+1}^j \) for \( j = 0, \ldots, T-1 \), \( G^\nu(j) := G^v(\text{int} j) \) for \( j = -T + 1, \ldots, -1 \), \( G^\nu(j) := 0 \) for \( |j| \geq T \) and \( \Gamma^v(j) := E v_{t-j} v_t^\prime \). 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:

\[ \max_{0 \leq j \leq H_T} \|G^v(j) - \Gamma^v(j)\| = O((\log T/T)^{1/2}), \quad (12) \]

\[ T^{-1}\langle x_t, x_{t+1} \rangle_{1}^T \xrightarrow{d} W, \quad (13) \]

\[ \max_{0 \leq j \leq H_T} \|\langle x_t, v_{t-j} \rangle_{j+1}^T\| = O_P(1). \quad (14) \]

Here it holds that \( W \) is a.s. and thus it follows that \( [T^{-1}\langle x_t, x_{t+1} \rangle_{1}]^{-1} = O_P(1) \), 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 (13) has to be changed appropriately in this case.

**Proof:**

**Proof of (i):** Equation (12) 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 since also our assumptions on the noise $(\varepsilon_t)_{t \in \mathbb{Z}}$ are sufficient.

The second result (13) 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 $t_1, \ldots, t_{2l}$ of the random vector $X_n(u, v, t_1, \ldots, t_{2l})$ are essentially (up to changes in the summation limits) equal to $\sqrt{2} \sum_{s=1}^{l-1} \sin(\theta_k s) \varepsilon_s$ and $\sqrt{2} \sum_{s=1}^{l-1} \cos(\theta_k s) \varepsilon_s$. Now,

$$x_{t,k} = \sum_{s=1}^{l-1} J_k^{s-1} K_{k,t-s} = J_k^{l-1} \sum_{s=1}^{l-1} J_k^{s-t} K_{k,t-s} = J_k^{l-1} \sum_{s=1}^{l-1} J_k^{s-t} K_{k,t-s}$$

where $[(u^1_s), (u^2_s)]' = K_{k} \varepsilon_s$. Thus, for any $x \neq 0$, $x' J_k^{l-t} x_{t,k}$ 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 $J_k^{l-t} x_{t,k}$, when scaled by $T^{-1/2}$, to a multivariate Brownian motion with zero mean and variance $V$.

For establishing non-singularity of the limiting variance $V$ 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 of the corresponding diagonal block of $V$ 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 $\omega_k$) 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 $J_k^{l-t} x_{t,k}$) is singular, if and only if there exists a non-zero vector $x' = [x'_1, x'_2]'$ such that the variances of both $x'_1 u^1_s + x'_2 u^2_s$ (corresponding to the cosine terms) and of $x'_1 u^2_s - x'_2 u^1_s$ (corresponding to the sine terms) are zero. This is equivalent to $x' [K_k, J_k' K_k] = 0$. The latter matrix has full row rank due to minimality and thus the contradiction is shown and consequently $V > 0$.

Now the continuous mapping theorem can be applied to show that

$$T^{-1} \langle x_t, x_t \rangle_1^T = T^{-1} \sum_{t=1}^T J_t^{l-1} (J_t^{l-1} x_t / \sqrt{T}) (J_t^{l-1} x_t / \sqrt{T})' (J_t^{l-1} x_t / \sqrt{T})' \overset{d}{\to} W.$$
The a.s. non-singularity of $W$ follows from the non-singularity of the limiting covariance matrices of $J^{1-\frac{t}{|t|}}x_t/\sqrt{T}$. 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 $\mathbb{P}\{\lambda_{\min}(T^{-1}\langle x_t, x_t \rangle_T^T) < \varepsilon\} \rightarrow \mathbb{P}\{\lambda_{\min}(W) < \varepsilon\}$ showing that $[T^{-1}\langle x_t, x_t \rangle_T^T]^{-1} = O_P(1)$.

The bounds formulated in (14) 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 $c_{\bullet}(z) = \sum_{j=0}^{\infty} c_{\bullet,j} z^j$ is such that $\sum_{j=0}^{\infty} j^{1/2}||c_{\bullet,j}|| < \infty$. Using this decomposition we obtain $\varepsilon_t = c(L)\varepsilon_t = \alpha_1 \varepsilon_t + \alpha_2 \varepsilon_{t-1} + \Delta_{\omega_k}(L)\varepsilon_t^*$, with $\varepsilon_t^* := c_{\bullet}(L)\varepsilon_t$. Hence

$$\langle x_{t,k}, \varepsilon_{t-j} \rangle_{j+1}^{T} = J_k^{1}(x_{t-j,k}, \varepsilon_{t-j})_{j+1}^{T} + J_k^{1} \sum_{i=0}^{j-1} J_k^{1} K_k^i K_k^{j-i-1} \langle \varepsilon_{t-i-1}, \varepsilon_{t-j} \rangle_{j+1}^{T}$$

$$= J_k^{1}(x_{t-j,k}, \varepsilon_{t-j})_{j+1}^{T} + O(j (\log T/T)^{1/2}) + O(1).$$

Here the first equality stems immediately from the definition of $x_{t,k}$. The second equality follows from $\langle \varepsilon_{t-i-1}, \varepsilon_{t-j} \rangle_{j+1}^{T} = O((\log T/T)^{1/2})$ for $i < j-1$ and $\langle \varepsilon_{t-j}, \varepsilon_{t-j} \rangle_{j+1}^{T} = O(1)$. This last result follows from the uniform convergence of the estimated covariance sequence, i.e. by applying (12) to the stacked process $([\varepsilon_t^*, \varepsilon_{t-j}^*])_{t \in \mathbb{Z}}$. Here it has to be noted that the difference between $\langle \varepsilon_{t-j}, \varepsilon_{t-j} \rangle_{j+1}^{T}$ and $\langle \varepsilon_t, \varepsilon_t \rangle_{j+1}^{T}$ 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 $T^{-1} \sum_{t=1}^{T-j} x_{t,k} \varepsilon_t^*$. This term can be developed as follows:

$$\frac{1}{T} \sum_{t=1}^{T-j} x_{t,k} \varepsilon_t^* = \frac{1}{T} \sum_{t=1}^{T-j} x_{t,k} (\alpha_1 \varepsilon_t + \alpha_2 \varepsilon_{t-1} + \Delta_{\omega_k}(L)\varepsilon_t^*)$$

$$= \frac{1}{T} \sum_{t=1}^{T-j} x_{t,k} \varepsilon_t^* \alpha_1 + \frac{1}{T} \sum_{t=1}^{T-j} x_{t,k} \varepsilon_{t-1}^* \alpha_2 + \frac{1}{T} \sum_{t=1}^{T-j} x_{t,k} (\varepsilon_t^* - 2 \cos(\omega_k)\varepsilon_{t-1}^* + \varepsilon_{t-2}^*)$$.

That the first two terms above converge for $j = 0$, i.e. when summation takes place from 1 to $T$, can be shown using Theorem 2.4 of Chan and Wei (1988) (see also Theorem 6 of Johansen and Schaumburg, 1999). 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 \(T^{-1} \sum_{t=T-j+1}^{T} x_{t,k} \varepsilon'_t \alpha'_t\). We obtain that

\[
\mathbb{E} \max_{1 \leq j \leq H_T} \|T^{-1} \sum_{t=T-j+1}^{T} \text{vec}(x_{t,k} \varepsilon'_t)\|_1 \leq \frac{\sqrt{\Sigma}}{T} \sum_{t=T-H_T+1}^{T} (\mathbb{E}\|x_{t,k} \varepsilon'_t\|_{F_\varepsilon})^{1/2}
\]

\[
\leq \frac{\sqrt{\Sigma}}{T} \sum_{t=T-H_T+1}^{T} (\mathbb{E}\|x_{t,k}\|_2^2 \mathbb{E}\|\varepsilon_t\|_2^2)^{1/2}
\]

\[
\leq \frac{c}{T} \sum_{t=T-H_T+1}^{T} t^{1/2} \leq \frac{cH_T}{\sqrt{T}} \to 0,
\]

where we have used the inequality \(\mathbb{E}\|x_{t,k} \varepsilon'_t\|_{F_\varepsilon}^2 \leq \mathbb{E}\|x_{t,k}\|_2^2 \mathbb{E}\|\varepsilon_t\|_2^2\), which follows from \(\mathbb{E}\{\varepsilon_t \varepsilon'_t | \mathcal{F}_{-1}\} = \Sigma\) using \(\mathbb{E}\|x_{t,k} \varepsilon'_t\|_{F_\varepsilon}^2 = \mathbb{E}(x_{t,k}^T x_{t,k} \varepsilon'_t \varepsilon_t) = \mathbb{E}(x_{t,k}^T x_{t,k} \text{tr}(\varepsilon_t \varepsilon'_t))\). Therefore we have established uniform convergence in \(0 \leq j \leq H_T\) of \(T^{-1} \sum_{t=j}^{T} x_{t,k} \varepsilon'_t\) 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

\[
\frac{1}{T} \sum_{t=1}^{T-j} x_{t,k} (v_{t} - 2 \cos(\omega_k) v_{t-1} + v_{t-2})' = \frac{1}{T} \sum_{t=1}^{T-j} x_{t,k} (v_{t}')' - 2 \cos(\omega_k) \sum_{t=1}^{T-j} x_{t,k} (v_{t-1})' + \frac{1}{T} \sum_{t=1}^{T-j} x_{t,k} (v_{t-2})' \\
= \frac{1}{T} \sum_{t=1}^{T-j} x_{t,k} (v_{t}')' - 2 \cos(\omega_k) \frac{1}{T} \sum_{t=1}^{T-j} x_{t,k} (v_{t-1})' + \frac{1}{T} \sum_{t=1}^{T-j} x_{t,k} (v_{t-2})' \\
= \frac{1}{T} \sum_{t=1}^{T-j} x_{t,k} (v_{t-1})' \left[ x_{t+1,k} (v_{t}')' + \frac{1}{T} \sum_{t=1}^{T-j} x_{t+2,k} (v_{t}')' + o(1) \right] \\
= \frac{1}{T} \sum_{t=1}^{T-j} (K_k \varepsilon_{t+1} + J_k' K_k \varepsilon_t) (v_{t}')' + o(1).
\]

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 \(\varepsilon_t\) 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 (12) this term is \(O(1)\) uniformly in \(0 \leq j \leq H_T\), which concludes the proof of (14).

**Proof of (ii):** The sharper result for the case of rational transfer functions is obtained by replacing Theorem 7.4.3 used above by Theorem 5.3.2 (p. 167) in HD.

**Proof of (iii):** Recall the definition of the variable \(\hat{v}_t = v_t - \langle v_t, d_t \rangle_T^T (\langle d_t, d_t \rangle_T^{1/2})^{-1} d_t\), 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 \(\langle v_t, d_t \rangle_T^T = O((T^{-1} \log \log T)^{1/2})\). This fact allows to derive equation (12) 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 (13) for \( \hat{x}_t \). This result follows from the observation that for all \( k = 1, \ldots, l \) it holds using the continuous mapping theorem and the results achieved in the proof of (i) that \( T^{-1/2}\langle J^{1/2}x_t, d_{t,k} \rangle_T^T = O_P(1) \). 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 (14) 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 \( \langle v_t, d_t \rangle_1^T = O(T^{-1/2}) \) 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 \( -\langle x_t, d_t \rangle_1^T(\langle d_t, d_t \rangle_1^T)^{-1}\langle d_t, \hat{v}_{t-j} \rangle_T^{T} \).

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 \( H_T/T^{1/2} \rightarrow 0 \) shows that this term is uniformly in \( 0 \leq j \leq H_T \) of order \( o_P(1) \). Therefore we have established \( \max_{0 \leq j \leq H_T} \|\langle \hat{x}_t, \hat{v}_{t-j} \rangle_{j+1}^T\| = O_P(1) \). \( \square \)

**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 a random variable with finite second moment.

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

**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

\[
X^{-1} = \begin{bmatrix} A & B \\ B' & C \end{bmatrix}^{-1} = \begin{bmatrix} 0 & 0 \\ 0 & C^{-1} \end{bmatrix} + \begin{bmatrix} I \\ -C^{-1}B' \end{bmatrix} (A - BC^{-1}B')^{-1} \begin{bmatrix} I, -BC^{-1} \end{bmatrix}.
\]
The following inequality holds true:

\[
\|(A - BC^{-1}B^t)^{-1}\|_2 \leq \|X^{-1}\|_2 + \|C^{-1}\|_2.
\]  

(15)

B Proofs of the Theorems

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 ‘\^’ by showing that the differences that occur are asymptotically sufficiently small. As a side remark note here that HD use the symbol ‘\^’ for the YW estimator, whereas we use it for the LS estimator. In this paper the YW estimators carry the symbol ‘\~’. Note for completeness that the symbol ‘\~’ is also used otherwise, this, however, should not lead to any confusion.

Proof of (i), (ii): In Theorem 7.4.5 (p. 331) of HD it is shown that \(\max_{1 \leq j \leq p} \|\hat{\Phi}_p^v(j) - \Phi_p^v(j)\| = O((\log T/T)^{1/2})\) 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 \(\hat{\Phi}_p^v(j)\) and \(\Phi_p^v(j)\) is ‘small enough’ asymptotically. For example for the first bound this means that we have to show that \(\max_{1 \leq j \leq H_T} \max_{1 \leq j \leq p} \|\hat{\Phi}_p^v(j) - \Phi_p^v(j)\| = O((\log T/T)^{1/2})\) 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 \(\hat{\Theta}_p^v\), quantities of the form \(\langle v_{t-i}, v_{t-j} \rangle = T^{-1} \sum_{t=p+1}^{T} \langle v_{t-i}, v_{t-j} \rangle\) for \(i, j = 1, \ldots, p\) appear, whereas the YW estimator uses \(G^v(i - j) = T^{-1} \sum_{t=1+j-i}^{T} v_{t} v_{t-j+i}\) for \(j \geq i\) and the corresponding expression for \(j < i\). Thus, the difference between these two terms is given (discussing here only the case \(j \geq i\); with the case \(j < i\) following analogously) by:

\[
\langle v_{t-i}, v_{t-j} \rangle - G^v(i - j) = -\frac{1}{T} \sum_{t=1+j-i}^{p-i} v_{t} v_{t-j+i} - \frac{1}{T} \sum_{t=T-i+1}^{T} v_{t} v_{t-j+i}.
\]

We know from equation (12) in Lemma 3 that \(T^{-1} \sum_{t=S}^{T} v_{t} v_{t-r} - \Gamma^v(-r) = O((\log T/T)^{1/2})\) uniformly in \(0 \leq r \leq H_T\). This directly implies \(\langle v_{t-i}, v_{t-j} \rangle - G^v(i - j) = O((\log T/T)^{1/2})\) 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 (12), and \(\max_{1 \leq p \leq H_T} \|E V_{t,p}^{-1}\|_\infty < \infty\) and \(\max_{1 \leq p \leq H_T} \|E V_{t,p}^{-1}\|_\infty < \infty\), see
Then we obtain concerning the properties of autoregressive approximations also holds for the LS estimator.

Further $p$ uniformly in $p$ the key element is equation (7.4.7) (also dealing with the YW estimator). Inspection of the proof of this theorem shows that instead of (i) leads to the tighter bound $\hat{\Phi}_p^v(j) \leq G_T = (\log T)^a$ 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.

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} \epsilon_t v_{t-k}'. Then we obtain

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

uniformly in $j, k \leq p$, which follows from the assumption of finite fourth moments of $(v_t)_{t \in \mathbb{Z}}$. Further $p^{-1} \sum_{t=1}^{p} (\tilde{g}(j, k) - \Phi^v(j)) v_{t-j} v_{t-k}' = O(1)$ 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 $\epsilon_t = \sum_{j=0}^{\infty} \Phi^v(j) v_{t-j}$ that:

$$\sum_{j=1}^{p} \left( \hat{\Phi}_p^v(j) - \Phi^v(j) \right) \tilde{g}(j, k) = \sum_{j=1}^{p} \left( \hat{\Phi}_p^v(j) - \Phi^v(j) \right) (v_{t-j}, v_{t-k}) + o(T^{-1/2})$$

$$= -\langle \epsilon_t, v_{t-k} \rangle + \sum_{j=p+1}^{\infty} \Phi^v(j) \frac{1}{T} \sum_{t=p+1}^{T} v_{t-j} v_{t-k}' + o(T^{-1/2})$$

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

$$= -\hat{u}_k + \sum_{j=p+1}^{\infty} \Phi^v(j) \tilde{g}(j, k) + o(T^{-1/2})$$
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 $\tilde{P}_sv_{t-p}$. Note that in this proof ‘˜’ is used to denote quantities relating to the augmented regression and not to the YW estimators. Using the block-matrix inversion formula and (15) from Lemma 4 (with the blocks corresponding to $V_{t,p-1}^{-}$ and $\tilde{P}_sv_{t-p}$) it is straightforward to show that $\| (\tilde{V}_{t,p}^{-}, \tilde{V}_{t,p}^{-})^{-1} \| < \infty$ and $\| (V_{t,p}^{-}, V_{t,p}^{-})^{-1} \|_{\infty} < \infty$ a.s. for $T$ large enough, uniformly in $1 \leq p \leq H_T$. This can be used to show the approximation properties of the autoregression including $\tilde{P}_sv_{t-p}$ as follows:

$$\tilde{\Theta}_p^v := \langle v_t, \tilde{V}_{t,p}^{-} \rangle (\tilde{V}_{t,p}^{-})^{-1}$$

$$= \mathbb{E}v_t(\tilde{V}_{t,p}^{-})'(\mathbb{E}V_{t,p}^{-}(\tilde{V}_{t,p}^{-}))^{-1} + \left[ \langle v_t, \tilde{V}_{t,p}^{-} \rangle - \mathbb{E}v_t(\tilde{V}_{t,p}^{-})' \right] (\tilde{V}_{t,p}^{-})^{-1} + \mathbb{E}v_t(\tilde{V}_{t,p}^{-})'(\mathbb{E}V_{t,p}^{-}(\tilde{V}_{t,p}^{-}))^{-1} \left[ \mathbb{E}V_{t,p}^{-}(\tilde{V}_{t,p}^{-})' - \langle \tilde{V}_{t,p}^{-}, \tilde{V}_{t,p}^{-} \rangle \right] (\tilde{V}_{t,p}^{-})^{-1}.$$ 

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 $\tilde{P}_sv_{t-p}$ on the approximation formula derived for $\tilde{IC}^v (p; C_T)$. By construction it holds that

$$\tilde{\Sigma}_{p-1} = \langle v_t - \tilde{\Theta}_p^{v-1} V_{t,p-1}, v_t - \tilde{\Theta}_p^{v-1} V_{t,p-1} \rangle_p \geq \tilde{\Sigma}_p = \langle v_t - \tilde{\Theta}_p^{v} \tilde{V}_{t,p}, v_t - \tilde{\Theta}_p^{v} \tilde{V}_{t,p} \rangle_{p+1} \geq \tilde{\Sigma}_p^v.$$ 

36
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 \( \Sigma_p^v \) 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 \( (C^\perp)'y_t - J(C^\perp)'y_{t-1} = x_t + (C^\perp)'e_t - Jx_{t-1} - J(C^\perp)'e_{t-1} \). These relationships also immediately establish the expression given for \( y_t - CJ(C^\perp)'y_{t-1} \) and thus also the definition of \( \tilde{c}_*(z) \). Furthermore, \( \tilde{c}_*(0) = I_s \) follows immediately from \( c_*(0) = I_s \). The summability properties of \( \tilde{c}_*(z) \) follow directly from the analogous properties of \( c_*(z) \). Since \( c_*(z) \) has no poles inside the unit circle, neither has \( \tilde{c}_*(z) \), 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 \( \tilde{c}_*(z) \) note that from the representation of \( \tilde{c}_*(z) \) given in the theorem the following relation is obtained for \( |z| < 1 \):

\[
\tilde{c}_*(z) = \tilde{c}^{-1} \begin{pmatrix} (I - zJ) & 0 \\ 0 & I \end{pmatrix} D(z)^{-1} \tilde{C}c(z).
\]

Now, since by assumption \( \det c(z) \neq 0 \) for all \( |z| < 1 \) and \( \det D(z) \neq 0, |z| < 1 \) it follows that \( \det \tilde{c}_*(z) \neq 0, |z| < 1 \). Further \( \det \tilde{c}_*(z) \neq 0, |z| = 1 \) by Assumption 4.

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

\[
\tilde{y}_{tq+i} = \begin{pmatrix} C J^{q-1} \\ \vdots \\ C J \end{pmatrix} x_{tq+i} + \begin{pmatrix} I_s & CK & \cdots & C J^{q-2} K \\ \cdots & \cdots & \cdots \\ 0 & \cdots & 0 & I_s \end{pmatrix} \begin{pmatrix} \varepsilon_{tq+q-1+i} \\ \varepsilon_{tq+1+i} \\ \varepsilon_{tq+i} \end{pmatrix} + \sum_{j=1}^{\infty} c_j^{(q)} \tilde{\varepsilon}_{(t-j)q+i}
\]
with \( \tilde{\epsilon}_t := \left[ \tilde{\epsilon}'_{t+q-1}, \ldots, \tilde{\epsilon}'_{t+1}, \tilde{\epsilon}'_t \right]' \) and where the coefficients in \( \sum_{j=1}^{\infty} \tilde{c}^{(q)}_j \tilde{\epsilon}_{(t-j)q+i} \) can be obtained by cumbersome but straightforward computations. It is clear that \( (\tilde{\epsilon}_{tq+i})_{t \in \mathbb{Z}} \) is a martingale difference sequence with respect to the filtration \( \mathcal{F}_{j, q+i}, j \in \mathbb{Z} \). 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 \( \bar{\epsilon}_t := \mathcal{E}_q^{-1} \tilde{\epsilon}_t \) which correspondingly defines \( c^{(q)}(z) = \sum_{j=0}^{\infty} \tilde{c}^{(q)}_j z^j = \sum_{j=0}^{\infty} \bar{c}^{(q)}_j \mathcal{E}_q^{-1} z^j \). Summability of the coefficients of \( c^{(q)}(z) \) follows from summability of \( c_s(z) \). Since \( \mathcal{E}_q \) is block upper triangular with diagonal blocks equal to the identity matrix it follows that the last block of \( \bar{\epsilon}_t \) equals \( \epsilon_t \).

Note also that \( D(L) \hat{y}_t = \tilde{v}_t, \) with \( (\tilde{v}_t)_{t \in \mathbb{Z}} \) defined analogously to \( (\bar{y}_t)_{t \in \mathbb{Z}} \). Thus, the subsampling argument leads to \( q \) processes \( \hat{y}_{tq+i}, i = 1, \ldots, q \) for which part (i) of the lemma can be applied, since by construction these processes all have an observability equal to 1. □

**B.3 Proof of Theorem 3**

The discussion above the theorem in the main text shows that the regression of \( \hat{\epsilon}_t \) 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 in the same way as the regressor vector we obtain:

\[
[\hat{\beta}_1, \hat{\beta}_2] := \langle \hat{\epsilon}_t, Z_{t,p}^- \rangle (Z_{t,p}^-)^{-1} \tag{16}
= \langle \hat{\epsilon}_t, z_t \rangle \langle z_t, z_t \rangle^{-1} \begin{bmatrix} I_c, -\langle z_t, Z_{t,p,2}^- \rangle (Z_{t,p,2}^-)^{-1} \\ \bar{c}^{sx,c}, \langle \hat{\epsilon}_t, Z_{t,p,2}^- \rangle (Z_{t,p,2}^-)^{-1} \end{bmatrix},
\]

with \( \hat{\beta}_1 := \langle \hat{\epsilon}_t, z_t \rangle \langle z_t, z_t \rangle^{-1} \) and \( z_t^\Pi := z_t - \langle z_t, Z_{t,p,2}^- \rangle (Z_{t,p,2}^-)^{-1} Z_{t,p,2} \). Thus, \( z_t^\Pi \) 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, \( \hat{\Theta}_p^\epsilon = \langle \hat{\epsilon}_t, Z_{t,p,2}^- \rangle (Z_{t,p,2}^-)^{-1} \), contains only stationary quantities. In particular \( Z_{t,p,2}^- \) contains \( \hat{\epsilon}_{t-j}, j = 1, \ldots, p-1 \) and a part of \( \hat{\epsilon}_{t-p} \) as blocks. Thus, the asymptotic behavior of this term is covered by Theorem 2, from which we obtain \( \hat{\Theta}_p^\epsilon = O(\log T)^{1/2} \). 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 \( \begin{bmatrix} I_c, -\langle z_t, Z_{t,p,2}^- \rangle (Z_{t,p,2}^-)^{-1} \end{bmatrix} \). Note first that \( z_t = x_{t-1} + (C^\dagger)' \epsilon_{t-1} \) and again that \( Z_{t,p,2}^- \) contains only stationary variables. Therefore equation (14)
of Lemma 3 shows that \( \langle x_{t-1}, Z_{t,p,2}^{-} \rangle \) is \( O_P(1) \) uniformly in \( p \). Furthermore, Theorem 6.6.11 (p. 267) of HD and Assumption 4 imply that \( \| (Z_{t,p,2}^{-})^{-1} \|_\infty < M \) a.s. for some constant \( M < \infty \) for \( T \) large enough. Equation (12) implies that \( \langle (C^\dagger) e_{t-1}, Z_{t,p,2}^{-} \rangle = O_P(1) \) and hence \( \langle z_t, Z_{t,p,2}^{-} \rangle (Z_{t,p,2}^{-})^{-1} = O_P(1) \) uniformly in \( p \).

Consider \( \hat{\beta}_1 = \langle \epsilon_t, z_t^H \rangle \langle z_t^H, z_t^H \rangle^{-1} \) next. We start with the first term, i.e. with \( \langle \epsilon_t, z_t^H \rangle = \langle \epsilon_t, Z_{t,p,2}^{-} \rangle (Z_{t,p,2}^{-})^{-1} \langle z_t^{-}, z_t \rangle \). Using again (14) of Lemma 3 it follows that both \( \langle \epsilon_t, x_{t-1} \rangle \) and \( \langle Z_{t,p,2}^{-}, x_{t-1} \rangle \) are \( O_P(1) \) uniformly in \( p \). Due to Lemma 3 it follows that \( \langle \epsilon_t, e_t \rangle = O_P(1) \). Then

\[
\langle \epsilon_t, Z_{t,p,2}^{-} \rangle (Z_{t,p,2}^{-})^{-1} = \Theta_p + O((T^{-1} \log T)^{1/2})
\]

shows that \( \langle \epsilon_t, z_t^H \rangle = O_P(1) \) uniformly in \( 1 \leq p \leq H_T \).

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

\[
T^{-1} \langle z_t^H, z_t^H \rangle = T^{-1} \langle z_t, z_t \rangle - T^{-1} \langle z_t, Z_{t,p,2}^{-} \rangle (Z_{t,p,2}^{-})^{-1} \langle Z_{t,p,2}^{-}, z_t \rangle.
\]

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 \( \langle z_t, \epsilon_{t-j} \rangle \) together with the established properties of \( \langle Z_{t,p,2}^{-}, Z_{t,p,2}^{-} \rangle \) immediately implies that it is of order \( O_P(p T^{-1}) \), which is due to our restriction that \( 1 \leq p \leq H_T \) in fact \( o_P(1) \).

Therefore we obtain

\[
P\left\{ \| \langle \epsilon_t, z_t^H \rangle (T^{-1} \langle z_t^H, z_t^H \rangle)^{-1} \| > M \right\} \leq P \left\{ \| (\epsilon_t, z_t^H) \| \| (T^{-1} \langle z_t^H, z_t^H \rangle)^{-1} \| > M \right\}
\leq P \left\{ \| (\epsilon_t, z_t^H) \| > \sqrt{M} \right\} + P \left\{ \| (T^{-1} \langle z_t^H, z_t^H \rangle)^{-1} \| > \sqrt{M} \right\}
\leq \eta/2 + P \left\{ \lambda_{\min}(T^{-1} \langle z_t^H, z_t^H \rangle) < 1/\sqrt{M} \right\} \leq \eta.
\]

In the above expression the first probability can be made arbitrarily small by choosing \( M \) large enough, since \( \langle \epsilon_t, z_t^H \rangle = O_P(1) \) and the second probability can be made arbitrarily small since \( T^{-1} \langle z_t^H, z_t^H \rangle = T^{-1} \langle z_t, z_t \rangle + o_P(1) \nrightarrow W \), 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 \( \hat{\Sigma}_p^y - \bar{\Sigma}_p^y = O_P(T^{-1}) \) uniformly in \( p \), where \( \bar{\Sigma}_p^y \) denotes the error covariance matrix from the LS regression of \( \hat{\epsilon}_t \) on \( Z_{t,p,2}^{-} \).

The results in item (ii) apply to the case when \( (y_t)_{t \in \mathbb{Z}} \) does not follow a finite order vector autoregression. Thus, it is important to first verify that the properties of \( (y_t)_{t \in \mathbb{Z}} \)
respectively \((\tilde{e}_t)_{t \in \mathbb{Z}}\) being finite order autoregressive processes are closely related. First, if 
\((\tilde{e}_t)_{t \in \mathbb{Z}}\) is generated by a finite order autoregression, so is \((y_t)_{t \in \mathbb{Z}}\) since \(y_t - C J (C^\dagger)' y_{t-1} = \tilde{e}_t\). A similar argument shows that if \((\tilde{e}_t)_{t \in \mathbb{Z}}\) is a VARMA process, then also \((y_t)_{t \in \mathbb{Z}}\) is a VARMA process and not a finite order autoregressive process, since \(\tilde{e}_t(z)\) has no determinantal zeros on the unit circle because of Assumption 4. Clearly, \(y_t - C J (C^\dagger)' y_{t-1} = \tilde{e}_t\) also implies that 
if \((\tilde{e}_t)_{t \in \mathbb{Z}}\) is not a rational process neither is \((y_t)_{t \in \mathbb{Z}}\).

These preliminary observations imply that the bounds in (ii) are established once \(\hat{\Sigma}^y_p - \tilde{\Sigma}^\epsilon_p = O_P(T^{-1})\) is shown, using Theorem 2(v) to obtain bounds for \(\tilde{\Sigma}^\epsilon_p\). To this end consider
\[
\hat{\Sigma}^y_p = \langle \tilde{e}_t - \hat{\beta}_1 z_t - \hat{\beta}_2 p Z_t, \tilde{e}_t - \hat{\beta}_1 z_t - \hat{\beta}_2 p Z_t \rangle \quad \text{and} \quad \tilde{\Sigma}^\epsilon_p = \langle \tilde{e}_t - \hat{\Theta}_p e_t, \tilde{e}_t - \hat{\Theta}_p e_t \rangle,
\]
where
\[
\tilde{e}_t - \hat{\Theta}_p e_t = \tilde{e}_t + \hat{\beta}_1 z_t + \hat{\beta}_2 p Z_t.
\]
These preliminary observations imply that the bounds in (ii) are established once \(\hat{\Sigma}^y_p - \tilde{\Sigma}^\epsilon_p = O_P(T^{-1})\) is shown, using Theorem 2(v) to obtain bounds for \(\tilde{\Sigma}^\epsilon_p\). To this end consider
\[
\hat{\Sigma}^y_p - \tilde{\Sigma}^\epsilon_p = \langle \tilde{e}_t - \hat{\Theta}_p e_t, \hat{\beta}_1 z_t \rangle + \langle \hat{\beta}_1 z_t, \tilde{e}_t - \hat{\Theta}_p e_t \rangle = \langle \hat{\beta}_1 z_t, \tilde{e}_t - \hat{\Theta}_p e_t \rangle.
\]
Recall that \(\hat{\beta}_1 = O_P(T^{-1}), \langle z_t, z_t \rangle = O_P(T), \langle e_t, e_t \rangle = O_P(1)\) and \(\langle z_t, Z_t \rangle = 0\) have been shown above. Using these results we obtain that \(\hat{\Sigma}^y_p - \tilde{\Sigma}^\epsilon_p = O_P(T^{-1})\) uniformly in \(p\).

To arrive at the result as formulated in the theorem it remains to show that all terms of order \(O_P(T^{-1})\) are in fact
\[
o_P \left( \frac{ps^2(C_T - 1)}{T} + \text{tr} \left[ (\Sigma^\epsilon)^{-1}(\Sigma^\epsilon_{p-1} - \Sigma^\epsilon) \right] \right).
\]
We show the claim by showing that \(T\) times the minimum over \(p \geq 0\) of (17) tends to infinity. Let \(p_T\) denote the largest integer such that \(\text{tr} \left[ (\Sigma^\epsilon)^{-1}(\Sigma^\epsilon_{p_T-1} - \Sigma^\epsilon) \right] > T^{-1/2}\). Since \(\Sigma^\epsilon\) converges monotonously to \(\Sigma^\epsilon\) with strict inequality for all finite \(p\) it follows that \(p_T \to \infty\) for \(T \to \infty\). Therefore, if \(C_T \to \infty\), then for \(p > 0\)
\[
T \frac{ps^2(C_T - 1)}{T} = ps^2(C_T - 1) \to \infty
\]
and the claim follows. If \(1 < C_T \leq M < \infty\) then consider
\[
\min_{p \geq 0} ps^2(C_T - 1) + T \text{tr} \left[ (\Sigma^\epsilon)^{-1}(\Sigma^\epsilon_{p-1} - \Sigma^\epsilon) \right] =
\]
\[
\min \left( \min_{0 \leq p \leq p_T} ps^2(C_T - 1) + T \text{tr} \left[ (\Sigma^\epsilon)^{-1}(\Sigma^\epsilon_{p-1} - \Sigma^\epsilon) \right], \min_{p_T \leq p} ps^2(C_T - 1) + T \text{tr} \left[ (\Sigma^\epsilon)^{-1}(\Sigma^\epsilon_{p-1} - \Sigma^\epsilon) \right] \right).
\]
For $0 \leq p \leq p_T$ by definition of $p_T$ the second term is bounded from below by $T T^{-1/2} = T^{1/2} \to \infty$ and for $p \geq p_T$ the first term is bounded from below by $p_T s(C_T - 1) \to \infty$ since $p_T \to \infty$. This shows the result.

**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 without Assumption 3.

**B.4 Proof of Theorem 4**

**Proof of (i):** In Lemma 1(ii) the stacked process $\tilde{y}_t := [y_{t+q-1}', \ldots, y_{t+2}', y_{t+1}', y_t']'$ is defined. The sub-sampled version $(\tilde{y}_{tq+i})_{t \in \mathbb{Z}}$, for $i = 1, \ldots, q$ fulfill assumptions 1 to 4 of Theorem 3 and have observability index equal to 1. Thus, for these processes the results established in Theorem 3 apply.

We consider autoregressive approximations of order $p = \tilde{p}q$ for some integer $\tilde{p}$ and consider for simplicity only sample sizes $T = \tilde{T}q$, i.e. for sample sizes that are integer multiples of $q$. First $p$ is treated as fixed integer and afterwards uniformity in this integer is established. The expressions for values of $T$ that are not integer multiples of $q$, for which we define $\tilde{T} = \lfloor T/q \rfloor$, differ from the expressions considered below only by at most $q - 1$ terms which does not change any of the error bounds provided below.

Theorem 3 provides uniform error bounds for the regressions of $(\tilde{y}_{tq+i})_{t \in \mathbb{Z}}$. Considering only the last block-row of the regression of $(\tilde{y}_{tq+i})_{t \in \mathbb{Z}}$ on $Y_{tq+i, \tilde{p}q}^-$ we obtain the coefficients of a regression of $y_{tq+i}$ on $p$ of its lags (i.e. on $y_{tq+i-1}, \ldots, y_{tq+i-p}$), for which due to sub-sampling the sample size is reduced to $\tilde{T}$. Denoting the corresponding coefficients of this sub-sampled regression with $\hat{\Theta}^{y,(i)}_p$ and using $Z_{tq+i, \tilde{p}q}^- = \tilde{T}_p^{-1} Y_{tq+i, \tilde{p}q}^-$ we obtain

\[
\hat{\Theta}^{y,(i)}_p = \tilde{T}_p^{-1} \langle \tilde{y}_{tq+i}, Z_{tq+i, \tilde{p}q}^- \rangle_{\tilde{p}}^{-1} \langle Z_{tq+i, \tilde{p}q}^-, Z_{tq+i, \tilde{p}q}^- \rangle_{\tilde{p}}^{-1} \tilde{T}_p
\]

for $i = 1, \ldots, q$. Here again in $Z_{tq+i, \tilde{p}q}^-$ only the first $c$ coordinates are integrated and do not depend on $p$ while the remaining coordinates are lags of a stationary process where with increasing $p$ more lags are added. Straightforward calculations show that

\[
\hat{\Theta}^y_p = \langle y_t, Y_{t,p/p+1}^- \rangle_{T} T^{-1} \langle Y_{t,p}, Y_{t,p/p+1}^- \rangle_{T} T^{-1} \sum_{i=1}^q \hat{\Theta}^{y,(i)}_p \langle Z_{tq+i, \tilde{p}q}^-, Z_{tq+i, \tilde{p}q}^- \rangle_{\tilde{p}}^{-1} \langle Z_{tq+i, \tilde{p}q}^-, Z_{tq+i, \tilde{p}q}^- \rangle_{\tilde{p}}^{-1} \tilde{T}_p
\]
which then leads to

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

From (the proof of) Theorem 3(i) we obtain that \((\hat{\Theta}_p^{y,(i)} - \Theta_p^{y}) \tilde{T}_p^{-1} = O_P((\tilde{T}^{-1} \log(\tilde{T}))^{1/2}) = O_P((\log T/T)^{1/2})\). The proof of Theorem 3(i) actually shows that the first \(c\) components of \(\hat{\Theta}_p^{y,(i)} - \Theta_p^{y}\) are \(O_p(T^{-1})\). Letting \(D_T = \text{diag}(T^{-1/2}I_c, I)\) one obtains \(D_T(Z_{t_{q+i}, \tilde{p}q}, Z_{t_{q+i}, \tilde{p}q})_{\tilde{p}}^{-1} D_T(D_T(Z_{t, p}, Z_{t, p})_{p+1}^{-1} D_{T})^{-1} \frac{d_i}{Z_{11}^{(i)}} \frac{0}{q} I\)

for \(i = 1, \ldots, q\), where \(\langle z_{t_{q+i}}, z_{t_{q+i}} \rangle_{\tilde{p}}^{-1}((z_t, z_t)_{p+1}^{-1} d_i \rightarrow Z_{11}^{(i)}\), with \(z_{t_{q+i}}\) and \(z_t\) denoting the \(c\) nonstationary components of \(Z_{t_{q+i}}\) and \(Z_t\). In the matrix above, the off diagonal terms are of order \(O_p(T^{-1/2})\) uniformly in \(1 \leq p \leq H_T\). All evaluations here follow directly from Lemma 3(i). The sum over \(q\) terms does not change any order of convergence and thus it follows that \(\hat{\Theta}_p^y - \Theta_p^y = O_P((\log T/T)^{1/2})\) for \(p = \tilde{p}q\).

**Proof of (ii):** The sub-sampled and stacked processes \((\tilde{e}_{t_{q+i}})_{t \in \mathbb{Z}}\) are defined based upon \(\tilde{e}_t := \tilde{z}_t^{(q)}(L^q)\tilde{e}_t\) where \(\tilde{z}_t := \mathcal{E}_f[\tilde{e}_{t+q-1}, \ldots, \tilde{e}_t]'\). Since \(\tilde{z}_t^{(q)}(L^q)\) is invertible due to Assumption 4 it follows that there exists a transfer function \(\tilde{\Phi}(L^q)\) such that \(\tilde{\Phi}(L^q)\tilde{e}_t = \tilde{e}_t\). The last block equation (note that \(\mathcal{E}_f\) is block upper triangular) here states that \(\tilde{e}_t\) can be obtained by filtering \(\tilde{e}_t\) which implies that \(\tilde{e}_t = \tilde{I}_s\tilde{\Phi}(L^q)\tilde{e}_t\). Therefore for \(p = \tilde{p}q, \tilde{p} \in \mathbb{N} \cup \{0\}\) consider

\[ \hat{\Sigma}_p^y = \langle y_t - \hat{\Theta}_p^y Z_{t,p}, y_t - \hat{\Theta}_p^y Z_{t,p} \rangle_{p+1}^{-1} \rightarrow \langle y_t - \Theta_p^y Z_{t,p}, y_t - \Theta_p^y Z_{t,p} \rangle_{p+1}. \]

Partition again \(Z_{t,p} = [z_t, Z_{tuman+2}]'\), where \(z_t \in \mathbb{R}^c\) contains the integrated components and \(Z_{tuman+2} \in \mathbb{R}^{bs-c}\) contains the stationary components, then it can be shown as in the proof of Theorem 3 but now using a sub-sampling argument, that uniformly in \(0 \leq \tilde{p} \leq H_T\)

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

In \(Z_{t,p+2}\) only a sub-block of \(\tilde{e}_t, \tilde{p}q\) is contained. Similarly as in the proof of Theorem 3(ii) irrespectively excluding \(\tilde{e}_t, \tilde{p}q\) leads to the upper and lower bounds for \(\hat{\Sigma}_p^y\).

Hence it remains to analyze \(\hat{\Sigma}_p^y\). Analogously to the corresponding part of 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 \((\tilde{e}_{t_{q+i}})_{t \in \mathbb{Z}}\) but only on a sub-vector obtained by pre-multiplying with \(\tilde{I}_s\). Second, the sub-sampled processes \((\tilde{e}_{t_{q+i}})_{t \in \mathbb{Z}}\) 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\(^1\) for \(k = 1, \ldots, p\)

\[
\sum_{j=1}^{p} \hat{I}_s \left( \hat{\Phi}_p^{(q)}(j) - \Phi^{(q)}(j) \right) \tilde{y}(j, k) = -\hat{u}_k + \sum_{j=p+1}^{\infty} \hat{I}_s \Phi^{(q)}(j) \tilde{y}(j, k) + o(T^{-1/2}),
\]

where \(\hat{u}_k = T^{-1} \sum_{t=qk+1}^{T} \varepsilon_t \tilde{e}_{t-kq}^p\) and \(\hat{\Phi}_p^{(q)}(j)\) denotes the least squares estimates in the regression for fixed \(p\). Let further \(\Phi^{(q)}(j)\) denote the true coefficients in the AR(\(\infty\)) representation. 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 \((\phi_p^{(2)})' \{\Gamma_{22} - \Gamma_2 \Gamma_1^{-1} \Gamma_{12}\} \phi_p^{(2)} = \hat{I}_s \Sigma_p \hat{\Sigma}_p - \Sigma\). Therefore in order to establish (7.4.32) on p. 341 of HD it is sufficient to show that \(\hat{G}_{22} - \hat{G}_{21} \hat{G}_{11}^{-1} \hat{G}_{12}\) can be replaced by its expectation introducing an error of magnitude \(o(p/T)\). For \(\hat{G}_{22}\) this again follows by sub-sampling and decomposing the sum over all \(t\) involved in the formation of \(\hat{G}_{22}\) 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

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

(here the scalar case is shown), which essentially involves replacing (in the notation of HD) \(\Gamma_{11}^{-1/2} y(t, p)\) with \(\varepsilon(t, p)\). In our setup this amounts to replacing \(\Gamma_{11}^{-1/2} [\varepsilon_{t-q}, \varepsilon_{t-2q}, \ldots, \varepsilon_{t-p}]'\) with \([\varepsilon'_{t-1}, \varepsilon'_{t-2}, \ldots, \varepsilon'_{t-p}]'\). 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 \(\tilde{e}_t = \hat{\Phi}(L^q) \varepsilon_t\) which follow from the assumptions on \(\hat{\Phi}^{(q)}(L^q)\). The rest of the proof of Theorem 7.4.6 of HD uses only properties of \(\varepsilon_t\). Then Theorem 7.4.6 of HD shows the required approximation for \(p = \tilde{p}q\) for any integer \(\tilde{p}\).

---

\(^1\)Note that HD use \(h\) instead of \(p\) to denote the lag length.
Now given any value of $p \in \mathbb{N}$ we use as in the proof of Theorem 2(v) with $\tilde{p} := \lfloor p/q \rfloor$

$$\langle y_t - \tilde{\Theta}_y^{\tilde{p}+1} (\tilde{p}+1)q \rangle \leq \langle y_t - \tilde{\Theta}_y^{\tilde{p}+1} (\tilde{p}+1)q \rangle^T (\tilde{p}+1)q \leq \hat{\Sigma}_y^p \leq \langle y_t - \tilde{\Theta}_y^{\tilde{p}+1} (\tilde{p}+1)q \rangle^T (\tilde{p}+1)q + 1.$$

Then using the result for $\tilde{pq}$ 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 $\Sigma_{\tilde{p}}^e > \Sigma^e$, $\Sigma_{\tilde{p}}^e \to \Sigma^e$ for $p \to \infty$ and the fact that the penalty term $C_T/T$ tends to zero by assumption.

**Proof of (ii):** The proof is based on the arguments outlined in 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

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

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

$$L_T(l_T(C_T)) \leq L_T(\tilde{p}(C_T)) \leq IC'^u(\tilde{p}(C_T); C_T)(1 + o_P(1)) \leq IC'^u(q(l_T(C_T) - 1); C_T)(1 + o_P(1)) \leq L_T(l_T(C_T))(1 + o_P(1)).$$

Here the first inequality follows from optimality of $l_C(C_T)$ with respect to $L_T$, the second from the lower bound of Theorem 3(ii) (or Theorem 4(ii) resp.) and $\tilde{p}(C_T) \to \infty$, the third from optimality of $\tilde{p}(C_T)$ with respect to $IC'^u(p, C_T)$ 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.