Persistence-Robust Surplus-Lag Granger Causality Testing

Dietmar Bauer
Arsenal Research
Vienna, Austria

Alex Maynard*
Department of Economics
University of Guelph, ON, Canada

June 29, 2010

Abstract

Previous literature has introduced causality tests with conventional limiting distributions in I(0)/I(1) vector autoregressive (VAR) models with unknown integration orders, based on an additional surplus lag in the specification of the estimated equation, which is not included in the tests. By extending this surplus lag approach to an infinite order VARX framework, we show that it can provide a highly persistence-robust Granger causality test that accommodates i.a. stationary, nonstationary, local-to-unity, long-memory, and certain (unmodelled) structural break processes in the forcing variables within the context of a single χ^2 null limiting distribution.

JEL Classification: C12, C32 *Keywords*: Granger causality, surplus lag, nonstationary VAR, local-to-unity, long-memory, structural breaks, predictive test, forward rate unbiasedness

---

*Corresponding author. Previously entitled "Robust Granger Causality Tests in the VARX Framework." We thank two anonymous referees, Zahid Asghar, Richard Baillie, Lynda Khalaf, J.M. Dufour, David Hendry, Peter Phillips, and participants at the Singapore Management University Conference in Honour of Peter C. B. Phillips, the 2006 NBER-NSF Time Series Conference, the 2nd International Workshop on Computational and Financial Econometrics, and the Canadian Econometric Study Group for useful discussion. This work was started while the authors were visiting the Cowles Foundation and we gratefully acknowledge their hospitality. The postdoc position of Bauer at the Cowles Foundation was financed by the Max Kade Foundation which is gratefully acknowledged. Maynard thanks the SSHRC for research funding.

1 Introduction

Since its introduction in Granger (1969), tests of Granger noncausality have become ubiquitous in economics, with recent applications ranging from the relationship between exchange rates and fundamentals (Engel and West, 2005) to tests for cycles of violence in the Palestinian-Israeli conflict (Jaeger and Paserman, 2008). Return predictability tests (e.g. Stambaugh, 1999) are also arguably interpretable as special cases of causality tests and there is a rich recent econometric literature on causality testing.[1]

Previous literature has shown that the addition of an untested surplus-lag leads to flexible inference in I(0)/I(1) VAR models with unknown integration orders (Toda and Yamamoto, 1995; Dolado and Lütkepohl, 1996; Saikkonen and Lütkepohl, 1996). By adapting this surplus-lag approach to an infinite order VARX setting, we show that it provides a Granger causality test that is particularly robust to the degree and nature of the persistence in the causing variables. The proposed causality test[2] has the same null limit distribution regardless of whether the causal variable is I(0), I(1), local-to-unity, long-memory/fractionally integrated, or subject to breaks in mean. Consequently, no pre-estimation or pre-test of persistence parameters is required.

These are desirable characteristics for several reasons. The practical difficulties associated with distinguishing $I(1)$ and $I(0)$ processes are well known. Moreover, processes with near unit roots may often be better modelled as local-to-unity (Phillips, 1987; Chan, 1988), against which unit root tests are inconsistent by design. Likewise, structural breaks can be confused with long-memory processes (e.g. Diebold and Inoue, 2001). Thus it can be difficult to determine with confidence the correct model for persistent data. As Phillips (2003, p. C35) puts it "no one really understands trends, even though most of us see trends when we look at economic data."

These distinctions are important to model specification, determining, e.g., whether a VAR is specified in levels, first-differences or error-correction format. Likewise, structural breaks require explicit modelling and long-memory processes are not easily accommodated in a VAR setting. Such choices have played an important practical role in some recent macroeconomic debates (e.g. Christiano *et al.*, 2003). Second stage inferences can also be sensitive to these choices in both theory (e.g. Elliott, 1998) and in applications, such as predictability tests (e.g. Stambaugh, 1999).

[1](e.g. Dufour and Renault, 1998; Hidalgo, 2000; Hidalgo, 2005; Dufour and Jouini, 2006; Saidi and Roy, 2008; Dufour *et al.*, 2006; McCracken, 2007; Hong *et al.*, 2009, to list just a few)

[2]We address only Granger's version of causality, despite the importance of several other definitions.

Our approach builds on a rich literature, originating in Park and Phillips (1989) and Sims *et al.* (1990), who show that parameters that may be expressed as coefficients on stationary regressors retain a standard root-T normal asymptotic distribution, even in I(1) systems. Similar results hold in cointegrating systems involving nonstationary fractional integration (Dolado and Marmol, 2004). The surplus lag approach uses this result to simplify inference. In the context of unit root testing, Choi (1993) recognized that, with the addition of an extra, unnecessary lag, the autoregressive model could be rewritten so that all the parameters of interest are expressed as coefficients on stationary transformations of the data. Thus, at some cost in terms of efficiency, inference procedures could be simplified via the avoidance of nonstandard distributions. Toda and Yamamoto (1995), and Dolado and Lütkepohl (1996) showed how the same surplus lag approach could be applied to provide inference in finite order vector autoregression, without knowing which components are I(0) and which are I(1). Saikkonen and Lütkepohl (1996) extended these results to infinite order VARs.

By incorporating an exogenously modelled component, we show that, in the context of Granger causality testing, the robustness features of the surplus lag approach can be considerably enhanced to accommodate a richer class of persistent processes for the forcing variable in the VARX framework, including those with long-memory/fractional integration or unmodelled structural breaks. Our results are not dependent on knowledge of the correct lag orders. In all cases, we allow for infinite lag orders under the null hypothesis, approximated by finite order models whose lag lengths increase with sample size. Thus we also build on the literature on reasonable approximability (Berk, 1974; Lewis and Reinsel, 1985; Lütkepohl and Saikkonen, 1997), providing some extensions to allow for exogenous regressors, including those with long-memory.[3]

Because Granger noncausality places no restriction on the coefficients in the equation describing the causal variable, a Granger causality test based on a VAR model can also be re-interpreted in terms of VARX based causality test. Therefore we conjecture that similar robustness results could be established for a causality test based on the surplus VAR methodology proposed by (Toda and Yamamoto, 1995; Dolado and Lütkepohl, 1996), despite the misspecification of the equation describing the causal variable when characterized by long-memory or structural breaks. A formal proof would require extensions of our results to allow the number of tested coefficients to

[3]Some related extensions are provided by Poskitt (2007), who establishes autoregressive approximations to (univariate) non-invertible and stationary long-memory processes.

increase with the lag order for the dependent variable, as required in the VAR setting. Of course, one advantage of the more general VARX based setting, is that the formulation of the alternative hypothesis is no longer tied to the lag order approximation for the dependent variable.

The generality of the surplus lag approach is not without cost. Naturally, the extra unnecessary lag reduces power relative to a correctly specified model. As previous literature reports, the magnitude of these effects varies considerably. Power losses are greatest in cointegration tests, in which consistency against $O(T^{-1})$ alternatives is lost, but can be far more moderate in other cases. When restricted to the I(0)/I(1) context there exist alternative tests, based on error correction (Toda and Phillips, 1993) and fully modified methods (Kitamura and Phillips, 1997)[4] that are arguably as general, but more powerful, than the existing results established for the surplus lag test. However, our results demonstrate that the surplus-lag causality test applies, without adjustment, to a considerably wider range of processes. This argues for its usefulness as a robust complement to tests that are more powerful in more restrictive settings.

A second limitation of our approach is that we allow for long-memory and structural breaks in the forcing processes but not in the intercepts or error processes for the dependent variables. The difficulty of weakening this assumption in the time domain is discussed by Hidalgo (2000; 2005), who provides frequency based non-parametric causality tests, which allow for covariance stationary long-memory in both.

The remainder of the paper is organized as follows. Sections 2, 3, 4, and 5 present the model, large sample results, simulations, and an empirical illustration, respectively. The tables are included at the back of the paper. Proofs, technical lemmas, and details of the numerical analysis are included in the appendix.

2 The model

We consider tests of the null hypothesis that z_{1t} ($k_{z1} \times 1$) does not Granger cause y_t ($k_y \times 1$) after controlling for z_{2t} ($k_{z2} \times 1$).[5] Using the notation $\mathcal{F}_{t,x}$ to denote the information set generated by $\{x_{t-j}, j \geq 0\}$, we test the Granger noncausality condition

$$\mathbb{E}\left[y_t | \mathcal{F}_{t-1,(y,z1,z2)'}\right] = \mathbb{E}\left[y_t | \mathcal{F}_{t-1,(y,z2)'}\right]. \tag{1}$$

[4] Kim and Phillips (2004) extend FM regression, but not causality tests, to fractional cointegration.

[5] While z_{2t} is optional, the results of Dufour and Renault (1998) underline its potential importance.

In practice this hypothesis is often tested by means of parameter restrictions on a joint VAR involving all three variables. However, in order to allow the forcing variable z_{1t} to exhibit long-memory or structural breaks, we instead model it exogenously, allowing for a number of alternative DGPs (see Section 3).[6]

Under the null hypothesis the true joint DGP for $w_t := [y_t', z_{2t}']'$ will be assumed to be approximable by a VAR model, i.e. we assume that

$$w_t = \sum_{j=1}^{\infty} \pi_{wj} w_{t-j} + \varepsilon_t \tag{2}$$

where $(\varepsilon_t)_{t\in\mathbb{Z}}$ is a martingale difference sequence (MDS; for detailed assumptions see Section 3). Our primary interest lies in the process for y_t, which is approximated by

$$y_t = \sum_{j=1}^{p} (\pi_{yj} y_{t-j} + \pi_{z2j} z_{2t-j}) + \varepsilon_{yt,p}. \tag{3}$$

In order to consider linear alternatives to Granger noncausality, we must also include lags of z_{1t} in the empirical specification. Thus, we estimate the VARX model[7]

$$y_t = \sum_{j=1}^{p} (\psi_{yj} y_{t-j} + \psi_{z2j} z_{2t-j}) + \sum_{j=1}^{p_{z1}+1} \psi_{z1j} z_{1t-j} + \varepsilon_{yt,p} \tag{4}$$

and test the joint restriction $\psi_{z1j} = 0$ for $1 \leq j \leq p_{z1}$ using a standard Wald test.

The estimated model includes a surplus lag of the forcing variable, $z_{1t-p_{z1}-1}$, which is not tested. Its role becomes apparent after reparameterizing (4) as

$$y_t = \sum_{j=1}^{p} (\psi_{yj} y_{t-j} + \psi_{z2j} z_{2t-j}) + \sum_{j=1}^{p_{z1}} \psi_{z1j} \left(z_{1t-j} - z_{1t-p_{z1}-1}\right) + \left(\sum_{j=1}^{p_{z1}+1} \psi_{z1j}\right) z_{1t-p_{z1}-1} + \varepsilon_{yt,p}. \tag{5}$$

When z_{1t} is integrated of order less than 1.5 the parameters restricted under the null hypothesis (i.e. ψ_{z1j} for $1 \leq j \leq p_{z1}$) are expressed as the coefficients on the covariance stationary variables $z_{1t-j} - z_{1t-p_{z1}-1}$ (recall that p_{z1} is fixed) and may be shown to follow a joint normal limiting distribution under suitable conditions.

Our analytic results are carried out under the null hypothesis. We will require p to increase with T in order to ensure that (3) approximates (2). In contrast, $p_{z1} \geq 1$ is necessarily over-specified under the null hypothesis in which all lags of z_{1t} are excluded

[6] Note, that although exogenously modelled, z_{t1} is not strictly exogenous in a statistical sense.

[7] When the null hypothesis holds $\psi_{yj} = \pi_{yj}$ and $\psi_{z2j} = \pi_{z2j}$.

from (2). Therefore, we do not require p_{z1} to be either a true value or to grow with T in order to approximate (2). In fact, because our analytic results pertain only to test size, we do not specify a true alternative model, but only an empirical alternative based on an arbitrary, but fixed value of p_{z1}. Of course, the choice of p_{z1} should matter for test power: larger choices of p_{z1} allow more general alternatives, but may reduce power against simpler alternatives. Also, p_{z1} need not be set equal to p, the lag order of w_t. In this way, the VARX provides additional flexibility. Even if modelling z_{1t} requires many lags, e.g. if z_{1t} has long-memory, it may still be possible to model w_t parsimoniously. Likewise, we require only an extra lag of z_{1t} rather than of (z_{1t}, w_t), improving efficiency, particularly when p is small, but the dimension of w_t is large.

In order to rewrite (4) in compact form define $y_t^- := [y'_{t-1}, \ldots, y'_{t-p}]'$, $z_{2t}^- := [z'_{2t-1}, \ldots, z'_{2t-p}]'$, $\psi_y := [\psi_{y1}, \ldots, \psi_{yp}]$, $\psi_{z2} := [\psi_{z21}, \ldots, \psi_{z2p}]$, and $z_{1t}^- := [z'_{1t-1}, \ldots, z'_{1t-p_{z1}}]'$, so that $\varepsilon_{yt,p} = y_t - \psi_y y_t^- - \psi_{z2} z_{2t}^-$. We define by $x_{1t}^- = z_{1t}^-$ the regressors whose coefficients ψ_{x1} are to be tested. The remaining regressors, including the surplus lag, are then grouped together as $x_{2t}^- := [(y_t^-)', (z_{2t}^-)', (z_{1t-p_{z1}-1})']'$. Thus, the estimated equation in (4) may be rewritten in single equation form as

$$y_t = \psi_{x1} x_{1t}^- + \psi_{x2} x_{2t}^- + \varepsilon_{yt,p} \tag{6}$$

where $\psi_{x1} \in \mathbb{R}^{k_y \times p_{z1} k_{z1}}$ and $\psi_{x2} \in \mathbb{R}^{k_y \times (k_y p + k_{z2} p + k_{z1})}$ or in stacked form as

$$Y = X_1 \psi'_{x1} + X_2 \psi'_{x2} + \mathcal{E}_p, \tag{7}$$

where $Y = \left[\begin{array}{ccc} y_{p_{max}+1}^-, & \ldots, & y_T^- \end{array} \right]'$, for $p_{max} = max\{p, p_{z1} + 1\}$, and X_1, X_2, and $\mathcal{E}_p$ stack x_{1t}^-, x_{2t}^- and $\varepsilon_{yt,p}^-$ in identical fashion.

Granger noncausality is imposed by $H_0 : \psi_{x1} = 0$. Defining $X_{1.2} = X_1 - X_2(X'_2 X_2)^{-1} X'_2 X_1$, with rows denoted by $\left(x_{1.2t}^-\right)'$, we estimate ψ_{x1} by $\hat{\psi}_{x1} = Y' X_{1.2} \left(X'_{1.2} X_{1.2}\right)^{-1}$ and the variance of $vec\left(\hat{\psi}_{x1}\right)$ by $\hat{\Sigma}_{x1} := \left((X'_{1.2} X_{1.2})^{-1} \otimes \hat{\Sigma}_\varepsilon\right)$, for $\hat{\Sigma}_\varepsilon := \frac{1}{T} \hat{\mathcal{E}}'_p \hat{\mathcal{E}}_p$, with the rows of $\hat{\mathcal{E}}_p$ given by $\hat{\varepsilon}'_{yt,p}$ for $\hat{\varepsilon}_{yt,p} := y_t - \hat{\psi}_{x1} x_{1t}^- - \hat{\psi}_{x2} x_{2t}^-$.[8] The Wald test takes the form:

$$\hat{W} := \text{vec}(\hat{\psi}_{x1})' \hat{\Sigma}_{x1}^{-1} \text{vec}(\hat{\psi}_{x1}) = \text{vec}(Y' X_{1.2})' \left((X'_{1.2} X_{1.2})^{-1} \otimes \hat{\Sigma}_\varepsilon^{-1}\right) \text{vec}(Y' X_{1.2}). \tag{8}$$

3 Large sample robustness results

In this section we show that the Wald statistic $\hat{W}$ for a test of Granger noncausality in the surplus lag VARX obeys a standard Chi-squared null limiting distribution under

[8]Here $\otimes$ stands for the Kronecker product corresponding to columnwise vectorization.

a variety of assumptions regarding the nature of the persistence in z_{1t}. We first state the assumptions on the innovation process for the endogenous variables w_t in (2):[9]

Assumption N: *The noise* $(\varepsilon_t)_{t\in\mathbb{Z}}$ *is a strictly stationary ergodic martingale difference sequence adapted to the increasing sequence of sigma algebras* $\mathcal{F}_t$ *generated by* $\varepsilon_t, \varepsilon_{t-1}, \ldots$. *Further assume that* $\mathbb{E}\{\varepsilon_t\varepsilon_t'|\mathcal{F}_{t-1}\} = \mathbb{E}\varepsilon_t\varepsilon_t' = \Sigma > 0$, $\mathbb{E}\{\varepsilon_{t,a}\varepsilon_{t,b}\varepsilon_{t,c}|\mathcal{F}_{t-1}\} = \omega_{a,b,c}$ *(constant) where* $\varepsilon_{t,a}$ *denotes the a-th coordinate of* ε_t, *and* $\mathbb{E}\{\varepsilon_{t,i}^4\} < \infty$.

Many of the results presented below may be proved under more general assumptions on the innovations. In particular, finite fourth moments are often unnecessary. The strict stationarity assumption may be seen as overly restrictive in many empirical macroeconomic applications. However, it is not easily relaxed in this setting and the above assumptions are standard in infinite order VAR models (Saikkonen and Lütkepohl, 1996, use similar but stronger assumptions) and provide a single set of assumptions that are sufficient for most of our results. A second restriction is the assumed conditional homoskedasticity of the innovations. If this restriction were dropped robust standard errors would be needed. This is not pursued.

Under the null hypothesis we have $y_t = \varepsilon_{yt,p} + \psi_{x2}x_{2t}^-$ and hence $Y'X_{1.2} = \mathcal{E}_p'X_{1.2}$. This motivates the following high level assumptions where $\hat{\Gamma}_{1.2} := T^{-1}X_{1.2}'X_{1.2}$ is used:

Assumption HL: *Let* $p = p(T)$, *let* p_{z1} *be a fixed integer, and assume that*
(i) $\hat{\Sigma}_\varepsilon \xrightarrow{p} \Sigma$.
(ii) $\hat{\Gamma}_{1.2} \xrightarrow{p} \Gamma_{1.2}$ *for some matrix* $\Gamma_{1.2} \in \mathbb{R}^{k_{z1}p_{z1}\times k_{z1}p_{z1}}, \Gamma_{1.2} > 0$.
(iii) $p(T)$ *is such that* $T^{-1/2}vec(\sum_{t=p+1}^T \varepsilon_{t,p}(x_{1.2t}^-)') \xrightarrow{d} N(0, \Gamma_{1.2}\otimes\Sigma)$.

From Assumption HL the standard asymptotics for $\hat{W}$ are immediate from (8).

Theorem 1 *Let Assumption HL hold for* $\varepsilon_{yt,p} = y_t - \psi_{x2}x_{2t}^- - \psi_{x1}x_{1t}^-$. *Then, under* $H_0 : \psi_{x1} = 0$, $\hat{W} \xrightarrow{d} \chi^2(k_y p_{z1} k_{z1})$.

We show below that in a multitude of circumstances Assumption HL is fulfilled.

3.1 Infinite Order Stationary $VARX$

We first extend the approximation results of Lewis and Reinsel (1985) from the VAR to the VARX model. We employ the following assumptions:[10]

[9]Note that $\mathcal{F}_{t-1,y,z2} = \mathcal{F}_{t-1}$ under the null hypothesis.

[10]We define $||\cdot||_2$ as the Euclidean norm $\|x\|_2 = \sqrt{x'x}$, when applied to the vector x and as the induced matrix norm $max\{\|Ax\|_2 : x(n\times 1), \|x\|_2 = 1\}$ when applied to the $m\times n$ matrix A.

Assumption P1:

(i) The noise $(\varepsilon_t)_{t\in\mathbb{Z}}$ *fulfills Assumption N.*

(ii) $\sum_{j=1}^{\infty}\|\pi_{w,j}\|_2 < \infty$ *and* $\det \pi_w(z) \neq 0, for |z| \leq 1,$ *where* $\pi_w(z) := I - \sum_{j=1}^{\infty}\pi_{w,j}z^j$.

(iii) The integer p *increases with* T *such that* $T^{1/2}\sum_{j=p+1}^{\infty}\|\pi_{w,j}\|_2 \to 0$ *and* $p^3/T \to 0$.

(iv) The process $(z_{1t})_{t\in\mathbb{Z}}$ *is generated according to the equation*

$$z_{1t} = \nu_t + \sum_{j=1}^{\infty}\theta_j\nu_{t-j} + \sum_{j=0}^{\infty}\phi_j\varepsilon_{t-j} \tag{9}$$

where $(\nu_t)_{t\in\mathbb{Z}}$ *fulfills Assumption N with* $\mathbb{E}\nu_t\nu_t' > 0$ *and is independent of the process* $(\varepsilon_t)_{t\in\mathbb{Z}}$*. Here* $\sum_{j=1}^{\infty}\|[\theta_j,\phi_j]\|_2 < \infty$ *is assumed.*

Assumptions (ii) and (iii) match those of Lewis and Reinsel (1985, Theorem 2, p. 398). However, the process $(z_{1t})_{t\in\mathbb{Z}}$ is not modelled endogenously, with the advantage of allowing the lag order p_{z1} for z_{1t} to vary freely, i.e. it is not tied to the approximation properties. Also, over-differenced processes are allowed for z_{1t}, as it does not require a $VAR(\infty)$ representation. The following result extends Theorem 3 of Lewis and Reinsel (1985) to the VARX framework:

Theorem 2 *Let* $x_{2t}^- := [y_{t-1}',\dots,y_{t-p}',z_{2t-1}',\dots,z_{2t-p}',z_{1t-p_{z1}-1}']'$ *and* $x_{1t}^- := [z_{1t-1}',\dots,z_{1t-p_{z1}}']'$. *Then Assumption P1 implies Assumption HL.*

The theorem shows that when the true process follows a VARX(∞,p_{z1}) the Wald test statistic can be used as if the true process was a $VARX(p,p_{z1})$. From the proof it is clear that in this special case the result also holds without the surplus lag $z_{1t-p_{z1}-1}$.

3.2 Infinite order I(1) and near-I(1) models

We next consider the near unit root model (Phillips, 1987; Chan, 1988), which approximates well the case in which the largest roots are indistinguishable from, but still less than, one. This often poses a challenge for inference since, in a local-to-unity model, the critical values of econometric tests designed for the I(0)/I(1) framework typically depend on the value of the localization parameter, which cannot be consistently estimated (see e.g. Elliott (1998)). We will use the following assumptions:

Assumption P2 :

(i) Define $A_{T,w} := I + C_w/T, C_w = diag(c_1,c_2,\dots,c_{k_y+k_{z2}-n})$ *and* $c_i \leq 0$ *for* $i =$

$1, \ldots c_{k_y+k_{z2}-n}$. *There exists a nonsingular matrix* $\Gamma = [\gamma_\perp, \gamma], \gamma \in \mathbb{R}^{(k_y+k_{z2})\times n}, 0 \leq n \leq k_y + k_{z2}$ *such that the process* $(v_t)_{t\in\mathbb{Z}}$ *obtained as (for suitable value* w_0*)*

$$v_t := \Big(\; (\gamma_\perp' w_t - A_{T,w}\gamma_\perp' w_{t-1})', \;\; (\gamma' w_t)' \; \Big)' \tag{10}$$

has an VAR(∞*) representation* $\sum_{j=0}^{\infty} \pi_{v,j} v_{t-j} = \varepsilon_t$ *where* $(\varepsilon_t)_{t\in\mathbb{Z}}$ *fulfills Assumption N.*
(ii) For $\pi_v(z) := \sum_{j=0}^{\infty} \pi_{v,j} z^j$ *we assume* $\det \pi_v(z) \neq 0, |z| \leq 1$.
(iii) Summability of the power series: $\sum_{j=1}^{\infty} j\|\pi_{v,j}\|_2 < \infty$.
(iv) The integer p *increases with* T *such that* $p^3/T \to 0$ *and* $T^{1/2}\sum_{j=p+1}^{\infty} \|\pi_{v,j}\|_2 \to 0$.
(v) Let $A_{T,z} := I + C_z/T$ *where* $C_z := S\,diag(c_{z,1}, \ldots c_{z,k_{z1}})S^{-1}$, $c_{z,i} \leq 0$ *for* $i = 1, \ldots k_{z1}$, *and* $S \in \mathbb{R}^{k_{z1}\times k_{z1}}$ *is nonsingular. The process* $(z_{1t} - A_{T,z} z_{1t-1})_{t\in\mathbb{Z}}$ *for some value* z_{10} *fulfills Assumption P1(iv) where additionally* $\sum_{j=1}^{\infty} j\|[\theta_j, \phi_j]\|_2 < \infty$ *holds.*

Under Assumption P2 y_t, z_{1t} and z_{2t} are all defined as triangular arrays[11] that can be either stationary, integrated, or near-integrated. Cointegrating relations may exist. The matrices of largest roots $A_{T,w}$ and $A_{T,z}$ depend on the matrices of local-to-unity parameters C_w and C_z respectively, allowing for a different local-to-unity parameter (c_i and $c_{z,i}$) in each element of $\gamma_\perp' w_t$ and z_{1t}. The matrix S generalizes the diagonal localization matrix to allow for a rotation of the coordinate system. It is not needed for C_w since $\gamma_\perp$ already allows for a rotation. The component $\gamma' w_t$ is stationary, allowing for cointegration in w_t with cointegration rank n. The no cointegration case ($n = 0$) is also included. Cointegration between w_t and z_{1t} is allowed for, but not explicitly modeled. Results for exact unit roots hold when $c_i = c_{z,i} = 0$.

The theorem below shows that $\hat{W}$ has an asymptotic normal null distribution that is invariant to both the local-to-unity parameters and the cointegrating rank.

Theorem 3 *Let* $x_{2t}^- := [y_{t-1}', \ldots, y_{t-p}', z_{2t-1}', \ldots, z_{2t-p}', z_{1t-p_{z1}-1}']'$ *and* $x_{1t}^- := [z_{1t-1}', \ldots, z_{1t-p_{z1}}']'$. *Then Assumption P2 implies Assumption HL.*

In the special case of exact unit roots ($C_w = 0, C_z = 0$) the theorem extends the robustness results of Saikkonen and Lütkepohl (1996) to the VARX model. The asymptotic normality result in the more general local-to-unity framework is a rare property that underlines the practical value of the surplus lag method as a robust test.

[11]For notational simplicity we follow common practice in suppressing the dependence on T.

3.3 Long-memory forcing variables

Models of fractional integration originating from (Granger and Joyeux, 1980; Hosking, 1981) provide another useful method of spanning the I(0)/I(1) divide. A variable z_{1t} is said to be integrated of order d if its fractional difference $(1-L)^d z_{1t}$ is I(0). Thus values of $0 < d < 1$ provide an intermediate between I(0) and I(1) models, in which shocks do decay, but only at a hyperbolic rate. These slow decay rates have been found useful for modelling a number of phenomena in economics and finance, such as volatilities (Baillie, 1996). For $d < 0.5$, the process fits into a larger class of stationary long-memory models. $d > 0.5$ corresponds to nonstationary fractional integration.

3.3.1 Stationary long-memory

Assumption P1 imposed short-memory via the summability assumptions on the MA(∞) representation of $(z_{1t})_{t\in\mathbb{N}}$. We now relax this condition.

Assumption P4 :
(i) Assumption P1, (i) - (iii) hold. Additionally $(\varepsilon_t)_{t\in\mathbb{Z}}$ *is assumed to be i.i.d.*
(ii) The process $(z_{1t})_{t\in\mathbb{Z}}$ *is generated according to the equation (9), where* $(\nu_t)_{t\in\mathbb{Z}}$ *fulfills Assumption N and is independent of the process* $(\varepsilon_t)_{t\in\mathbb{Z}}$*. Here* $\|[\theta_j,\phi_j]\|_2 \le cj^{d-1}$ *for some constant* $0 < c < \infty$ *and* $-0.5 < d < 0.5$ *is assumed.*
(iii) p *is chosen such that* $p = o(T^{1-2d})$ *and Assumption P1(iii) is fulfilled.*

Since the squared coefficients for $d \approx 0.5, d \le 0.5$ are just summable, the conditions on the impulse response sequences are close to minimal. Thus, the assumptions on the exogenous inputs include many long-memory processes, including fractionally integrated processes and sums of fractionally integrated processes. On the other hand, we now require an additional condition on p, the number of lags included in the approximation for $1/3 < d < 1/2$ since in this case the estimates of the covariance sequence, including the cross covariance with lags of y_t and z_{2t}, are extremely unreliable. In fact, their covariances are of order $O(T^{4d-2})$ and hence arbitrarily small fractions of the sample size are obtained as convergence orders for values close to $d = 0.5$. This in turn limits the range of admitted processes via the assumption that $T^{1/2}\sum_{j=p+1}^{\infty}\|\pi_{w,j}\|_2 \to 0$. In some situations this is not a severe limitation. If the joint process w_t is a VARMA process then any rate of the form $p = T^{\delta}$ will fulfill the approximation restriction and choosing $\delta < 1-2d$ the condition on p is met.

In this setting the advantage of the VARX framework is clearly visible. If instead one modelled the process $[y_t', z_{1t}', z_{2t}']'$ using a VAR(p) then a large p would be required for a small approximation error $\varepsilon_{yt,p} - \varepsilon_{yt}$ due to the slow decay of the coefficients in the true VAR(∞) representation. Again it can be shown that Assumption HL holds. The following result also holds if the surplus lag $z_{1t-p_{z1}-1}$ is omitted.

Theorem 4 *Let* $x_{2t}^- := [y_{t-1}', \ldots, y_{t-p}', z_{2t-1}', \ldots, z_{2t-p}', z_{1t-p_{z1}-1}']'$ *and* $x_{1t}^- := [z_{1t-1}', \ldots, z_{1t-p_{z1}}']'$. *Then Assumption P4 implies Assumption HL.*

3.3.2 Nonstationary long-memory

We next establish that the surplus lag test also retains robustness under the following set of assumptions, which allow for forcing variables with nonstationary long-memory.

Assumption P5 :

(i) Assumption P1, (i) - (iii) hold. Additionally $(\varepsilon_t)_{t\in\mathbb{Z}}$ *is assumed to be i.i.d. and* $\sum_{j=1}^{\infty} j^{1+\delta}\|\pi_{w,j}\| < \infty$ *for some* $\delta > 0$.

(ii) There exists full column rank matrices $\beta \in \mathbb{R}^{k_{z1}\times(k_{z1}-c_{z1})}$ *and* $\beta_\perp \in \mathbb{R}^{k_{z1}\times c_{z1}}$, $\beta'\beta_\perp = 0$ *such that for* $\beta_\perp' z_{10} = 0$

$$\begin{bmatrix} \beta_\perp'(z_{1t} - z_{1t-1}) \\ \beta' z_{1t} \end{bmatrix} = v_t, t \in \mathbb{N}, \text{ where } v_{i,t} = \sum_{j=0}^{\infty} L_i(j) \frac{\Gamma(j + d_i)}{\Gamma(d_i)\Gamma(j+1)} \alpha_i' \begin{pmatrix} \nu_{t-j} \\ \varepsilon_{t-j} \end{pmatrix}, \quad (11)$$

for $-0.5 < d_i < 0.5$, $\|\alpha_i\|_2 = 1$, $\lim_{j\to\infty} L_i(j) = 1$, *and* $(\nu_t)_{t\in\mathbb{Z}}$ *i.i.d. and independent of* ε_t, *with* $\mathbb{E}\nu_t = 0$, $\mathbb{E}\nu_t\nu_t' > 0$ *and finite fourth moments.*

(iii) Defining $d_{\max} := \max(d_1, \ldots, d_{k_{z1}})$, *and* $d_{\min} := \min(d_1, \ldots, d_{c_{z1}})$, *p is chosen such that* $p = o\left(T^{\min\{1/3, 1-2d_{\max}, 1/3(1+2d_{\min})\}}\right)$ *and* $T^{1/2}\sum_{j=p+1}^{\infty}\|\pi_{w,j}\|_2 \to 0$.

Type I Nonstationary fractional integration (see Marinucci and Robinson (1999)) in the forcing variable is allowed for through the hyperbolic rates of decay on $\beta_\perp'(z_{1t} - z_{1t-1})$, through (11), which allows for different values of d in each element of $\beta_\perp' z_{1t}$. The cointegrating residuals, $\beta' z_{1t}$, may be fractionally integrated of order $-0.5 < d_i < 0.5$. The inclusion of the slowly varying coefficients, $L_i(j)$, lends flexibility to the short-memory dynamics, allowing for models such as the ARFIMA(p,d,q) (see Davidson and Hashimzade, 2007). The required restrictions on the increase of p as a function of T are striking. Assumption P4 showed problems for d_i close to 0.5 due to the bad estimates of the covariance sequence. Assumption P5 indicates difficulties for d_i near -0.5, which results from the slow divergence rate of the nonstationary component, with integration $1 + d_i$ only slightly above 0.5. The borderline case $d_i = 0.5$ has not been analyzed. The next theorem shows that when the forcing variables are fractionally integrated of order $0.5 < d < 1.5$ the null asymptotics remain standard.[12]

Theorem 5 *Let* $x_{2t}^{-} := [y_{t-1}', \ldots, y_{t-p}', z_{2t-1}', \ldots, z_{2t-p}', z_{1t-p_{z1}-1}']'$ *and* $x_{1t}^{-} := [z_{1t-1}', \ldots, z_{1t-p_{z1}}']'$. *Then Assumption P5 implies Assumption HL.*

3.4 Structural breaks

We next expand the stationary infinite VARX process to allow for the occurrence of a fixed number (J) of historical breaks in the intercept of the exogenously modelled variable z_{1t}, which occur at fixed fractions of the sample size. Although the true data generating process includes breaks, we do not assume that any breaks are included in the estimated model. In particular, we wish to avoid any first stage inference regarding the existence of and/or number of breaks. Breaks in the process for the endogenously modelled variables w_t would have to be explicitly modelled and thus are not considered. Breaks in the coefficients ψ_{x1} governing the impact of x_{1t} on y_t are also excluded under the null hypothesis, under which these coefficients are fixed at zero.

Assumption P6 :
(i) Assumption P1, (i) - (iii) hold. Additionally $(\varepsilon_t)_{t\in\mathbb{Z}}$ *is assumed to be i.i.d.*
(ii) Let J be a fixed integer denoting the number of breaks. Defining $\omega_0 := 0$ *and letting* ω_j $j = 1, \ldots, J$ *denote the fraction of the sample spent in regime j with* $\sum_{j=1}^{J} \omega_j = 1$, *the process* $(z_{1t})_{t\in\mathbb{Z}}$ *is generated according to the equation*

$$z_{1t} = \sum_{j=1}^{J} \bar{\phi}_j \mathbf{I}\left(1 + \left\lfloor \sum_{k=0}^{j-1} \omega_k T \right\rfloor \leq t \leq \left\lfloor \sum_{k=1}^{j} \omega_k T \right\rfloor\right) + \nu_t + \sum_{j=1}^{\infty} \theta_j \nu_{t-j} + \sum_{j=0}^{\infty} \phi_j \varepsilon_{t-j}$$

where $\mathbf{I}(\cdot)$ *denotes an indicator function* $\lfloor x \rfloor$ *denotes the greatest integer less than x,* $(\nu_t)_{t\in\mathbb{Z}}$ *fulfills Assumption N with* $\mathbb{E}\nu_t\nu_t' > 0$ *and is independent of the process* $(\varepsilon_t)_{t\in\mathbb{Z}}$. *Here* $\sum_{j=1}^{\infty} \|[\theta_j, \phi_j]\| < \infty$ *is assumed.*

The estimated VARX must either include an intercept or z_{1t} must be demeaned prior to estimation. It will be convenient to work with deviations from means. Define

[12] In the special case when the lag length p is known and finite, the validity of the excess lag test may be partially anticipated by the results of Dolado and Marmol (2004) who generalize the findings of Sims *et al.* (1990) to allow for nonstationary fractional integration. However, the above result appears to be the first to directly establish the validity of the surplus lag method with nonstationary fractionally integrated regressors. The allowance for unknown and possibly infinite order models complicates the analysis non-trivially.

$S_j = \left\{ p_{z1} + 2 + \lfloor \sum_{k=0}^{j-1} \omega_k T \rfloor, \ldots, \lfloor \sum_{k=1}^{j} \omega_k T \rfloor \right\}$ as the set of time periods for which all elements of z_{1t}^- belong to regime j. Let $x_t^- := \left[\begin{array}{cc} (x_{1t}^-)' & (x_{2t}^-)' \end{array} \right]'$ denote the full set of regressors and define $\mu(j) := \mathbb{E}\left[x_t^- \mathbf{I}\left(t \in S_j\right) \right]$ and $\bar{\mu} := \sum_{j=1}^{J} \omega_j \mu(j)$ as the mean within regime j and the average mean across regimes, respectively.

Let $\widehat{W}\left(\bar{x}^-\right)$ denote the value of the Wald statistic introduced earlier when the original data x_t^- is replaced by $x_t^- - \bar{x}^-$. We first show that the infeasible estimator $\widehat{W}\left(\bar{\mu}\right)$ has the correct large sample distribution. The result is then easily extended to the feasible statistic $\widehat{W}\left(\bar{x}^-\right)$ in the corollary that follows, the proof of which is omitted.

Theorem 6 *Let* $x_{2t}^- := [y_{t-1}', \ldots, y_{t-p}', z_{2t-1}', \ldots, z_{2t-p}', z_{1t-p_{z1}-1}']'$ *and* $x_{1t}^- := [z_{1t-1}', \ldots, z_{1t-p_{z1}}']'$. *Then Assumption P6 implies Assumption HL is satisfied for* $\widehat{W}\left(\bar{\mu}\right)$.

Corollary Let Assumption P6 hold. Then under $H_0 : \psi_{x1} = 0$, $\widehat{W}\left(\bar{x}^-\right) \xrightarrow{d} \chi^2(k_y p_{z1} k_{z1})$.

4 Simulation results

Below we conduct a small scale simulation. These complement those of Dolado and Lütkepohl (1996) and Swanson *et al.* (2003), who investigate the I(0)/I(1) cases. We consider three methods, the Toda and Phillips (1993) approach, based on a vector error correction model of lag-order $\hat{p}$, with pre-tests for unit roots and cointegration rank (VECM), and two variants of the surplus lag causality test: a VAR($\hat{p}+1$), in which only the first $\hat{p}$ lags are tested (surplus-VAR), and a VARX($\hat{p}$,$\bar{p}_{z1}+1$) in which only the first $\bar{p}_{z1}$ lags of the exogenous component z_{1t} are tested (surplus-VARX).

Table 1 provides a detailed list of the simulation models employed in Tables 2 and 3. In all cases we test the null hypothesis ($H_0 : \delta = 0$) that z_{1t} does not Granger cause y_t against $\delta \neq 0$. Both test size ($\delta = 0$) and size-adjusted power based on critical value adjustments ($\delta \neq 0$) are reported. We consider a broad range of models, including I(0) [DGP 1], I(1) [DGP 2], cointegrated I(1) [CI(1), DGPs 3-4], near I(1)/local-to-unity [NI(1), DGP 5], cointegrated near-I(1) [CNI(1), DGP 6-7], co-structural break [CB, DGP 8], fractionally integrated [I(d), DGPs 9-10], and co-fractionally integrated [CI(d)] models. Models 1-8 are based on specializations of

$$(1 - qI_2L)(y_t, z_{1t})' = \tau_{T,t}(0,1)' + A_T(y_{t-1}, z_{1t-1})' + C(\Delta y_{t-1}, \Delta z_{1t-1})' + u_t, \qquad (12)$$

while the fractionally integrated DGPs (9-12) are based on

$$(1-qL)y_t = A_{1,\cdot}(y_{t-1}, z_{1t-1})' + C_{1,\cdot}(\Delta y_{t-1}, \Delta z_{1t-1})' + u_{1t}, \quad z_{1t} = (1-L)^{-d} u_{2t}. \qquad (13)$$

The error process is specified as

$$u_t' = (u_{1t}, u_{2t}) = \varepsilon_t + B\varepsilon_{t-1}, \quad \varepsilon_t' \sim \text{i.i.d. } N(0,\Sigma), \Sigma_{11} = \Sigma_{22} = 1, \Sigma_{12} = -0.8. \qquad (14)$$

In Table 2 we employ known lag-lengths ($\hat{p} = p = 2$, $\hat{p}_z = p_z = 2$) and white noise errors ($B = I_2$). In Table 3 we allow for infinite values of both p and p_z in the true autoregressive specification, via the vector moving average errors in (14), by setting $B = \begin{bmatrix} -0.3 & 0.5\delta \\ 0 & -0.3 \end{bmatrix}$. In order to preserve test size, the lag-length selection is carried out with the null hypothesis $H_0 : \delta = 0$ imposed. Specifically, for all three tests, $\hat{p}$ is selected by the Akaike (AIC) criterion in an autoregression of y_t alone. The true order of p_{z1} is infinite under $H_A : \delta \neq 0$, but zero under H_0. Therefore it cannot be estimated by AIC with H_0 imposed. Instead, we set it to the same fixed, but now incorrect, value of $\bar{p}_{z1} = 2 \neq p_{z1}$.

Several general findings emerge from Tables 2 and 3. Both surplus lag methods provide fairly reliable test size over the full range of DGPs, both when the lag length p is known and when it is estimated. Likewise, the VECM provides appropriate size in all of the models (DGPs 1-4) for which it was designed, as well as many for which it was not, most notably the fractionally integrated models (DGPs 9-12). However, moderate size distortion is observed in the structural break model (DGP 8) and larger distortions are observed in certain near unit root specifications, particularly DGP 7.

On the other hand, in cases when it has correct size the VECM generally provides better power. The power loss associated with the surplus lag approaches varies considerably across DGP specifications. As anticipated, it can be severe in cases where the Granger causality test corresponds to a cointegration test (e.g. DGPs 3 and 12), whereas it is quite moderate in many of the other cases (e.g. DGPs 1, 5, 9, 10).

It is also interesting to compare the power of the two surplus-lag approaches. The tests differ in two main ways. First, when the number of lags tested in both models are the same ($\bar{p}_{z1} = \hat{p}$), the VARX variant may be expected to have more power since it employs only an extra lag of z_{1t}, whereas the VAR employs an extra lag of $(y_t, z_{1t})'$. This is observed in Table 2, in which the power of the surplus-VARX is always as good as and often much better than that of the surplus-VAR. On the other hand, when $\hat{p} \neq \bar{p}_{z1}$ the two methods test a different number of lags and depending on the form of the alternative this effect can favor either test. In the tests of Table 3, we generally observed $\hat{p} > \bar{p}_{z1}$ and since the true model is infinite this effect tends to favor the surplus VAR. As a result of these two competing effects the power comparison now varies across DGPs, with no clear overall choice between the two.

5 Empirical Illustration

As empirical illustration we test the forward rate unbiasedness hypothesis in a VAR setting. We denote by $s_{i,t}$, $f_{i,t}$, $fp_{i,t} = f_{i,t} - s_{i,t}$, and $r^e_{i,t} = \Delta s_{i,t} - fp_{i,t-1}$ the log spot and forward exchange rates with respect to the US Dollar, the forward premium and the excess return to holding foreign currency, respectively. Here $i = 1, 2$ denote the British Pound (\$/£) and German Mark (\$/DM), respectively. The forward rate unbiasedness hypothesis $\mathbb{E}[s_{i,t}|\mathcal{F}_{t-1}] = f_{i,t-1}$, a risk neutral market efficiency condition, implies that there are no expected excess returns to holding foreign currency: $\mathbb{E}[r^e_{i,t}|\mathcal{F}_{t-1}] = 0$. An immediate implication is that the forward premium does not Granger cause the excess return, i.e. $\mathbb{E}[r^e_{i,t}|\mathcal{F}_{t-1,fp}] = \mathbb{E}[r^e_{i,t}|\mathcal{F}_{t-1,(fp,r^e)'}] = 0$.[13] Rejections of these hypotheses underly the forward premium anomaly, a major puzzle in international finance (see Engel, 1996, for a survey).

The forward premium is highly persistent and there has been debate as to whether it is best modelled via a root near unity (Crowder, 1994), long-memory (Baillie and Bollerslev, 1994; Maynard and Phillips, 2001), or structural breaks (Choi and Zivot, 2007). Consequently, Bekaert and Hodrick (2001) perform a calibrated small sample simulation and report over-rejections of VAR based Wald tests of unbiasedness. Similar concerns have been expressed in a regression based tests of unbiasedness (Baillie and Bollerslev, 2000; Maynard and Phillips, 2001). To ensure reliable inference, a persistence robust test may therefore be required. While robust predictive tests, such as sign tests, have been applied in a simple bivariate regression based tests (e.g. Maynard, 2006), few methods exist for addressing this problem in the VAR-type frameworks. Below we apply the surplus-lag VARX test to address this problem.

Defining $y_t = (r^e_{1,t}, r^e_{2,t})'$ and $z_{1t} = (fp_{1,t}, fp_{2,t})'$, we test the hypothesis that z_{1t} does not Granger cause y_t using 322 end-of-month observations from June, 1973 to March, 2000.[14] We select p by AIC in a VAR of y_t alone (enforcing H_0) and set $\bar{p}_{z1} = 2$ exactly as in Table 3. This yields $\hat{p} = 0$, which precludes a standard VAR based test. However, using values of $p = 1$ and $p = 2$, the causality test based on a standard VAR (without surplus lag) yields p-values of 0.0004 and 0.0054 respectively. This is

[13] Here we use the definition of $\mathcal{F}_{t,x}$ given directly above (1). Section A.2 of the appendix discusses in more detail the relationship the unbiasedness and Granger non-causality hypotheses.

[14] See Maynard (2006) for further details on the data.

strong rejection of unbiasedness. Nonetheless, given the discussion above, its validity could be questioned. Employing the surplus-VARX($\hat{p} = 0, \bar{p}_{z1} + 1 = 3$), in which only $\bar{p}_{z1} = 2$ lags are tested[15] we obtain a larger, but still significant, p-value of 0.0103. This provides a more definitive rejection, whose significance cannot be question based on the persistence of the causal variable.

References

Baillie, R. T (1996). Long memory processes and fractional integration in econometrics. *Journal of Econometrics* **73**, 5–59.

Baillie, R. T and T Bollerslev (1994). The long memory of the forward premium. *Journal of International Money and Finance* **13**(5), 565–571.

Baillie, R. T and T Bollerslev (2000). The forward premium anomaly is not as bad as you think. *Journal of International Money and Finance* **19**, 471–488.

Bekaert, G and R Hodrick (2001). Expectations hypothesis tests. *Journal of Finance* **56**, 1357–1394.

Berk, K. N (1974). Consistent autoregressive spectral estimates. *Ann. Statist.* **2**, 489–502.

Chan, N. H (1988). The parameter inference for nearly nonstationary time series. *Journal of the American Statistical Association* **83**(403), 857–862.

Choi, I (1993). Asymptotic normality of the least-squares estimates for higher order autorgressive integrated processes with some applications. *Econometric Theory* **9**, 263–282.

Choi, K and E Zivot (2007). Long memory and structural changes in the forward discount: An empirical investigation. *Journal of International Money and Finance* **26**, 342–363.

Christiano, L, M Eichenbaum and R. J Vigfusson (2003). What happens after a technology shock?. NBER Working Papers 9819.

Crowder, W. J (1994). Foreign exchange market efficiency and common stochastic trends. *Journal of International Money and Finance* **13**(5), 551–564.

Davidson, J and N Hashimzade (2007). Convergence to stochastic integrals with fractionally integrated processes: Theory, and applications to cointegrating regres-

[15] A simple example in Section A.2 of the appendix illustrates why $p_{z1} = 2$ could be a reasonable choice in this application. It also picked to match the value used in the simulations.

sion. Technical report. University of Exeter.

Diebold, F. X and A Inoue (2001). Long memory and regime switching. *Journal of Econometrics* **105**, 131–159.

Dolado, J and F Marmol (2004). Asymptotic inference results for multivariate long-memory processes. *Econometrics Journal* **7**, 168–190.

Dolado, J and H Lütkepohl (1996). Making Wald tests work for cointegrated VAR systems. *Econometric Reviews* **15**, 369–386.

Dufour, J and E Renault (1998). Short run and long run causality in time series: Theory. *Econometrica* **66**, 1099–1125.

Dufour, J. M and T Jouini (2006). Finite-sample simulation-based inference in VAR models with application to granger causality testing. *Journal of Econometrics* **135**, 229–254.

Dufour, J. M, D Pelletier and E Renault (2006). Short run and long run causality in time series: inference. *Journal of Econometrics* **132**, 337–362.

Elliott, G (1998). On the robustness of cointegration methods when regressors have almost unit roots. *Econometrica* **66**, 149–158.

Engel, C (1996). The forward discount anomaly and the risk premium: A survey of recent evidence. *Journal of Empirical Finance* **3**, 123–191.

Engel, C and K. D West (2005). Exchange rates and fundamentals. *Journal of Political Economy* **113**, 485–517.

Granger, C. W. J (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* **37**, 424–459.

Granger, C. W. J and R Joyeux (1980). An introduction to long memory time series models and fractional differencing. *Journal of Time Series Analysis* **1**, 15–39.

Hidalgo, F. J (2000). Nonparametric test for causality with long-range dependence. *Econometrica* **68**, 1465–1490.

Hidalgo, F. J (2005). A bootstrap causality test for covariance stationary processes. *Journal of Econometrics* **126**, 115–143.

Hong, Y, Y Liu and S Wang (2009). Granger causality in risk and detection of extreme risk spillover between financial markets. *Journal of Econometrics* **150**, 271–287.

Hosking, J. R. M (1981). Fractional differencing. *Biometrika* **68**, 165–176.

Jaeger, D. A and M. D Paserman (2008). The cycle of violence? an empirical analysis of fatalities in the palestinian-israeli conflict. *American Economic Review* **98**, 1591–1604.

Kim, C. S and P. C Phillips (2004). Fully modified estimation of fractional cointegration models. Mimeo.

Kitamura, Y and P. C. B Phillips (1997). Fully modified IV, GIVE and GMM estimation with possibly nonstationary regressors and instruments. *Journal of Econometrics* **80**, 85–123.

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

Lütkepohl, H and P Saikkonen (1997). Impulse response analysis in infinite order cointegrated vector autoregressive processes. *Journal of Econometrics* **81**, 127–157.

Marinucci, D and P. M Robinson (1999). Alternative forms of fractional Brownian motion. *Journal of Statistical Planning and Inference* **80**, 111–122.

Maynard, A (2006). The forward premium anomaly: statistical artifact or economic puzzle? New evidence from robust tests. *Canadian Journal of Economics* **39**(4), 1244–1281.

Maynard, A and P. C. B Phillips (2001). Rethinking an old empirical puzzle: Econometric evidence on the forward discount anomaly. *Journal of Applied Econometrics* **16**(6), 671–708.

McCracken, M. W (2007). Asymptotics for out of sample tests of granger causality. *Journal of Econometrics* **140**, 719–752.

Park, J. Y and P. C Phillips (1989). Statistical inference in regressions with integrated processes: Part II. *Econometric Theory* **5**, 95–131.

Phillips, P. C. B (1987). Towards a unified asymptotic theory for autoregression. *Biometrika* **74**, 535–547.

Phillips, P. C. B (2003). Laws and limits of econometrics. *Economic Journal* **113**, pp. C26–C52.

Poskitt, D. S (2007). Autoregressive approximation in nonstandard situations: The fractionally integrated and non-invertible cases. *Annals of the Institute of Statistical Mathematics* **59**, 697–725.

Saidi, A and R Roy (2008). Robust optimal tests for causality in multivariate time series. *Econometric Theory* **24**, 948–87.

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

Sims, C. A, J. H Stock and M. W Watson (1990). Inference in linear time series models

Table 1: Details of the data generating processes used in Tables 2 and 3

DGPs based on (12)

Model:		q_t	$\tau_{T,t}$	A_T	C	Model:		q_t	$\tau_{T,t}$	A_T	C
1	I(0)	0	0	$H(\delta)$	0	5	NI(1)	1	0	$(c/T)I_2$	$H(\delta)$
2	I(1)	1	0	0	$H(\delta)$	6	CNI(1)	1	0	$G_{1,T}(c)$	$H(0.5\delta)$
3	CI(1)	1	0	$\alpha(1)\beta(0.5\delta)'$	$H(0)$	7	CNI(1)	1	0	$G_{2,T}(0.5\delta, c)$	$H(0)$
4	CI(1)	1	0	$\alpha(0)\beta(1)'$	$H(0.5\delta)$	8	CB	0	$b(\frac{t}{T})$	$H(0.4\delta)$	0

DGPs based on (13)

DGP:		q_t	$A_{1,\cdot}$	$C_{1,\cdot}$	d	DGP:		q_t	$A_{1,\cdot}$	$C_{1,\cdot}$	d
9	I(d)	1	0	$H_{1,\cdot}(\delta)$	0.4	11	CI(d)	1	$-0.5\beta(\delta)'$	0	0.4
10	I(d)	1	0	$H_{1,\cdot}(\delta)$	0.8	12	CI(d)	1	$-0.5\beta(0.5\delta)'$	0	0.8

$c = -5$, $\alpha(\delta)' = (-\delta, 1)$, $\beta(\delta)' = (1, -\delta)$, $b(r) = 2 - 4 \times \mathbf{1}(r \leq 1/2)$

$$H(\delta) = \begin{bmatrix} H_{1,\cdot}(\delta) \\ H_{2,\cdot} \end{bmatrix} = \begin{bmatrix} 0.5 & \delta \\ 0.3 & 0.5 \end{bmatrix}, \; G_{1,T}(c) = \begin{bmatrix} c/T & 0 \\ 1 + c/T & -1 \end{bmatrix}, \; G_{2,T}(\delta, c) = \begin{bmatrix} -1 & (1 + c/T)\delta \\ 0 & c/T \end{bmatrix}$$

with some unit roots. *Econometrica* **58**, 113–144.

Stambaugh, R. F (1999). Predictive regressions. *Journal of Financial Economics* **54**, 375–421.

Swanson, N. R, A Ozyildirim and M Pisu (2003). A comparison of alternative causality and predictive ability tests in the presence of integrated and cointegrated economic variables. In: *Computer Aided Econometrics* (David Giles, Ed.). pp. 91–148. Springer Verlag. New York.

Toda, H. Y and P. C Phillips (1993). Vector autoregressions and causality. *Econometrica* **61**, 1367–1393.

Toda, H. Y and T Yamamoto (1995). Statistical inference in vector autoregressions with possibly integrated processes. *Journal of Econometrics* **66**, 225–250.

Table 2: Null rejection rates and size-adjusted power (fixed lag lengths).

Method	T	$\delta = 0$	0.2	0.4	$\delta = 0$	0.2	0.4	$\delta = 0$	0.2	0.4	$\delta = 0$	0.2	0.4
		DGP 1. I(0)			DGP 2. I(1)			DGP 3. CI(1)			DGP 4. CI(1)		
Toda-	100	0.080	0.230	0.884	0.120	0.948	0.995	0.059	0.554	0.997	0.059	0.272	0.920
Phillips	200	0.071	0.446	0.995	0.075	0.999	1.000	0.050	0.915	1.000	0.047	0.548	0.997
Surplus	100	0.088	0.261	0.840	0.066	0.149	0.470	0.064	0.098	0.258	0.068	0.230	0.741
VAR	200	0.062	0.487	0.989	0.061	0.278	0.820	0.046	0.169	0.485	0.056	0.399	0.966
Surplus	100	0.083	0.259	0.844	0.069	0.234	0.756	0.062	0.173	0.540	0.057	0.264	0.840
VARX	200	0.060	0.544	0.990	0.064	0.450	0.973	0.045	0.377	0.905	0.057	0.488	0.991
		DGP 5. NI(1)			DGP 6. CNI(1)			DGP 7. CNI(1)			DGP 8. CB		
Toda-	100	0.120	0.255	0.889	0.071	0.451	0.986	0.369	0.324	1.000	0.157	0.364	0.921
Phillips	200	0.119	0.475	0.993	0.045	0.921	1.000	0.309	1.000	1.000	0.109	0.783	0.999
Surplus-	100	0.069	0.270	0.887	0.063	0.105	0.261	0.075	0.238	0.805	0.063	0.126	0.284
VAR	200	0.057	0.503	0.994	0.047	0.158	0.484	0.053	0.560	0.997	0.052	0.138	0.435
Surplus-	100	0.074	0.307	0.913	0.059	0.215	0.566	0.082	0.257	0.865	0.058	0.124	0.400
VARX	200	0.055	0.566	0.994	0.043	0.404	0.918	0.056	0.574	0.997	0.055	0.182	0.702
		DGP 9. I(d), d =0.4			DGP 10. I(d), d=0.8			DGP 11. CI(d), d=0.4			DGP 12. CI(d), d=0.8		
Toda-	100	0.108	0.322	0.848	0.073	0.157	0.563	0.072	0.103	0.446	0.100	0.675	0.946
Phillips	200	0.082	0.670	0.993	0.082	0.260	0.836	0.042	0.266	0.889	0.100	0.867	1.000
Surplus-	100	0.090	0.152	0.497	0.081	0.142	0.502	0.080	0.106	0.256	0.076	0.078	0.121
VAR	200	0.065	0.353	0.864	0.069	0.287	0.819	0.060	0.163	0.468	0.060	0.098	0.192
Surplus-	100	0.089	0.308	0.799	0.092	0.184	0.589	0.083	0.112	0.301	0.076	0.087	0.162
VARX	200	0.076	0.635	0.985	0.075	0.407	0.905	0.061	0.200	0.586	0.066	0.115	0.261

Table entries show both empirical rejection rates under the null hypothesis ($\delta = 0$) and size adjusted power under the alternative ($\delta \neq 0$) for a nominal 5% test. The surplus-VAR uses a lag order of 3, but tests only the first 2 lags, and the surplus-lag VARX is based on ARX(2,3), in which only the first 2 lags on the exogenous component are tested. Innovations are drawn from (14) with $B = I_2$. Further details of the DGPs are provided in Table 1. The results are based on 1,000 replications.

Table 3: Null rejection rates and size-adjusted power (AIC).

Method	T	$\delta = 0$	0.2	0.4	$\delta = 0$	0.2	0.4	$\delta = 0$	0.2	0.4	$\delta = 0$	0.2	0.4
		DGP 1. I(0)			DGP 2. I(1)			DGP 3. CI(1)			DGP 4. CI(1)		
Toda-	100	0.049	0.493	1.000	0.091	0.392	1.000	0.116	0.944	0.999	0.025	0.426	0.938
Phillips	200	0.056	0.756	1.000	0.072	0.738	1.000	0.079	0.998	1.000	0.032	0.598	0.999
Surplus	100	0.044	0.319	0.997	0.091	0.374	0.999	0.076	0.408	0.985	0.056	0.274	0.838
VAR	200	0.041	0.666	1.000	0.063	0.758	1.000	0.064	0.891	1.000	0.046	0.436	0.991
Surplus	100	0.056	0.547	1.000	0.072	0.552	1.000	0.070	0.208	0.944	0.031	0.302	0.859
VARX	200	0.050	0.844	1.000	0.053	0.833	1.000	0.057	0.500	0.999	0.033	0.376	0.986
		DGP 5. NI(1)			DGP 6. CNI(1)			DGP 7. CNI(1)			DGP 8. CB		
Toda-	100	0.104	0.252	0.985	0.033	0.379	0.915	0.310	0.570	1.000	0.194	0.899	1.000
Phillips	200	0.101	0.635	1.000	0.027	0.547	0.995	0.289	1.000	1.000	0.081	0.996	1.000
Surplus-	100	0.066	0.328	0.994	0.054	0.291	0.845	0.070	0.634	0.920	0.040	0.198	0.999
VAR	200	0.059	0.732	1.000	0.054	0.402	0.990	0.059	0.970	0.980	0.052	0.587	1.000
Surplus-	100	0.054	0.569	0.999	0.035	0.295	0.847	0.061	0.698	0.994	0.056	0.449	0.881
VARX	200	0.053	0.834	1.000	0.031	0.384	0.985	0.062	0.882	1.000	0.034	0.598	0.999
		DGP 9. I(d), d =0.4			DGP 10. I(d), d=0.8			DGP 11. CI(d), d=0.4			DGP 12. CI(d), d=0.8		
Toda-	100	0.099	0.164	0.508	0.068	0.138	0.460	0.081	0.069	0.294	0.092	0.650	0.928
Phillips	200	0.067	0.338	0.900	0.071	0.205	0.753	0.048	0.242	0.888	0.087	0.878	0.999
Surplus-	100	0.095	0.180	0.561	0.087	0.131	0.424	0.067	0.094	0.195	0.060	0.065	0.093
VAR	200	0.062	0.331	0.879	0.071	0.222	0.773	0.048	0.152	0.420	0.055	0.076	0.182
Surplus-	100	0.071	0.182	0.515	0.074	0.156	0.464	0.055	0.101	0.302	0.060	0.093	0.168
VARX	200	0.060	0.364	0.868	0.052	0.320	0.839	0.044	0.203	0.618	0.050	0.121	0.276

Table entries show both empirical rejection rates under the null hypothesis ($\delta = 0$) and size adjusted power under the alternative ($\delta \neq 0$) for a nominal 5% test. For all three tests the Lag length $\hat{p}$ is selected under the null hypothesis based on an autoregression in y_t alone. In the VARX $\bar{p}_{z1}$ is set equal to 2. Innovations are drawn from (14) with $B = [(-0.3, 0)', (0.5\delta, -0.3)']$. Further details of the DGPs are provided in Table 1. The results are based on 1,000 replications.

A Additional details of the numerical analysis

A.1 Further explanation of the Monte Carlo DGPs

Here we provide further explanation and detail on the DGPs described in Table 1 and equations (12) and (13).

The first four DGPs represent the I(0), I(1) and cointegrated I(1) models. DGP 1 is a stationary VAR(1) in the levels of the data, in which both y_t and z_{1t} are I(0):

$$\begin{bmatrix} y_t \\ z_{1t} \end{bmatrix} = \begin{bmatrix} 0.5 & \delta \\ 0.3 & 0.5 \end{bmatrix} \begin{bmatrix} y_{t-1} \\ z_{1t-1} \end{bmatrix} + \begin{bmatrix} u_{1t} \\ u_{2t} \end{bmatrix}. \qquad \text{[DGP 1]}$$

DGP2 is a difference VAR in which both y_t and z_{1t} are I(1) and there are no cointegrating vectors:[16]

$$\begin{bmatrix} \Delta y_t \\ \Delta z_{1t} \end{bmatrix} = \begin{bmatrix} 0.5 & \delta \\ 0.3 & 0.5 \end{bmatrix} \begin{bmatrix} \Delta y_{t-1} \\ \Delta z_{1t-1} \end{bmatrix} + \begin{bmatrix} u_{1t} \\ u_{2t} \end{bmatrix}. \qquad \text{[DGP 2]}$$

DGPs 3 and 4 are both based on vector error correction models (VECMs), in which y_t and z_{1t} are I(1) and cointegrated. In DGP 3 the causality is due to the presence of cointegration under the alternative

$$\begin{bmatrix} \Delta y_t \\ \Delta z_{1t} \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \end{bmatrix} \begin{bmatrix} 1 & -0.5\delta \end{bmatrix} \begin{bmatrix} y_{t-1} \\ z_{1,t-1} \end{bmatrix} + \begin{bmatrix} 0.5 & 0 \\ 0.3 & 0.5 \end{bmatrix} \begin{bmatrix} \Delta y_{t-1} \\ \Delta z_{1t-1} \end{bmatrix} + \begin{bmatrix} u_{1t} \\ u_{2t} \end{bmatrix}, \qquad \text{[DGP 3]}$$

whereas in DGP 4 the causality instead results from the coefficient on the lagged first differences:

$$\begin{bmatrix} \Delta y_t \\ \Delta z_{1t} \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \begin{bmatrix} 1 & -1 \end{bmatrix} \begin{bmatrix} y_{t-1} \\ z_{1,t-1} \end{bmatrix} + \begin{bmatrix} 0.5 & 0.5\delta \\ 0.3 & 0.5 \end{bmatrix} \begin{bmatrix} \Delta y_{t-1} \\ \Delta z_{1t-1} \end{bmatrix} + \begin{bmatrix} u_{1t} \\ u_{2t} \end{bmatrix}. \qquad \text{[DGP 4]}$$

DGPs 5-7 all represent models with near unit root models, in which we define $c \leq 0$ as the local-to-unity coefficient and $a_T = 1 + c/T$. In DGP 5 (y_t, z_{1t}) are modelled as non-cointegrated near unit roots:

$$\begin{bmatrix} \Delta y_t \\ \Delta z_{1t} \end{bmatrix} = \begin{bmatrix} a_T - 1 & 0 \\ 0 & a_T - 1 \end{bmatrix} \begin{bmatrix} y_{t-1} \\ z_{1t-1} \end{bmatrix} + \begin{bmatrix} 0.5 & \delta \\ 0.3 & 0.5 \end{bmatrix} \begin{bmatrix} \Delta y_{t-1} \\ \Delta z_{1t-1} \end{bmatrix} + \begin{bmatrix} u_{1t} \\ u_{2t} \end{bmatrix}. \qquad \text{[DGP 5]}$$

[16] In order to provide a basis of comparison to the previous literature, we choose the parameters of this model to match a special case of the simulations in Dolado and Lütkepohl (1996).

In DGPs 6-7, we allow for cointegration between near unit roots. In DGP 6:

$$\begin{bmatrix} \Delta y_t \\ \Delta z_{1t} \end{bmatrix} = \begin{bmatrix} a_T - 1 & 0 \\ a_T & -1 \end{bmatrix} \begin{bmatrix} y_{t-1} \\ z_{1t-1} \end{bmatrix} + \begin{bmatrix} 0.5 & 0.5\delta \\ 0.3 & 0.5 \end{bmatrix} \begin{bmatrix} \Delta y_{t-1} \\ \Delta z_{1t-1} \end{bmatrix} + \begin{bmatrix} u_{1t} \\ u_{2t} \end{bmatrix} \qquad \text{[DGP 6]}$$

(y_t, z_{1t}) have cointegrating vector $(1, -1)$ and z_{1t} adjusts to restore long-run equilibrium. In DGP 7, specified by,

$$\begin{bmatrix} \Delta y_t \\ \Delta z_{1t} \end{bmatrix} = \begin{bmatrix} -1 & 0.5a_T\delta \\ 0 & c/T \end{bmatrix} \begin{bmatrix} y_{t-1} \\ z_{1t-1} \end{bmatrix} + \begin{bmatrix} 0.5 & 0 \\ 0.3 & 0.5 \end{bmatrix} \begin{bmatrix} \Delta y_{t-1} \\ \Delta z_{1t-1} \end{bmatrix} + \begin{bmatrix} u_{1t} \\ u_{2t} \end{bmatrix}, \qquad \text{[DGP 7]}$$

it is y_t that performs this adjustment and therefore the alternative hypothesis defines cointegration between y_t and z_{1t}. Both DGP 6 and 7 are specializations of (Elliott 1998, eq. 2).

DGP 8 is a stationary model with a four standard deviation structural break to the intercept for z_{1t} in the middle of the sample. Under the null hypothesis, the break effects only z_{1t}. Under the alternative hypothesis, it also implies a break in the mean value of y_{1t}. Therefore, in this case, the causality test has the interpretation of a test for co-breaking. Specifically it is specified as:

$$\begin{aligned} \begin{bmatrix} y_t \\ z_{1t} \end{bmatrix} &= \tau_{T,t} \begin{bmatrix} 0 \\ 1 \end{bmatrix} + \begin{bmatrix} 0.5 & 0.4\delta \\ 0.3 & 0.5 \end{bmatrix} \begin{bmatrix} y_{t-1} \\ z_{1t-1} \end{bmatrix} + \begin{bmatrix} u_{1t} \\ u_{2t} \end{bmatrix} \\ \tau_{T,t} &= 2 - 4 \times \mathbf{1}(t \leq T/2). \qquad \text{[DGP 8]} \end{aligned}$$

In DGPs 9-12 z_{1t} is fractionally integrated of order d and modelled as:

$$z_{1t} = (1 - L)^{-d} u_{2t} \qquad \text{(A.1)}$$

for $d = 0.4$ and $d = 0.8$. We consider two models. In DGPs 9 and 10 we employ

$$\Delta y_t = 0.5\Delta y_{t-1} + \delta \Delta z_{1t-1} + u_{1t}, \qquad \text{[DGPs 9 and 10]}$$

in which y_t is I(1) under both the null and alternative. Because z_{1t} is I(d) for $d < 1$, y_t and z_{1t} cannot cointegrate even under the alternative hypothesis. We set $d = 0.4$ (stationary case) in DGP 9 and $d = 0.8$ (nonstationary case) in DGP 10. Finally, in DGPs 11-12 we employ

$$\Delta y_t = -0.5(y_{t-1} - b\delta z_{1t-1}) + u_{1t}, \qquad \text{[DGPs 11 and 12]}$$

in which y_t is I(0) under H_0 and I(d) and cointegrated with z_{1t} (with cointegrating vector $(1, -b\delta)$) under the alternative. In DGP 11, $d = 0.4$ and $b = 1$. In DGP 12, $d = 0.8$ and $b = 0.5$.

A.2 Further details relating to the empirical illustration

Here we detail the relationship between the forward rate unbiasedness hypothesis and the Granger causality test employed in the empirical illustration. The forward rate unbiasedness hypothesis, given by[17] $\mathbb{E}[s_t|\mathcal{F}_{t-1}] = f_{t-1}$ also implies that the lagged forward premium ($fp_{t-1} = f_{t-1} - s_{t-1}$) provides an unbiased forecast of the spot return

$$\mathbb{E}[\Delta s_t|\mathcal{F}_{t-1}] = fp_{t-1}. \tag{A.2}$$

In practice, (A.2) is often tested by $H_0 : \beta_1 = 1$ in a regression of

$$s_t - s_{t-1} = \beta_0 + \beta_1 fp_{t-1} + \varepsilon_t \tag{A.3}$$

as in Fama (1984) or by an equivalent restriction in a larger VAR model as in Bekaert and Hodrick (2001), in which lags of Δs_t are included but not tested. For example, in the simplest case, a VAR(1) in $y_t = (\Delta s_t, fp_t)$ would be specified as

$$\begin{bmatrix} \Delta s_t \\ fp_t \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} \Delta s_{t-1} \\ fp_{t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{bmatrix} \tag{A.4}$$

with the hypothesis $H_0 : A_{12} = 1$ tested. These are the most common forms in which the hypothesis has been tested. However, (A.2) is equivalent to the non-predictability of the excess return defined by $r_t^e = \Delta s_t - fp_{t-1}$

$$\mathbb{E}[r_t^e|\mathcal{F}_{t-1}] = 0. \tag{A.5}$$

Likewise (A.3) can be re-expressed as

$$r_t^e = \beta_0 + \alpha_1 fp_{t-1} + \varepsilon_t \tag{A.6}$$

where $\alpha_1 = 1 - \beta_1$ and $H_0 : \alpha_1 = 0$ in (A.6) is equivalent to $H_0 : \beta_1 = 1$ in (A.3). Similarly, we may transform (A.4) into a VAR(2) in $(r_t^e, fp_t)'$:

$$\begin{bmatrix} r_t^e \\ fp_t \end{bmatrix} = \begin{bmatrix} A_{11} & \widetilde{A}_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} r_{t-1}^e \\ fp_{t-1} \end{bmatrix} + \begin{bmatrix} 0 & \widetilde{A}_{12} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} r_{t-2}^e \\ fp_{t-2} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{bmatrix} \tag{A.7}$$

where $\widetilde{A}_{12} = A_{12} - 1$ and in which $H_0 : \widetilde{A}_{12} = 0$ implies that fp_t does not Granger cause r_t^e in a VAR(2). This hypothesis may also be tested more parsimoniously using a VARX(1,2).

[17] Here we use the definition of $\mathcal{F}_{t,x}$ given directly above (1) and denote $\mathcal{F}_{t,(s,f)}$ by $\mathcal{F}_t$.

While this is just a special case, the noncausality restriction is also a theoretical restriction. By the law of iterated expectations, (A.5) implies both

$$\mathbb{E}[r_t^e|\mathcal{F}_{t-1,(fp,r^e)}] = 0, \quad \text{and} \tag{A.8}$$

$$\mathbb{E}[r_t^e|\mathcal{F}_{t-1,fp}] = 0, \tag{A.9}$$

the latter being a condition that is closer to what is actually tested in practice by (A.6). Then a joint implication of (A.8) and (A.9) is that fp_t does not Granger cause r_t^e:

$$\mathbb{E}[r_t^e|\mathcal{F}_{t-1,(fp,r^e)'}] = \mathbb{E}[r_t^e|\mathcal{F}_{t-1,fp}], \tag{A.10}$$

which matches (1) when setting $y_t = r_t^e$ and $z_{1t} = fp_t$ and omitting z_{2t}.

B Technical lemmas

Lemma 1 *Let $w_t = \sum_{j=0}^{\infty} \phi_{w,j}\varepsilon_{t-j}$ where $(\varepsilon_t)_{t\in\mathbb{Z}}$ is an i.i.d. sequence of random variables having zero mean and finite fourth moments. Let $\hat{\gamma}_j := T^{-1}\sum_{t=1+p}^{T} w_t w_{t-j}'$ and $\gamma_j := \mathbb{E}w_t w_{t-j}'$. Assume that $\phi_{w,j} = O(j^{d-1})$ where $-0.5 < d < 0.5$. Then:*

$$\mathbb{E}vec(\hat{\gamma}_j - \mathbb{E}\hat{\gamma}_j)vec(\hat{\gamma}_k - \mathbb{E}\hat{\gamma}_k)' = \begin{cases} O(T^{4d-2}) & , \ \textit{for} \quad 0.25 < d < 0.5 \\ O(T^{-1}\log T) & , \qquad d = 0.25, \\ O(T^{-1}) & , \quad -0.5 < d < 0.25 \end{cases}$$

All $O(.)$ terms hold uniformly in $1 \le j,k \le p$ and $1 \le p \le T$.

Proof: We employ Theorem 1, 3 and 5 of Hosking (1996). However, these results apply to fixed lags, whereas we require expressions uniformly in the lag. First note that using $\Omega := \mathbb{E}\varepsilon_t\varepsilon_t'$ we have for some constant $0 < K < \infty$ not depending on $j \in \mathbb{Z}$ and for $0 < d < 1/2$

$$\|\gamma_j\|_2 = \|\sum_{i=j}^{\infty}\phi_{w,i}\Omega\phi_{w,i-j}'\|_2 \le C\sum_{i=j}^{\infty}\|\phi_{w,i}\|_2\|\phi_{w,i-j}\|_2 \le Kj^{2d-1}$$

since $\|\Omega\|_2 < C, \|\phi_{w,i}\|_2 \le C_k i^{d-1}$ for some $K < \infty$ (using Lemma 3a of Palma and Zevallos (2004) with $b = 1-d$). For $-1/2 < d \le 0$ we obtain $\|\gamma_j\|_2 \le Kj^{\epsilon-1}$ for every $\epsilon > 0$. The vector case is only notationally more complex and hence we only show the result for the case of scalar w_t. Then we obtain

$$\mathbb{E}\hat{\gamma}_j\hat{\gamma}_k = T^{-2}\sum_{t,s=1+p}^{T}\mathbb{E}w_{t+j}w_t w_s w_{s+k}.$$

Note that $\mathbb{E}w_t w_s w_r w_0 = \gamma_{t-s}\gamma_r + \gamma_{t-r}\gamma_s + \gamma_t\gamma_{s-r} + \kappa_4(t,s,r)$ for

$$\kappa_4(t,s,r) := \sum_{a=-\infty}^{\infty} \phi_{w,a+t}\phi_{w,a+s}\phi_{w,a+r}\phi_{w,a}(\mathbb{E}\varepsilon_t^4 - 3(\mathbb{E}\varepsilon_t^2)^2)$$

where for notational simplicity $\phi_{w,a} = 0, a < 0$ is used. It follows that $\mathbb{E}w_0^4 \le M_4 < \infty$ since $\|\phi_{w,a}^4\|_2 = O(a^{4d-4}) = o(a^{-2})$. Next

$$T^{-2}\sum_{t,s=1+p}^{T} \mathbb{E}w_{t+j}w_t w_s w_{s+k} = T^{-2}\sum_{t,s=1+p}^{T} \gamma_j\gamma_k + \gamma_{t-s+j}\gamma_{t-s-k} + \gamma_{t+j-s-k}\gamma_{t-s} + \kappa_4(t-s, t-s+j, k). \tag{A.11}$$

The first term is equal to $(T-p)^2 T^{-2}\gamma_j\gamma_k = \mathbb{E}\hat{\gamma}_j\mathbb{E}\hat{\gamma}_k$ independent of the value of d. The derivation of the bounds for the remaining terms in (A.11) will be done separately for the different cases for d. First consider $0.25 < d < 0.5$. The last term in (A.11) is majorized by the first term in (A.2) of Hosking (1996) and hence can be bounded by $M_{4\epsilon}T^{-1}\gamma_j\gamma_k$ where $M_{4\epsilon}$ is the fourth cumulant of ε_t. In fact this holds for any $d < 0.5$. The two middle terms can be dealt with using $\|\gamma_l\|_2 \le Kl^{2d-1}$ as shown above:

$$\begin{aligned} \left| T^{-2}\sum_{t,s=1+p}^{T} \gamma_{t-s+j}\gamma_{t-s-k} \right| &\le T^{-1}\sum_{l=1-T+p}^{T-1-p} |\gamma_{l+j}\gamma_{l-k}| \frac{T-|l|-p}{T} \\ &\le T^{-1}\left(\sum_{l=1-T+p}^{T-1-p} \gamma_{l+j}^2\right)^{1/2}\left(\sum_{l=1-T+p}^{T-1-p} \gamma_{l-k}^2\right)^{1/2} \end{aligned}$$

and for $j \ge 0$, using Lemma 3.2. (i) of Chan and Palma (1998), we have

$$\sum_{l=1-T+p}^{T-1-p} \gamma_{l+j}^2 \le \sum_{l=1-T+p}^{T-1+2j-p} \gamma_{l+j}^2 = \sum_{l=1-T-j+p}^{T-1+j-p} \gamma_l^2 = O((T-p+j)^{4d-1}) = O(T^{4d-1}).$$

This holds for $0.25 < d < 0.5$ For $d = 0.25$ the same argument shows the bound $O(\log T)$ (cf. Hosking, 1996, top of p. 278). For $j \le 0$ the analogous argument can be used extending the sum to the negative integers. Combining these expressions we obtain $\mathbb{E}\hat{\gamma}_j\hat{\gamma}_k - \mathbb{E}\hat{\gamma}_j\mathbb{E}\hat{\gamma}_k = \Delta_{j,k}$ where $\mathbb{E}|\Delta_{j,k}| \le MT^{4d-2}$ for $0.25 < d < 0.5$.

For $d = 0.25$ the same bound on the last term in (A.11) applies as for $0.25 < d < 0.5$. Further $\mathbb{E}|\Delta_{j,k}| \le M(\log T)/T$ for $d = 0.25$ by standard summability arguments showing that $\sum_{j=1}^{T} j^{-1} = O(T\log T)$ (see e.g. Hosking, 1996, top of p. 278). This shows the claim for $d = 0.25$.

For $d < 0.25$ it follows that the middle two terms are of order $O(T^{-1})$ independent of j, k, p. Hence $\mathbb{E}|\Delta_{j,k}| \le M/T$ for $d < 0.25$. All bounds hold uniformly in $1 \le j, k \le p$

and $1 \leq p \leq T$. $\square$

Inspecting the proof it follows that it also applies (with $d = 0$) to linear processes $v_t = \sum_{j=0}^{\infty} \theta_{v,j}\varepsilon_{t-j}$ where $(\varepsilon_t)_{t\in\mathbb{Z}}$ fulfills Assumption N if $\sum_{j=0}^{\infty} \|\theta_{v,j}\|_2 < \infty$.

Lemma 2 *Let $(\varepsilon_t)_{t\in\mathbb{Z}}$ fulfill Assumption N. Let $v_{t,p} = \sum_{j=0}^{\infty} \phi_{p,j}\varepsilon_{t-j}, t \in \mathbb{Z}, p \in \mathbb{N}$. Then if $\sup_{p\in\mathbb{N}} \sum_{j=0}^{\infty} \|\phi_{p,j}\|_2^2 < \infty$ it follows that $\sup_{p\in\mathbb{N}} \mathbb{E}\|v_{t,p}\|_2^4 < \infty$.*

Proof: The proof for the multivariate case is only notationally more complex, hence only the univariate case will be dealt with. Then $\mathbb{E}v_{t,p}^4 = 3(\mathbb{E}v_{t,p}^2)^2 + \kappa_{4,p}$ (see e.g. the proof of Lemma 1 given above). Next since $\mathbb{E}v_{t,p}^2 = \sum_{j=0}^{\infty} \phi_{p,j}^2 \mathbb{E}\varepsilon_t^2$ it follows that $\sup_{p\in\mathbb{N}} \mathbb{E}v_{t,p}^2 < \infty$. Further

$$\kappa_{4,p} = \sum_{j=0}^{\infty} \phi_{p,j}^4 \mathbb{E}\varepsilon_t^4 \leq \mathbb{E}\varepsilon_t^4 \left(\sum_{j=0}^{\infty} \phi_{p,j}^2 \right)^2 .$$

Hence $\sup_p \kappa_{4,p} < \infty$. $\square$

Lemma 3 *Let Γ denote the Gamma function and let $L_i(j)$ satisfy $lim_{j\to\infty} L_i(j) = 1$ for $i = 1, \ldots, k_u$. Then define v_t by $\Delta v_t = u_t$, $t > 0$ and $v_t = 0$, $t \leq 0$, where $u_{i,t} = \sum_{j=0}^{\infty} \theta_{u,j,i}(\alpha_i' \varepsilon_{t-j})$, $\|\alpha_i\|_2 = 1$, $(\varepsilon_t)_{t\in\mathbb{Z}}$ is i.i.d. with mean zero and finite fourth moments and $\theta_{u,j,i} := \Gamma(d_i)^{-1}(j+1)^{(d_i-1)}L_i(j)$, for $0 < d_i < 1/2$ and $\theta_{u,j,i} := a_{j,i} - a_{j-1,i}$ for $j > 0$ and $\theta_{u,0,i} := a_{0,i}$ for $a_{j,i} := \Gamma(1+d_i)^{-1}(j+1)^{d_i} L_i(j)$ for $-1/2 < d_i < 0$. Further let $w_t = \sum_{j=0}^{\infty} \theta_{w,j}\varepsilon_{t-j}$ for $0 < \| \sum_{j=0}^{\infty} \theta_{w,j}\|_2 < \infty$ and $\theta_{w,j} := O(j^{-1-\delta})$ for $\delta > 0$. Then using $D_T := diag\left(T^{-(d_1+1)}, \ldots, T^{-(d_{k_u}+1)}\right)$ and $D_{T,0} := diag\left(T^{-(d_{1,0}+1)}, \ldots, T^{-(d_{k_u,0}+1)}\right)$, for $d_{i,0} := \max(d_i, 0)$, we have (uniformly in $p = o(T^{1/3})$)*

$$(i) \qquad D_T \sum_{t=p+1}^{T} v_t v_t' D_T \xrightarrow{d} \Xi_d, \quad \textit{where } \det \Xi_d \neq 0 \textit{ a.s.}$$

$$(ii) \qquad \max_{0\leq j\leq H_T} \|D_{T,0} \sum_{t=p+1}^{T} v_t w_{t-j}'\|_2 = O_P(1), \quad \textit{where } H_T = o(T^{1/3})$$

$$(iii) \qquad T^{-(1+\max(d_i+d_j,0))} \sum_{t=p+1}^{T} v_{i,t} u_{j,t}' = O_P(1),$$

$$(iv) \qquad D_T \sum_{t=p+1}^{T} v_{t-1} \varepsilon_t' = O_P(1).$$

Proof: (i), (iii), and (iv) follow from Proposition 4.1 and Theorem 4.1 of Davidson and Hashimzade (2007). For (ii), the convergence in distribution of $T^{-(d_{i,0}+1)} \sum_{t=p+1}^{T} v_{i,t} w_{j,t+1}'$

follows from Theorem 4.1. of Davidson and Hashimzade (2007). The uniform (in j) result can be derived from the following argument:

$$\begin{aligned}T^{-(d_{i,0}+1)}\sum_{t=p+1}^{T} v_{i,t}w'_{t-j} &= T^{-(d_{i,0}+1)}\sum_{t=p+1}^{T}(v_{t,i}-v_{t-j-1,i})w'_{t-j}+T^{-(d_{i,0}+1)}\sum_{t=p+1}^{T} v_{t-j-1,i}w'_{t-j}\\ &= T^{-(d_{i,0}+1)}\sum_{r=0}^{j}\sum_{t=p+1}^{T}\Delta v_{t-r,i}w'_{t-j}+T^{-(d_{i,0}+1)}\sum_{t=p+1}^{T} v_{t-j-1,i}w'_{t-j}\\ &= T^{-d_{i,0}}\sum_{r=0}^{j}\left(T^{-1}\sum_{t=p+1}^{T}u_{t-r,i}w'_{t-j}\right)+T^{-(d_{i,0}+1)}\sum_{t=p+1}^{T} v_{t-j-1,i}w'_{t-j}.\end{aligned}$$

The first term is the sum of $j+1$ estimated covariances to which we apply Lemma 1:

$$\sum_{r=0}^{j}\left(T^{-1}\sum_{t=p+1}^{T}u_{t-r,i}w'_{t-j}\right)=\sum_{r=0}^{j}\mathbb{E}u_{t-r,i}w'_{t-j}+\sum_{r=0}^{j}\left(T^{-1}\sum_{t=p+1}^{T}\left[u_{t-r,i}w'_{t-j}-\mathbb{E}u_{t-r,i}w'_{t-j}\right]\right)+O(pT^{-1})$$

which is of order $O(p^{d_{0,i}})+O_P((j+1)f_T)$ where $f_T=T^{2d_{0,i}-1}$ for $0.25<d_{0,i}<0.5$, $f_T=T^{-1/2}\sqrt{\log T}$ for $d_{0,i}=0.25$ and $f_T=T^{-1/2}$ for $d_{0,i}<0.25$. Here $\sum_{r=1}^{j}\mathbb{E}u_{t-r-1,i}w'_{t-j}=O(p^{d_{0,i}})$ is used which is straightforward to derive. Hence the first term above is of order $o(1)+O_P(jf_TT^{-d_{0,i}})=o_P(1)$ for $d_i>0$ and of order $O(1)+O_P(jT^{-1/2})=O_P(1)$ for $d_i<0$ uniformly in $0\le j\le T^{1/3}$.□

Lemma 4 *Let $v_{t,T}-A_Tv_{t-1,T}=u_t, t\in\mathbb{N}, A_T=I-diag(c_1,\dots,c_k)/T, c_i\ge 0$ for $i=1,\dots k$, where u_t is stationary and ergodic with finite second moments generated according to $\sum_{j=0}^{\infty}\pi_{u,j}u_{t-j}=\varepsilon_t$ where $(\varepsilon_t)_{t\in\mathbb{Z}}$ fulfills Assumption N, and where, for $\pi_u(z):=\sum_{j=0}^{\infty}\pi_{u,j}z^j$, we have $\det\pi_u(z)\neq 0,|z|\le 1$ and $\sum_{j=0}^{\infty}\|\pi_{u,j}\|_2<\infty$. The recursions are started at $v_{0,T}=v_0, T\in\mathbb{N}$ which is assumed to be deterministic. Further let $w_t=\sum_{j=0}^{\infty}\phi^{\varepsilon}_{w,j}\varepsilon_{t-j}+\sum_{j=0}^{\infty}\phi^{\eta}_{w,j}\eta_{t-j}$ where $\sum_{j=0}^{\infty}j\|\phi^{\varepsilon}_{w,j}\|_2<\infty,\sum_{j=0}^{\infty}\|\phi^{\eta}_{w,j}\|_2<\infty$ and $(\eta_t)_{t\in\mathbb{Z}}$ fulfills Assumption N and is independent of $(\varepsilon_t)_{t\in\mathbb{Z}}$. Then:*

(i) $\mathbb{E}\|v_{t,T}\|_2^2=O(t)$ *uniformly in T.*

(ii) $\mathbb{E}\|T^{-3/2}\sum_{t=p+1}^{T}v_{t,T}w'_t\|_2^2=O(T^{-1})$.

(iii) $T^{-2}\sum_{t=p+1}^{T}v_{t,T}v'_{t,T}\xrightarrow{d}\int_0^1 J_c(w)J_c(w)'dw$ *where $J_c(w)$ denotes an Ornstein-Uhlenbeck process.*

(iv) $T^{-1}\sum_{t=p+1}^{T}v_{t,T}u'_t\xrightarrow{d}\int_0^1 J_c(w)dB(w)'+\sigma_u$ *for some matrix σ_u. Here $B(w)$ denotes the Brownian motion associated with $T^{-1/2}u_t$.*

Proof: (i) According to the assumptions it follows that $u_t = \sum_{j=0}^{\infty} \phi_{u,j}\varepsilon_t$ (Lewis and Reinsel, 1985,p. 395, l.3). Further $\sum_{j=-\infty}^{\infty} \|\mathbb{E}u_0 u_j'\|_2 < \infty$ follows. The recursive definition of $v_{t,T}$ implies that $v_{t,T} = A_T^t v_0 + \sum_{i=0}^{t-1} A_T^i u_{t-i}$. Consequently

$$\mathbb{E}\|v_{t,T}\|_2^2 = \mathbb{E}(A_T^t v_0 + \sum_{i=0}^{t-1} A_T^i u_{t-i})'(A_T^t v_0 + \sum_{i=0}^{t-1} A_T^i u_{t-i}) = \mathbb{E}v_0'(A_T^t)' A_T^t v_0 + \sum_{i,j=0}^{t-1} \mathbb{E}u_{t-i}'(A_T^i)' A_T^j u_{t-j}.$$

Since $c_i \geq 0$ for $i = 1, \ldots k$, it follows that the elements of the diagonal matrix A_T are all less than one and hence $v_0(A_T^t)' A_T^t v_0 = O(1)$. For the second term note that

$$|\sum_{i,j=0}^{t-1} \mathbb{E}u_{t-i}'(A_T^i)' A_T^j u_{t-j}| \leq \sum_{i,j=0}^{t-1} \|\mathbb{E}u_{t-i}u_{t-j}'\|_2 \leq t \sum_{j=-\infty}^{\infty} \|\mathbb{E}u_0 u_j'\|_2 = O(t).$$

(ii) We will only deal with the univariate case. The multivariate case is only notationally more difficult. The process $(w_t)_{t\in\mathbb{N}}$ can be decomposed as $w_t := w_t^{\varepsilon} + w_t^{\eta} = (\sum_{j=0}^{\infty} \phi_{w,j}^{\varepsilon} \varepsilon_{t-j}) + (\sum_{j=0}^{\infty} \phi_{w,j}^{\eta} \eta_{t-j})$. Since ε_s and η_t are independent it follows that

$$\mathbb{E}v_{t,T}v_{s,T}w_t w_s = \mathbb{E}v_{t,T}v_{s,T}w_t^{\varepsilon} w_s^{\varepsilon} + \mathbb{E}v_{t,T}v_{s,T}\mathbb{E}w_t^{\eta} w_s^{\eta} \tag{A.12}$$

because $\mathbb{E}v_{t,T}v_{s,T}w_t^{\varepsilon} w_s^{\eta} = \mathbb{E}v_{t,T}v_{s,T}w_t^{\varepsilon}\mathbb{E}w_s^{\eta} = 0$ and expectations exist by Assumption N. We bound the contribution to $\mathbb{E}\|T^{-3/2}\sum_{t=p+1}^{T} v_{t,T}w_t\|_2^2$ of the second term in (A.12) by

$$T^{-3}\sum_{t=1+p}^{T}\sum_{s=1+p}^{T} |\mathbb{E}v_{t,T}v_{s,T}\mathbb{E}w_t^{\eta} w_s^{\eta}| \leq T^{-3}\sum_{t=1+p}^{T}\sum_{s=1+p}^{T} t^{1/2}s^{1/2}|\mathbb{E}w_t^{\eta} w_s^{\eta}| = O(T^{-1})$$

due to $\sum_{j=-\infty}^{\infty} \|\mathbb{E}w_t^{\eta} w_{t-j}^{\eta}\|_2 < \infty$.

For the first term in (A.12), we use the Beveridge-Nelson decomposition (Phillips and Solo, 1992) $w_t^{\varepsilon} = \phi_w(1)\varepsilon_t + w_t^* - w_{t-1}^*$. We then rewrite $\sum_{j=p+1}^{T} v_{t,T}w_t^{\varepsilon}$ as a sum of several terms and show that the expectation of the square of each summand is of the required order. Of course, the cross terms are then of the same order,. It follows that

$$\begin{aligned} T^{-3/2}\sum_{t=1+p}^{T} v_{t,T}w_t^{\varepsilon} &= T^{-3/2}\sum_{t=1+p}^{T} v_{t,T}\varepsilon_t\phi_w(1) + T^{-3/2}\sum_{t=1+p}^{T} v_{t,T}(w_t^* - w_{t-1}^*) \\ &= T^{-3/2}\sum_{t=1+p}^{T} v_{t,T}\varepsilon_t\phi_w(1) - T^{-3/2}\sum_{t=p}^{T-1}(v_{t+1,T} - v_{t,T})w_t^* \\ &\quad + T^{-3/2}v_{T,T}w_T^* - T^{-3/2}v_{p,T}w_p^* \ . \end{aligned} \tag{A.13}$$

Since $v_{T,T} = A_T^T v_0 + \sum_{i=0}^{T-1} A_T^i u_{T-i}$ it follows from finite fourth moments of u_t that $\mathbb{E}v_{T,T}^4 = O(T^4)$ and finite fourth moments of w_T^* (see the proof of Lemma 1) then imply via the Cauchy-Schwartz inequality that $\mathbb{E}v_{T,T}^2(w_T^*)^2 = O(T^2)$. Therefore the two last terms in the expression above contribute terms of the order $O(T^{-1})$ to $\mathbb{E}\|T^{-3/2}\sum_{t=p+1}^{T} v_{t,T}w_t\|_2^2$ as required. Further $v_{t,T} = A_T v_{t-1,T} + u_t$ and

$$\mathbb{E}\left(T^{-3/2}\sum_{t=1+p}^{T} v_{t-1,T}\varepsilon_t\right)^2 = T^{-3}\sum_{t,s=1+p}^{T} \mathbb{E}v_{t-1,T}\varepsilon_t v_{s-1,T}\varepsilon_s = T^{-3}\sum_{t=1+p}^{T} \mathbb{E}v_{t-1,T}^2 \mathbb{E}\varepsilon_t^2 = O(T^{-1})$$

due to $\mathbb{E}\{\varepsilon_t\varepsilon_t'|\mathcal{F}_{t-1}\} = \mathbb{E}\varepsilon_t\varepsilon_t'$ and $\mathbb{E}v_{t,T}^2 = O(t)$. Obviously $\mathbb{E}(T^{-3/2}\sum_{t=p+1}^{T} u_t\varepsilon_t)^2 = O(T^{-1})$. Finally $v_{t,T} - v_{t-1,T} = v_{t,T} - A_T v_{t-1,T} + (A_T - 1)v_{t-1,T} = u_t - c/T v_{t-1,T}$ and therefore the square of the second term in (A.13) equals

$$T^{-3}\sum_{t,s=1+p}^{T} u_{t+1}u_{s+1}w_t^* w_s^* - \frac{c}{T}(v_{t,T}u_{s+1}w_t^* w_s^* + v_{s,T}u_{t+1}w_t^* w_s^*) + \frac{c^2}{T^2}v_{t,T}v_{s,T}w_s^* w_t^*.$$

Now $\mathbb{E}v_{t,T}^4 = O(t^4)$ and hence $\mathbb{E}v_{t,T}u_{s+1}w_t^* w_s^* \leq (\mathbb{E}v_{t,T}^4)^{1/4}(\mathbb{E}u_{s+1}^4)^{1/4}(\mathbb{E}(w_t^*)^4)^{1/2} = O(t)$. Therefore (ii) follows.

The proofs for (iii) and (iv) are omitted since they closely follow previously established results. (iii) and (iv) are proved in Lemma 1 (c) and (d) of (Phillips, 1987) for the univariate case ($k = 1$) and in Lemma 1 (iii) and (iv) of (Elliott, 1998) for the multivariate case, in both cases under different assumptions on the process u_t. The main fact used in both cases, however, is that the process $X_T(t) = T^{-1/2}\sigma^{-1}\sum_{s=1}^{\lfloor tT\rfloor} u_s, 0 \leq t \leq T$ converges weakly to a Brownian motion. It is a standard result that this holds under our assumptions (see e.g. Hall and Heyde, 1980, Theorem 4.1.). □

Lemma 5 *Let the process* $(w_t)_{t\in\mathbb{Z}}$ *be generated according to Assumption P2 (i)-(ii) and be partitioned as* $w_t' = [y_t', z_{2t}']'$. *Accordingly let* ε_{yt} *denote the first block of* $(\Gamma')^{-1}\varepsilon_t$. *Define* $\pi_{w,0,T} := I, \Gamma' := \begin{pmatrix} \gamma_\perp' \\ \gamma' \end{pmatrix}, \pi_{w,j,T} := (\Gamma')^{-1}[\pi_{v,j}\Gamma' - \pi_{v,j-1}\begin{pmatrix} A_{T,w}\gamma_\perp' \\ 0 \end{pmatrix}], j \geq 1$. *Let* $\varepsilon_{yt,p} := \sum_{j=0}^{p-1}[I_s, 0]\pi_{w,j,T}w_{t-j} - [I_s, 0](\Gamma')^{-1}\pi_{v,p-1}\begin{pmatrix} A_{T,w}\gamma_\perp' \\ 0 \end{pmatrix} w_{t-p} = \varepsilon_{yt} - \sum_{j=p}^{\infty}[I_s, 0](\Gamma')^{-1}\pi_{v,j}v_{t-j}$. *Then, for a suitable constant* $c < \infty$ *not depending on* p,

$$\mathbb{E}(\|\varepsilon_{yt,p} - \varepsilon_{yt}\|_2^2)^{1/2} \leq c\sum_{j=p}^{\infty}\|\pi_{v,j}\|_2 \tag{A.14}$$

Proof: Using (10) and the definition of $\pi_{w,j,T}$ to substitute for w_t and $\pi_{w,j,T}$ respectively in the equation for $\varepsilon_{yt,p}$ we obtain $\varepsilon_{t,p} = \sum_{j=0}^{p-1}\pi_{v,j}v_{t-j}$ where $\varepsilon_t = \sum_{j=0}^{\infty}\pi_{v,j}v_{t-j}$. Then (A.14) follows by Lewis and Reinsel (1985), p. 397, (2.9) and $\varepsilon_{yt,p} = [I_s, 0](\Gamma')^{-1}\varepsilon_{t,p}$. □

Remark 1 *The Lemma holds for both the stationary (see Assumption P1) and (co)-integrated I(1) processes as special cases when $\gamma_\perp = 0$ and $c = 0$, respectively.*

Lemma 6 *Let $R_T \in \mathbb{R}^{g_T \times g_T}$ denote a sequence of (possibly random) nonsingular matrices whose dimension g_T depends on the sample size T. Let $\hat{R}_T$ denote a sequence of random matrices such that $\|\hat{R}_T - R_T\|_2 = O_P(b_T)$ where $b_T \to 0$. Then if $\sup_{T\in\mathbb{N}} \|R_T^{-1}\|_2 < \infty$ a.s. it follows that $\|\hat{R}_T^{-1} - R_T^{-1}\|_2 = O_P(b_T)$.*

Proof: See Lewis and Reinsel (1985), p. 397, l. 11. □

Lemma 7

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} A^{-1} & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} -A^{-1}B \\ I \end{bmatrix} \left[D - CA^{-1}B\right]^{-1} \begin{bmatrix} -CA^{-1} & I \end{bmatrix} \tag{A.15}$$

Proof: This can be verified by simple algebraic manipulations. □

Lemma 8 *Under Assumption P1(i), (ii) and (iv) let $\Gamma_p := \mathbb{E}(x_t^-)(x_t^-)'$ where $x_t^- = [(x_{2t}^-)', (x_{1t}^-)']'$ as defined in Theorem 2. Then $\sup_{p\in\mathbb{N}} \|\Gamma_p^{-1}\|_2 < \infty$.*

Proof: Since $z_{1t} = z_{1t}^\nu + z_{1t}^\varepsilon$ where $z_{1t}^\nu = \nu_t + \sum_{j=1}^\infty \theta_j \nu_{t-j}$ and $z_{1t}^\varepsilon = \sum_{j=0}^\infty \phi_j \varepsilon_{t-j}$ are mutually independent, we have $\mathbb{E} z_{1t-i} z_{1t-j}' = \mathbb{E} z_{1t-i}^\nu (z_{1t-j}^\nu)' + \mathbb{E} z_{1t-i}^\varepsilon (z_{1t-j}^\varepsilon)'$. Let x_{1t}^ε and x_{1t}^ν denote the components of x_{1t}^- generated from ε_t and ν_t respectively. Then

$$\begin{aligned}
\Gamma_p &= \mathbb{E} \begin{bmatrix} y_t^-(y_t^-)' & y_t^-(z_{2t}^-)' & y_t^-(z_{1t-p_{z1}-1}^\varepsilon)' & y_t^-(x_{1t}^\varepsilon)' \\ z_{2t}^-(y_t^-)' & z_{2t}^-(z_{2t}^-)' & z_{2t}^-(z_{1t-p_{z1}-1}^\varepsilon)' & z_{2t}^-(x_{1t}^\varepsilon)' \\ z_{1t-p_{z1}-1}^\varepsilon(y_t^-)' & z_{1t-p_{z1}-1}^\varepsilon(z_{2t}^-)' & z_{1t-p_{z1}-1}^\varepsilon(z_{1t-p_{z1}-1}^\varepsilon)' & z_{1t-p_{z1}-1}^\varepsilon(x_{1t}^\varepsilon)' \\ x_{1t}^\varepsilon(y_t^-)' & x_{1t}^\varepsilon(z_{2t}^-)' & x_{1t}^\varepsilon(z_{1t-p_{z1}-1}^\varepsilon)' & x_{1t}^\varepsilon(x_{1t}^\varepsilon)' \end{bmatrix} \\
&\quad + \mathbb{E} \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & z_{1t-p_{z1}-1}^\nu(z_{1t-p_{z1}-1}^\nu)' & z_{1t-p_{z1}-1}^\nu(x_{1t}^\nu)' \\ 0 & 0 & x_{1t}^\nu(z_{1t-p_{z1}-1}^\nu)' & x_{1t}^\nu(x_{1t}^\nu)' \end{bmatrix} \stackrel{def}{=} \Gamma_p^\varepsilon + \Gamma_p^\nu.
\end{aligned}$$

Clearly $0 \leq \Gamma_p^\varepsilon, 0 \leq \Gamma_p^\nu$. Also the largest eigenvalues of both matrices are bounded uniformly in p (see Theorem 6.6.10. of Hannan and Deistler (1988) for Γ_p^ε; the nonzero eigenvalues of Γ_p^ν do not depend on p). Furthermore the matrix in the third and fourth block row and block column of Γ_p^ν is positive definite, since z_{1t} contains the term ν_t. For the heading subblock built from the first and second block row and columns of Γ_p^ε the smallest eigenvalue is bounded uniformly in p by Theorem 6.6.10. on p. 265 of

Hannan and Deistler (1988). Suppose then that the uniform bound on the eigenvalues of Γ_p does not hold. Then there exists a sequence $p_T \to \infty$ and a sequence of unit norm vectors x_p such that $x_p'\Gamma_p x_p \to 0$. Then $x_p'\Gamma_p^{\varepsilon} x_p + x_p'\Gamma_p^{\nu} x_p \to 0$ and hence partitioning $x_p = [x_{p,1}', x_{p,2}', x_{p,3}', x_{p,4}']'$ where $x_{p,i}$ corresponds to the partitioning used previously it follows that $\mathbb{E}(x_{p,3}' z_{1t-p_{z1}-1}^{\nu} + x_{p,4}' x_{1t}^{\nu})(x_{p,3}' z_{1t-p_{z1}-1}^{\nu} + x_{p,4}' x_{1t}^{\nu})' \to 0$. It follows that $\|x_{p,3}\|_2 + \|x_{p,4}\|_2 \to 0$. From Theorem 6.6.10 of Hannan and Deistler (1988) it also follows that $\mathbb{E}(x_{p,1}' y_t^- + x_{p,2}' z_{2t}^-)(x_{p,1}' y_t^- + x_{p,2}' z_{2t}^-)' \to 0$ implies $\|x_{p,1}\|_2 + \|x_{p,2}\|_2 \to 0$. But this produces a contradiction to $\|x\|_2 = 1$. This shows the claim. $\square$

Lemma 9 *Let $(w_t)_{t\in\mathbb{Z}}$, $(\varepsilon_{yt,p})_{t\in\mathbb{Z}}$, and $\pi_{w,j,T}$, $j \geq 0$ be defined as in Lemma 5. Then, under $H_0 : \gamma_{z1j} = 0$ for all j, and for $T > max(c_i)$, (4) can be reformulated as*

$$\Delta y_t = \Psi_{0,p,T}(\gamma_\perp' w_{t-1}) + \sum_{j=1}^{p} \Xi_{j,p,T} v_{t-j} + \left(\sum_{j=1}^{p_{z1}+1} \psi_{z1j}\right) z_{1t-p_{z1}-1} + \sum_{j=1}^{p_{z1}} \psi_{z1j}(z_{1t-j} - z_{1t-p_{z1}-1}) + \varepsilon_{yt,p}, \tag{A.16}$$

where $\sup_{p,T}(\sum_{j=1}^{\infty} \|\Xi_{j,p,T}\|_2) < \infty$, $\Psi_{0,p,T} := -[I:0](\Gamma')^{-1}[I:0]' - \sum_{j=1}^{p-1} \pi_{\perp,j} A_{T,w}^{-(j-1)} - [I:0](\Gamma')^{-1}\pi_{v,p-1}[I:0]' A_T^{2-p}$, *and* $\Xi_{j,p,T} := [\Xi_{1,j,p,T}, \Xi_{2,j,T}]$ *for* $\Xi_{1,j,p,T} := \sum_{h=j+1}^{p-1} \pi_{\perp,h} A_{T,w}^{-(h-j)} + (\Gamma')^{-1}\pi_{v,p-1}[I:0]' A_T^{j-p+1}$ *for* $j = 1, \ldots, p-1$, *and* $\Xi_{1,p,p,T} := 0$, $\Xi_{2,1,T} := -[I:0](I + \pi_{w,1,T})(\Gamma')^{-1}[0:I]'$, $\Xi_{2,j,T} := -[I:0]\pi_{w,j,T}(\Gamma')^{-1}[0:I]'$ *for* $j = 2, \ldots p-1$, $\Xi_{2,p,T} = 0$, *and* $\pi_{\perp,j} := [I:0]\pi_{w,j,T}(\Gamma')^{-1}[I:0]'$.

Remark 2 *A similar reformulation is employed in (A.2) of Saikkonen and Lütkepohl (1996) for the VAR case with $A_{T,w} = I$. However, the derivations and notation differ.*

Proof: Using $[\psi_{yj}, \psi_{z2j}] = -[I,0]\pi_{w,j,T}, j = 1, \ldots, p-1, [\psi_{yp}, \psi_{z2p}] = [I_s, 0](\Gamma')^{-1}\pi_{v,p-1}\left(\gamma_\perp A_{T,w}', 0\right)'$ (since $\gamma_{z1j} = 0$ under H_0) and subtracting $y_{t-1} = [I:0]w_{t-1}$ from both sides of (4) and using $w_t = (\Gamma')^{-1}\Gamma' w_t = (\Gamma')^{-1}((\gamma_\perp' w_t)', v_{2,t}')'$, for $v_{2,t} = [0:I]v_t$, we obtain

$$\Delta y_t = [I:0]\left[-(\Gamma')^{-1}\begin{bmatrix}\gamma_\perp' w_{t-1}\\ v_{2,t-1}\end{bmatrix} - \sum_{j=1}^{p} \pi_{w,j,T}(\Gamma')^{-1}\begin{bmatrix}\gamma_\perp' w_{t-j}\\ v_{2,t-j}\end{bmatrix}\right] + \sum_{j=1}^{p_{z1}+1} \psi_{z1j} z_{1t-j} + \varepsilon_{yt,p}. \tag{A.17}$$

Defining $v_{1,t} := [I:0]v_t = \gamma_\perp' w_t - A_{T,w}\gamma_\perp' w_{t-1}$ and noting that $A_{T,w}$ is invertible for $T > max(c_i)$, the terms involving $\gamma_\perp' w_{t-j}$ in (A.17) can be re-expressed as:

$$\left[-[I:0](\Gamma')^{-1}[I:0]' - \sum_{j=1}^{p} \pi_{\perp,j} A_{T,w}^{-(j-1)}\right] \gamma_\perp' w_{t-1} - \sum_{j=1}^{p-1} \sum_{h=j+1}^{p} \pi_{\perp,h} A_{T,w}^{-(h-j)} v_{1,t-j}.$$

Likewise, the terms involving z_{1t-j} may be re-expressed as in (5), yielding (A.16). Since, by using (10) to substitute for v_j $j = 0, 1, 2 \ldots$ in $\sum_{j=0}^{\infty} \pi_{v,j} v_{t-j} = \varepsilon_t$, $\pi_{w,j,T}$ may be expressed as a linear finite lag function of $\pi_{v,j}$, $\sum_{j=1}^{\infty} j\|\pi_{w,j,T}\| < \infty$ follows by Assumption P2 (iii). $\sup_{p,T}(\sum_{j=1}^{\infty}\|\Xi_{1,j,p,T}\|_2) \leq \sup_T([I:0]\sum_{j=1}^{\infty}\sum_{h=j+1}^{\infty}\|\pi_{w,h,T}\|_2(\Gamma')^{-1}[I:0]') < \infty$ and absolute summability of $\Xi_{2,j}$ both follow. $\square$

C Proof of Theorems

The proof of the theorems will be given based on the following lemma, which introduces a new set of high level conditions sufficient for Assumptions HL to hold:

Lemma 10 *Let $(w_t)_{t\in\mathbb{Z}}$, $(\varepsilon_{yt,p})_{t\in\mathbb{Z}}$, and $\pi_{w,j,T}$, $j \geq 0$ be defined as in Lemma 5. Assume that $z_t^- \in \mathbb{R}^{k_{zp}}$ is a vector, which is $\mathcal{F}_{t-1}$ measurable such that $y_t = A(p)z_t^- + \varepsilon_{yt,p} = [A_1(p), A_2(p), A_3(p)][(z_{t,1}^-)', (z_{t,2,p}^-)', z_{3,t}']' + \varepsilon_{yt,p}$ where $z_t^- \in \mathbb{R}^{k_{zp}}$ is partitioned as $z_t^- = [(z_{1,t}^-)', (z_{2,t,p}^-)', z_{3,t}']'$ such that $z_{t,1}^- = [z_{t-1,1}', \dots, z_{t-p_1,1}']' \in \mathbb{R}^{k_{z1}}$ (where p_1 is fixed) and $z_{3,t} \in \mathbb{R}^{k_{z3}}$ do not depend on p and $z_{2,t,p} = [z_{2t-1}', \dots, z_{2t-p}']'$ depends on p. Further let p tend to infinity as a function of the sample size such that $p^3/T \to 0$ and $T^{1/2}\sum_{j=p+1}^{\infty}\|\pi_{v,j}\|_2 \to 0$ such that $\mathbb{E}(\|\varepsilon_{yt,p} - \varepsilon_{yt}\|_2^2)^{1/2} = o(T^{-1/2})$.*
Then the following conditions are sufficient for Assumption HL to hold: There exists a matrix R_T and a scaling matrix $D_T = diag(I_{k_{z1}}T^{-1/2}, IT^{-1/2}, F_T)$ (where $F_T = diag(f_{t,1}, \dots, f_{tk_{z3}})$) such that ($\lambda_{max}$ denotes a maximal eigenvalue)

$$\sup_{T\in\mathbb{N}} \lambda_{max}(\mathbb{E}R_T) = O(1), \;\; \lambda_{max}(R_T) = O_P(1), \;\; \lambda_{max}(R_T^{-1}) = O_P(1), \tag{A.18}$$

$$R_T = \begin{bmatrix} R_{1,1} & R_{T,1,2} & 0 \\ R_{T,2,1} & R_{T,2,2} & 0 \\ 0 & 0 & R_{T,3,3} \end{bmatrix}, \tag{A.19}$$

$$\hat{R}_T := D_T \sum_{t=p+1}^{T} z_t^-(z_t^-)'D_T, \;\; \text{such that } \|\hat{R}_T - R_T\|_2 = o_P(p^{-1/2}), \;\; \text{and } \mathbb{E}\hat{R}_T = O(1) \;\; \text{elementwise} \tag{A.20}$$

$$\sup_{l\in\mathbb{R}^{k_{zp}},\|l\|_2=1} T^{-1/2}\sum_{t=p+1}^{T}(\mathbb{E}\|l'D_T z_t^-\|_2^2)^{1/2} = O(1), \tag{A.21}$$

$$vec\left[\sum_{t=p+1}^{T}\varepsilon_{yt}(z_t^-)'D_T R_T^{-1}\begin{pmatrix} I & 0 & 0 \end{pmatrix}'\right] \xrightarrow{d} Z, \tag{A.22}$$

where $Z \sim N(0, \Gamma_{1.2}^{-1} \otimes \Sigma)$, where $\Gamma_{1.2} := \lim_{T\to\infty} R_{1,1} - R_{T,1,2}R_{T,2,2}^{-1}R_{T,2,1} > 0$.

Proof: Consider[18]

$$\hat{A}(p) \;:=\; \sum_{t=p+1}^{T} y_t(z_t^-)'(\sum_{t=p+1}^{T} z_t^-(z_t^-)')^{-1} = A(p) + \sum_{t=p+1}^{T} \varepsilon_{yt,p}(z_t^-)'D_T(D_T \sum_{t=p+1}^{T} z_t^-(z_t^-)'D_T)^{-1}D_T$$
$$+ \;\; O(T^{-1}) \;=\; A(p) + \left(\sum_{t=p+1}^{T} \varepsilon_{yt,p}(z_t^-)'D_T \right) \hat{R}_T^{-1} D_T + O(T^{-1}),$$

where $\sum_{t=p+1}^{T} \varepsilon_{yt,p}(z_t^-)'D_T = \sum_{t=p+1}^{T} \varepsilon_{yt}(z_t^-)'D_T + \sum_{t=p+1}^{T} (\varepsilon_{yt,p} - \varepsilon_{yt})(z_t^-)'D_T$ and

$$\mathbb{E}\| \sum_{t=p+1}^{T} (\varepsilon_{yt,p} - \varepsilon_{yt})(z_t^-)'D_T\|_2 \;\leq\; \sum_{t=p+1}^{T} (\mathbb{E}\|\varepsilon_{yt,p} - \varepsilon_{yt}\|_2^2)^{1/2}(\mathbb{E}\|D_T(z_t^-)\|_2^2)^{1/2}$$
$$= \;(T^{1/2} \left(\mathbb{E}\|\varepsilon_{y1,p} - \varepsilon_{y1}\|_2^2\right)^{1/2}) \left(T^{-1/2} \sum_{t=p+1}^{T} (\mathbb{E}\|D_T(z_t^-)\|_2^2)^{1/2} \right) \;=\; o(p^{1/2}). \quad \text{(A.23)}$$

Here (A.21) and Lemma 5 are used. Moreover letting $\varepsilon_{yt(i)}, i = 1, \ldots, k_y$, denote a coordinate of ε_{yt} we have

$$\mathbb{E}(\sum_{t=p+1}^{T} \varepsilon_{yt(i)}(z_t^-)'D_T)'(\sum_{t=p+1}^{T} \varepsilon_{yt(i)}(z_t^-)'D_T) = \sum_{t=p+1}^{T} \mathbb{E}\varepsilon_{yt(i)}^2 \mathbb{E}D_T z_t^-(z_t^-)'D_T = \mathbb{E}\varepsilon_{y1(i)}^2 \mathbb{E}\hat{R}_T$$

using the martingale difference property. Therefore $\| \sum_{t=p+1}^{T} \varepsilon_{yt,p}(z_t^-)'D_T\|_2 = O_P(p^{1/2})$. Consequently, $\|(\hat{A}(p) - A(p))D_T^{-1}\|_2 = O_P(p^{1/2})$ using (A.21), (A.18, A.20) and Lemma 6. Then consider $\hat{\Sigma}_\varepsilon := T^{-1} \sum_{t=p+1}^{T} \hat{\varepsilon}_t \hat{\varepsilon}_t'$: We obtain

$$\begin{aligned}
\hat{\Sigma}_\varepsilon &= \frac{1}{T} \sum_{t=p+1}^{T} (y_t - \hat{A}(p)z_t^-)(y_t - \hat{A}(p)z_t^-)' \\
&= \frac{1}{T} \sum_{t=p+1}^{T} (\varepsilon_{yt,p} - (\hat{A}(p) - A(p))z_t^-)(\varepsilon_{yt,p} - (\hat{A}(p) - A(p))z_t^-)' \\
&= \frac{1}{T} \sum_{t=p+1}^{T} \varepsilon_{yt,p}\varepsilon_{yt,p}' - \frac{1}{T} \sum_{t=p+1}^{T} \varepsilon_{yt,p}(z_t^-)'(\hat{A}(p) - A(p))' - \frac{1}{T} \sum_{t=p+1}^{T} (\hat{A}(p) - A(p))z_t^- \varepsilon_{yt,p}' \\
&\quad + (\hat{A}(p) - A(p)) \left(\frac{1}{T} \sum_{t=p+1}^{T} z_t^-(z_t^-)' \right) (\hat{A}(p) - A(p))' \\
&= \Sigma + o_P(1) + O_P(p/T) = \Sigma + o_P(1).
\end{aligned}$$

[18] The $O(T^{-1})$ term is due to the dependence of $A(p)$ on $A_{T,z}/T, A_{T,w}/T$ in the local-to-unity case, see Lemma 5.

Here the bound follows from $T^{-1}\sum_{t=p+1}^{T}\varepsilon_{yt,p}\varepsilon_{yt,p}' \to \Sigma$, which can be shown using Lemma 5 and the ergodicity of $(\varepsilon_t)_{t\in\mathbb{Z}}$, implying that $T^{-1}\sum_{t=p+1}^{T}\varepsilon_t\varepsilon_t' \to \Sigma$ almost surely. Further $\|(\hat{A}(p)-A(p))D_T^{-1}\|_2 = O_P(p^{1/2})$, $\|\hat{R}_T\|_2 = O_P(1)$ and $\|\sum_{t=p+1}^{T} D_T z_t^- \varepsilon_{yt,p}'\|_2 = O_P(p^{1/2})$ are used. This shows HL (i).

Next, note that $\hat{\Gamma}_{1.2}^{-1}$ equals the (1,1) block of $\hat{R}_T^{-1}$. Then (A.20) and (A.18) imply HL (ii).

With respect to HL (iii) note that $x_{1.2t}^- = [\hat{\Gamma}_{1.2}, 0]D_T^{-1}\hat{R}_T^{-1}D_T z_t^-$. Therefore

$$T^{-1/2}\sum_{t=p+1}^{T}\varepsilon_{yt,p}(x_{1.2t}^-)' = \sum_{t=p+1}^{T}\varepsilon_{yt,p}(z_t^-)'D_T\hat{R}_T^{-1}\begin{bmatrix}\hat{\Gamma}_{1.2}\\0\\0\end{bmatrix} = \sum_{t=p+1}^{T}\varepsilon_{yt}(z_t^-)'D_T R_T^{-1}\begin{bmatrix}\hat{\Gamma}_{1.2}\\0\\0\end{bmatrix} + o_P(1)$$

since $\|\sum_{t=p+1}^{T}(\varepsilon_{yt}-\varepsilon_{yt,p})(z_t^-)'D_T l\|_2 = o_P(1)$ similar to (A.23) and $\sum_{t=p+1}^{T}\varepsilon_{yt}(z_t^-)'D_T = O_P(p^{1/2})$ as used above. Then (A.22) and HL (ii) imply HL (iii). □

C.1 Proof of Theorem 2

Proof: The proof uses a number of results of Lewis and Reinsel (1985), henceforth LR. We verify the conditions of Lemma 10 where $z_{1,t}^- := x_{1t}^-, z_{2,t,p}^- := x_{2t}^-$ and $z_{3,t}$ does not occur. Thus $k_{zp} = k_{z1}p_{z1} + p(k_y + k_{z2}) + k_{z1}$ and $D_T = T^{-1/2}I$. Also, by assumption, all variables are stationary with bounded variance. Then $\mathbb{E}\hat{R}_T = (T-p)/TR_T$. The maximum eigenvalue of R_T is bounded uniformly in $T \in \mathbb{N}$ since z_t^- is a vector containing only lags of the vector process $[w_t', z_{1t}']'$, which has bounded spectrum due to the summability assumptions on the autoregression coefficients (see e.g. Hannan and Deistler, 1988, p. 265). The bound on the minimum eigenvalue of R_T is derived in Lemma 8. This verifies (A.18), (A.19) and (A.21).

Each entry in $\hat{R}_T - R_T$ is equivalent to an estimated covariance at some lag up to an approximation error due to the different limits of summation. Lemma 1 shows that the variance of the estimators of the covariances are of order $O(T^{-1})$, see also Hannan (1976), Chapter 4. The change in the summation introduces an error of order $O_P(pT^{-1})$ since the difference is a sum of a maximum of p terms each of variance $O(T^{-2})$. Thus all entries in $\hat{R}_T - R_T$ are of order $O_P(T^{-1/2})$ and therefore $\|\hat{R}_T - R_T\|_2 = O_P(pT^{-1/2})$. Then $p/T^3 \to 0$ implies that $pT^{-1/2} = o(p^{-1/2})$ showing (A.20).

Finally (A.22) follows as in Theorem 3 of LR (see also Theorem 7.4.9. of Hannan and Deistler, 1988). The only change in the arguments lies in the different definition of the regressors and correspondingly the replacement of Γ_p of LR by R_T. In the proof the uniform bound on $\lambda_{max}(R_T^{-1})$ derived above is crucial. Details are omitted. □

C.2 Proof of Theorem 3

Proof: The proof builds on Saikkonen and Lütkepohl (1996) henceforth SP96. We re-parameterize the auxiliary model (4) using (A.16), which is permissible for our purpose since we test only $\psi_{z1j}, j = 0, \ldots, p_{z1}$, whose estimates coincide in (4) and (A.16).

Note that in (A.16) there are two variables containing nonstationary regressors: $(\gamma_\perp' w_{t-1})$ and $z_{1t-p_{z1}-1}$. Assumption P2 allows for full column rank matrices $\beta \in \mathbb{R}^{(n+k_{z1})\times n_z}$ with $0 \leq n_z \leq n + k_{z1}$ and $\beta_\perp \in \mathbb{R}^{(n+k_{z1})\times(n+k_{z1}-n_z)}$ such that $\beta'\beta_\perp = 0$[19] where $(\tilde{n}_{t,\perp})_{t\in\mathbb{N}}, \tilde{n}_{t,\perp} := \beta'[(\gamma_\perp' w_{t-1})', z_{1t-p_{z1}-1}']'$ is stationary and $(\tilde{n}_t)_{t\in\mathbb{N}}, \tilde{n}_t := \beta_\perp'[(\gamma_\perp' w_{t-1})', z_{1t-p_{z1}-1}']'$ is integrated (but not cointegrated). Thus instead of (A.16), we consider

$$\Delta y_t = [\psi_{x1}, \tilde{\psi}_{x2}, \tilde{\Psi}_0][(z_{1,t}^-)', (z_{2,t,p}^-)', z_{3,t}']' + \varepsilon_{yt,p} = A(p,T)z_t^- + \varepsilon_{yt,p}, \qquad \text{(A.24)}$$

where $z_{1,t}^- := \tilde{x}_{1t}^- = [(z_{1t} - z_{1t-p_{z1}-1})', \ldots, (z_{1t-p_{z1}} - z_{1t-p_{z1}-1})']', z_{2,t,p}^- := [\tilde{n}_{t,\perp}', v_{t-1}', \ldots, v_{t-p+1}', (\gamma' w_{t-p})']'$ and $z_{3,t} := \tilde{n}_t$ analogously to the definition in Lemma A.3. of SP96. Here $(z_{2,t,p}^-)_{t\in\mathbb{Z}}$ is stationary for given value of p. $z_{1,t}^- := [(z_{1t} - z_{1t-p_{z1}-1})', \ldots, (z_{1t-p_{z1}} - z_{1t-p_{z1}-1})']'$ behaves essentially as a stationary process since $z_{1t-j} - A_{T,z}^{p_{z1}+1-j} z_{1t-p_{z1}-1}$ is stationary (as a finite sum of stationary terms) and therefore

$$z_{1t-j} - z_{1t-p_{z1}-1} = z_{1t-j} - A_{T,z}^{p_{z1}-j+1} z_{1t-p_{z1}-1} + (A_{T,z}^{p_{z1}-j+1} - 1)z_{1t-p_{z1}-1},$$

where $A_{T,z}^{p_{z1}-j+1} - 1 = O(T^{-1})$. Thus it follows from Lemma 4 that the second term is negligible and it is sufficient to verify the conditions of Lemma 10.

Define $\hat{R}_T := D_T(\sum_{t=p+1}^T z_t^-(z_t^-)')D_T$ for $D_T := \text{diag}(T^{-1/2}I, T^{-1/2}I, T^{-1}I)$, with partitioning corresponding to that of z_t^- in (A.24). The last k_{z3} coordinates of z_t^- are integrated. The rest are stationary, apart from lower order remainders. Further let

$$R_T := \begin{bmatrix} \mathbb{E}z_{1,t}^-(z_{1,t}^-)' & \mathbb{E}z_{1,t}^-(z_{2,t}^-)' & 0 \\ \mathbb{E}z_{2,t}^-(z_{1,t}^-)' & \mathbb{E}z_{2,t}^-(z_{2,t}^-)' & 0 \\ 0 & 0 & T^{-2}\sum_{t=p+1}^T \tilde{n}_t\tilde{n}_t' \end{bmatrix},$$

such that obviously (A.19) holds. Here the submatrix built of the first two block rows and columns of R_T has uniformly bounded eigenvalues (both from below and from

[19] Cointegration between $\gamma_\perp' w_{t-1}$ and $z_{1t-p_{z1}-1}$ is allowed for, but not imposed. The no cointegration case is accommodated by taking $n_z = 0$.

above) due to Lemma 8 as in the proof of Theorem 2. The nonsingularity (in probability) of the (3,3) block of R_T follows from the convergence in distribution (cf. Lemma 4 (iii)) to an almost sure positive definite random matrix. Therefore $\lambda_{max}(R_T) = O_P(1)$ and $\lambda_{max}(R_T^{-1}) = O_P(1)$ establishing (A.18). $\mathbb{E}\hat{R}_T = O(1)$ is easy to verify from the results of the proof of Theorem 2 and $\mathbb{E}\tilde{n}_t\tilde{n}_t' = O(t)$ from standard theory.
Next, Lemmas 1 (for $d = 0$) and 4 (ii) imply that each entry in $\hat{R}_T - R_T$ has variance uniformly of order $O(T^{-1})$. Thus $\|\hat{R}_T - R_T\|_2 = O_P(p/T^{-1/2})$ showing (A.20) for $p = o(T^{1/3})$. Then consider $\mathbb{E}\|l'D_T z_t^-\|_2^2 = \mathbb{E}(T^{-1}\|l_1' z_{1,t}^-\|_2^2 + T^{-1}\|l_2' z_{2,t}^-\|_2^2 + T^{-2}\|l_3' z_{3,t}\|_2^2)$ where $l' = [l_1', l_2', l_3']$ is partitioned in accordance with z_t^-. By Lemma 4 (i), $\mathbb{E}\|z_{3,t}\|_2^2 = O(t)$. Due to stationarity of the remaining terms $\mathbb{E}\|l'D_T z_t^-\|_2^2 = O(T^{-1})$, analogously to the proof in Theorem 2, and (A.21) follows. Finally, in $\sum_{t=p+1}^{T} \varepsilon_{yt}(z_t^-)' D_T R_T^{-1}[I, 0, 0]'$ the nonstationary terms do not occur due to the block diagonal structure of R_T. Thus analogous arguments as in the proof of Theorem 2 imply that (A.22) holds. □

C.3 Proof of Theorem 4

Proof: The proof follows that of Theorem 2, except that the impulse response sequence corresponding to z_{1t} is not summable. (Note that w_t is short-memory.) Hence let $D_T = T^{-1/2}I$, $R_T = \mathbb{E}z_t^-(z_t^-)'$, and $\hat{R}_T := T^{-1}\sum_{t=p+1}^{T} z_t^-(z_t^-)'$, where z_t^- is defined as in the proof of Theorem 2. To show $\|\hat{R}_T - R_T\|_2 = o_P(p^{-1/2})$, note that every entry in this matrix converges in mean square since, by Lemma 1, the variances are of order $O(T^{\max(4d-2,-1)})$ for $d \neq 0.25$ and of order $O(T^{-1}\log T)$ for $d = 0.25$. Note that $\mathbb{E}\hat{\gamma}_j = (T-p)/T\gamma_j$. Hence $\mathbb{E}\hat{R}_T = (T-p)/TR_T$. Thus the expectation of the sum of squared entries of $\hat{R}_T - R_T$ is of order $O(T^{4d-2}p + p^2T^{-1})$, $O(pT^{-1}\log T + p^2T^{-1})$, and $O(pT^{-1} + p^2T^{-1})$ for $0.25 < d < 0.5$, $d = 0.25$, and $d < 0.25$, respectively. This follows since there are only $O(p)$ terms involving the long-memory processes, as w_t has short memory and contributes p^2 terms of order $O(T^{-1})$. Hence, for obtaining $\|\hat{R}_T - R_T\|_2 = o_P(p^{-1/2})$ it suffices that $p^2T^{4d-2} + p^3T^{-1} \to 0$ for $0.25 < d < 0.5$, $(p^2\log T + p^3)/T \to 0$ for $d = 0.25$, and $p^3T^{-1} \to 0$ otherwise. This shows (A.20). The bounds in (A.18) follow from Lemma 8 (which did not use the short memory assumption on z_{1t}) as in the proof of Theorem 2. Since $z_{3,t}$ does not occur (A.19) follows trivially. Stationarity and finite variances of $(z_{1t})_{t\in\mathbb{N}}$ implies (A.21) as in the proof of Theorem 2.

It remains to verify (A.22). In the following we will only deal with the scalar output case (i.e. $k_y = 1$). The multivariate case is only notationally more difficult. It is sufficient to show that $T^{-1/2}\sum_{t=p+1}^{T}\varepsilon_{yt}(\alpha_p' z_t^-)$ is asymptotically normal with $\alpha_p' R_T \alpha_p \to \alpha_\infty' R_\infty \alpha_\infty$ for vector sequences α_p such that $0 < c < \inf_{p\in\mathbb{N}} \|\alpha_p\|_2 \leq \sup_{p\in\mathbb{N}} \|\alpha_p\|_2 \leq C$ for some constants $0 < c < C < \infty$ and $\|[\alpha_p', 0]' - \alpha_\infty\|_2 \to 0$ holds. Clearly the columns of R_T^{-1} fulfill these requirements.

In this respect we use the three series criterion of Hall and Heyde (1980, Theorem 3.2, p. 58): With $X_{Tt} = \varepsilon_{yt}(\alpha_p' z_t^-)/\sqrt{T}$ we obtain that $(X_{Tt})_{1\leq t\leq T}$ is a martingale difference sequence with respect to the sigma field generated by $\varepsilon_s, \nu_s, s \leq t$. Below we deal only with the univariate case. The multivariate case follows as usual from the Cramer-Wold device (see e.g. Davidson, 1994, Theorem 25.5.). Then Theorem 3.2. states that $\sum_{t=1}^{T} X_{Tt} \xrightarrow{d} \mathcal{N}(0, \eta^2)$ if

$$(i) \max_{1\leq t\leq T} |X_{Tt}| \xrightarrow{p} 0, \quad (ii) \sum_{t=1}^{T} X_{Tt}^2 \xrightarrow{p} \eta^2 (\text{a constant}), \quad (iii)\ \mathbb{E} \max_{1\leq t\leq T} X_{Tt}^2 \text{ is bounded in } T\ .$$

Assume that $\alpha_p' R_T \alpha_p \to \tilde{\eta}^2$ (for some constant $\tilde{\eta}$) as $p \to \infty$. Then it holds that $\mathbb{E}\varepsilon_{yt}^2(\alpha_p' z_t^-)^2 = \mathbb{E}\varepsilon_{yt}^2 \mathbb{E}(\alpha_p z_t^-)^2 < M$ for some constant $0 < M < \infty$ uniformly in $p \in \mathbb{N}$ due to the conditional homoskedasticity and the assumption of finite second moments of z_t^-. Then $\mathbb{E}\max_{1\leq t\leq T} X_{Tt}^2 \leq \sum_{t=1}^{T}\mathbb{E}X_{Tt}^2 \leq M$ such that (iii) follows. Secondly,

$$\sum_{t=1}^{T} X_{Tt}^2 = T^{-1}\sum_{t=1}^{T}\varepsilon_{yt}^2(\alpha_p' z_t^-)^2 = T^{-1}\sum_{t=1}^{T}(\varepsilon_{yt}^2 - \mathbb{E}\varepsilon_{yt}^2)\alpha_p' z_t^-(z_t^-)'\alpha_p + \left(T^{-1}\sum_{t=1}^{T}\alpha_p' z_t^-(z_t^-)'\right)\alpha_p \mathbb{E}\varepsilon_{yt}^2$$

where $\alpha_p'(T^{-1}\sum_{t=1}^{T} z_t^-(z_t^-)')\alpha_p = \alpha_p'\hat{R}_T\alpha_p \to \tilde{\eta}^2$ since $\|\hat{R}_T - R_T\|_2 \to 0$. Therefore it is sufficient to show that $T^{-1}\sum_{t=1}^{T}(\varepsilon_{yt}^2 - \mathbb{E}\varepsilon_{yt}^2)\alpha_p' z_t^-(z_t^-)'\alpha_p$ converges to zero. According to Davidson (1994, Theorem 19.7) this hold for our assumptions if $|(\varepsilon_{yt}^2 - \mathbb{E}\varepsilon_{yt}^2)(\alpha_p' z_t^-)^2|$ can be shown to be uniformly integrable (uniformly over t and p). Now $\mathbb{E}(\varepsilon_{yt}^2 - \mathbb{E}\varepsilon_{yt}^2)^2(\alpha_p' z_t^-)^4 = (\mathbb{E}(\varepsilon_{yt}^2 - (\mathbb{E}\varepsilon_{yt}^2))^2)(\mathbb{E}\alpha_p' z_t^-)^4$ due to the i.i.d. assumption on $(\varepsilon_t)_{t\in\mathbb{Z}}$. But $\mathbb{E}(\varepsilon_{yt}^2 - (\mathbb{E}\varepsilon_{yt}^2))^2 < \infty$ due to finite fourth moments. In order to show that $\sup_{p\in\mathbb{N}}\mathbb{E}(\alpha_p' z_t^-)^4 < \infty$ for $\sup_p \|\alpha_p\|_2 < \infty$ we use Lemma 2: Clearly $\alpha_p' z_t^- = \sum_{j=0}^{\infty}\phi_{p,j}^{\nu}\nu_{t-j} + \sum_{j=0}^{\infty}\phi_{p,j}^{\varepsilon}\varepsilon_{t-j}$. Thus it suffices to show that $\sup_p \sum_{j=0}^{\infty}\|[\phi_{p,j}^{\nu}, \phi_{p,j}^{\varepsilon}]\|_2^2 < \infty$, which follows since $\sup_p \|\alpha_p\|_2$ is bounded by assumption and for each of y_t, z_{1t} and z_{2t} the summability assumption is easily verified. Uniform integrability then follows from Davidson (1994, Theorem 12.10.). It follows that (ii) holds.

Finally (i) holds since it is implied by ($\mathbf{I}(.)$ denoting the indicator function)

$$\sum_{t=1}^{T}\mathbb{E}\left[X_{Tt}^2\mathbf{I}(X_{Tt}^2 > \epsilon)\right] = T\mathbb{E}\left[X_{T1}^2\mathbf{I}(X_{T1}^2 > \epsilon)\right] \to 0$$

for each $\epsilon > 0$ (see Hall and Heyde ,1980, (3.6), p. 53). Here convergence is implied by $\mathbb{E}[\varepsilon_{y1}(\alpha_p' z_1)]^4 = \mathbb{E}\varepsilon_{y1}^4 \mathbb{E}(\alpha_p' z_1)^4 < \infty$ as shown previously. This concludes the proof. $\square$

C.4 Proof of Theorem 5

Proof: The proof of Theorem 5 combines the arguments from the proof of Theorems 3 and 4. Analogously to equation (A.16) we obtain

$$y_t = \sum_{j=1}^{p-1} \pi_j y_{t-j} + \sum_{j=1}^{p} \psi_j z_{2t-j} + \left(\sum_{j=1}^{p_{z1}+1} \psi_{z1j} \right) B^{-1}(B z_{1t-p_{z1}-1}) + \sum_{j=1}^{p_{z1}} \psi_{z1j}(z_{1t-j} - z_{1t-p_{z1}-1}) + \varepsilon_{yt,p}$$

where $B := [\beta, \beta_\perp]$. Note that $z_{1t-j} - z_{1t-p_{z1}-1} = \sum_{i=j}^{p_{z1}} \Delta z_{1t-i} = \sum_{i=j}^{p_{z1}} x_{1t-i}$ is stationary for each $1 \le j < p_{z1}$. Define $z_{1t}^- := \left[z_{1t-1}' - (z_{1t-p_{z1}-1})', \ldots, z_{1t-p_{z1}}' - (z_{1t-p_{z1}-1})' \right]'$, $z_{2,t,p}^- := [(y_t^-)', (z_{2t}^-)', (\beta' z_{1t-p_{z1}-1})']'$ and $z_{3,t} := \beta_\perp' z_{1t-p_{z1}-1}$. Then in $z_t^- := [(z_{1,t}^-)', (z_{2,t,p}^-)', z_{3,t}']'$ the last coordinates (i.e. $z_{3,t}$) are fractionally integrated while the rest are stationary. Let $D_T := \text{diag}\left(T^{-1/2} I, T^{-(d_1+1)}, \ldots, T^{-(d_{c_{z1}}+1)}\right)$, $\hat{R}_T := D_T \sum_{t=p+1}^{T} z_t^- (z_t^-)' D_T$, and

$$R_T := \begin{bmatrix} \mathbb{E} z_{1,t}^-(z_{1,t}^-)' & \mathbb{E} z_{1,t}^-(z_{2,t}^-)' & 0 \\ \mathbb{E} z_{2,t}^-(z_{1,t}^-)' & \mathbb{E} z_{2,t}^-(z_{2,t}^-)' & 0 \\ 0 & 0 & [\hat{R}_T]_{3,3} \end{bmatrix}.$$

Obviously (A.19) holds with this choice. The uniform bound on the eigenvalues of R_T follows as in the proof of Theorem 4 and from

$$\text{diag}\left(T^{-(d_1+1)}, \ldots, T^{-(d_{c_{z1}}+1)}\right) \sum_{t=p+1}^{T} z_{3,t} z_{3,t}' \text{diag}\left(T^{-(d_1+1)}, \ldots, T^{-(d_{c_{z1}}+1)}\right) \xrightarrow{d} \Xi \quad (A.25)$$

where Ξ is a.s. positive definite by Lemma 3 (i). Consequently (A.18) holds.

Next we show that (A.20) also holds. $\hat{R}_T - R_T$ consists of six types of subblocks: The terms involving only $z_{1,t}^-$ and $z_{2,t}^-$ can be analyzed exactly as in the proof of Theorem 4, with $d_{\max} := \max(d_1, \ldots, d_{k_{z1}})$ replacing d. The upper bound on the increase of p as a function of T shows that the sum of squares of these entries is of order $O_P(p^{-1})$. The $(3,3)$ block of $\hat{R}_T - R_T$ is zero by definition. The remaining two terms include terms of the form $T^{-(d_r+3/2)} \sum_{t=p+1}^{T} [z_{3,t}]_r [(\beta' z_{1t-j})']_s = O_p(T^{\max(d_r+d_s,0)-d_r-1/2})$ $T^{-(d_r+3/2)} \sum_{t=p+1}^{T} [z_{3,t}]_r [\Delta z_{1t-j}']_s = O_p(T^{\max(d_r+d_1,\ldots,d_r+d_{cz},0)-d_r-1/2})$ by Lemma 3 (iii). Both terms are $o_p(p^{-1/2})$ since $|d_s|, |d_r| < 0.5$ and, by Assumption P5 (iii), $p < T^{\min_s(1-2d_s,(1+2d_r)/3,1/3)}$ for $r = 1, \ldots, c_{z1}$ and $s = 1, \ldots, k_{z1}$. Likewise, defining $d_{r,0} :=$

$\max(0, d_r)$, it follows from Lemma 3 (ii) that[20]

$$\max_{0\le j\le H_T}\left\|T^{-d_r-3/2}\sum_{t=p+1}^{T}[z_{3,t}]_r[y'_{t-j}, z'_{2t-j}]\right\|_2 = O_P(T^{d_{r,0}-d_r-1/2}), \quad \text{for } H_T = o(T^{1/3}),\ r = 1,\dots,c_{z1}.$$

Thus the sum over these terms is $O_P(pT^{d_{r,0}-d_r-1/2}) = o_p(p^{-1/2})$ since, by Assumption P5 (iii), $p < T^{1/3}$ (covers $0 \le d_r < 1/2$) and $p < T^{2/3(1/2+d_r)}$ (covers $-1/2 < d_r < 0$).

Further $\mathbb{E}[z_{3,t}]_r^2 = O(T^{2d_r+1})$ follows from Davidson and Hashimzade (2007). Thus (A.20) holds under the restrictions on p imposed in Assumption P5. From (A.25) it also follows that the contribution of this block to $\mathbb{E}\|l'D_T z_t^-\|_2^2$ is $O(1)$, showing (A.21). Finally the arguments to show (A.22) are analogous to those used in the proof of Theorem 4 since the nonstationary components are not involved. This concludes the proof. □

C.5 Proof of Theorem 6

Proof: The strategy of the proof is to apply, where possible, the previously proved results within each regime. We will verify the conditions of Lemma 10 where $z_{1,t}^- := x_{1t}^-, z_{2,t,p}^- := x_{2t}^-$ and $z_{3,t}$ does not occur. Thus $k_{zp} = k_{z1}(p_{z1}+1) + p(k_y + k_{z2})$ and $D_T = T^{-1/2}I$. S_j (defined in the main text) omits $p_{z1}+1$ discarded lags, which we denote by $D_j := \left\{\lfloor\sum_{k=0}^{j-1}\omega_k T\rfloor + 1, \dots,\ p_{z1} + 1 + \lfloor\sum_{k=0}^{j-1}\omega_k T\rfloor\right\}$. Let $D := \bigcup_{j=1}^{J} D_j$. Define the within-regime variance $\Gamma(j) := \mathbb{E}\left[z_t^-(z_t^-)'\mathbf{I}(t\in S_j)\right] - \mu(j)\mu(j)'$ and define $R := \sum_{j=1}^{J}\omega_j R(j)$, where $R(j) := \mathbb{E}\left[\left(z_t^- - \bar{\mu}\right)\ \left(z_t^- - \bar{\mu}\right)'\mathbf{I}(t\in S_j)\right]$ as a measure of the overall average variation. Noting that $R(j) = \Gamma(j) + (\mu(j)-\bar{\mu})(\mu(j)-\bar{\mu})'$ we decompose R as $R = \sum_{j=1}^{J}\omega_j\Gamma(j) + \sum_{j=1}^{J}\omega_j(\mu(j)-\bar{\mu})(\mu(j)-\bar{\mu})'$.

Using the same argument as was used directly for R in the proof of Theorem 4 for $\Gamma(i)$ we have $\lambda_{\max}(\Gamma(i)), \lambda_{\max}(\Gamma(i)^{-1}) = O(1)$. We also have $\lambda_{max}\left((\mu(j)-\bar{\mu})(\mu(j)-\bar{\mu})'\right) = O(1)$ despite the fact that the dimension $\mu(j)-\bar{\mu}$ grows in p, since it consists of $p_{z_1}+1$ repeated copies of the same vector extended to the correct dimension by adding zeros. Here p_{z_1} is fixed independently of the sample size. Then it follows (Lütkepohl, 1996, p. 74) that

$$\lambda_{max}(R) \le \sum_{j=1}^{J}\omega_j\lambda_{max}(\Gamma(j)) + \sum_{j=1}^{J}\omega_j\lambda_{max}\left((\mu(j)-\bar{\mu})(\mu(j)-\bar{\mu})'\right) = O(1) \text{ and}$$

$$\lambda_{max}\left(R^{-1}\right) \le \left(\sum_{j=1}^{J}\omega_j\lambda_{min}(\Gamma(j))\right)^{-1} = O(1)$$

[20] The summability condition of Assumption P5 (i) implies the rate condition on $\theta_{w,j}$ in Lemma 3.

where J is fixed. This shows (A.18).

Next define sample counterparts (recall that here $\bar{\mu}$ is treated as known):

$$\begin{aligned}\bar{z}(j) &:= \lfloor \omega_j T \rfloor^{-1} \sum_{t \in S_j} z_t^-, \quad \bar{z} := T^{-1} \sum_{t=p+1}^{T} z_t^-, \quad \hat{\Gamma}(j) := \lfloor \omega_j T \rfloor^{-1} \sum_{t \in S_j} \left(z_t^- - \mu\left(j\right)\right)\left(z_t^- - \mu\left(j\right)\right)', \\ \hat{R}(j) &:= \lfloor \omega_j T \rfloor^{-1} \sum_{t \in S_j} \left(z_t^- - \bar{\mu}\right)\left(z_t^- - \bar{\mu}\right)'\end{aligned}$$

and note that

$$\begin{aligned}\mathbb{E}\left\|\hat{R}(j) - R(j)\right\|_2 &\leq \mathbb{E}\left\|\hat{\Gamma}(j) - \Gamma(j)\right\|_2 + 2\mathbb{E}\left\|\left(\bar{z}(j) - \mu\left(j\right)\right)\right\|_2 \left(\left\|\mu\left(j\right)'\right\|_2 + \left\|\bar{\mu}'\right\|_2\right) \\ &\quad + (p_{z1}+1)\left\|[\omega_j T]^{-1}\bar{\mu}\bar{\mu}'\right\|_2\end{aligned}$$

where the last term results from the sum over the $p_{z1}+1$ discarded lags in D_j.

$$\begin{aligned}\mathbb{E}\left\|\left(\bar{z}(j) - \mu\left(j\right)\right)\mathbf{I}\left(t \in S_j\right)\right\|_2^2 &= (p_{z1}+1)\sum_{i=1}^{k_{z1}} \mathbb{E}\left[\left(\bar{x}_{1ti}(j) - \mathbb{E}\left[x_{1ti}\mathbf{I}\left(t \in S_j\right)\right]\right)^2 \mathbf{I}\left(t \in S_j\right)\right] \\ &+ p\sum_{i=1}^{k_y+k_{z2}} \mathbb{E}\left[\left(\bar{x}_{2ti}(j) - \mathbb{E}\left[x_{2ti}\mathbf{I}\left(t \in S_j\right)\right]\right)^2 \mathbf{I}\left(t \in S_j\right)\right] = O\left(pT^{-1}\right).\end{aligned} \tag{A.26}$$

By similar argument $\|\bar{\mu}\|_2$, $\|\mu(j)\|_2 = O_P(1)$ and $\|p_{z1}[\omega_j T]^{-1}\bar{\mu}\bar{\mu}'\|_2 \leq O_P\left(pT^{-1}\right)$. Thus

$$\left\|\hat{R}(j) - R(j)\right\|_2 \leq \left\|\hat{\Gamma}(j) - \Gamma(j)\right\|_2 + O_P\left(\sqrt{pT^{-1}}\right). \tag{A.27}$$

Next define $\hat{R} := T^{-1}\sum_{t=p+1}^{T}\left(z_t^- - \bar{\mu}\right)\left(z_t^- - \bar{\mu}\right)'$ and note that

$$\begin{aligned}\hat{R} = &\sum_{j=1}^{J} \omega_j \hat{R}(j) + \sum_{j=1}^{J} T^{-1} \sum_{t \in D_j} \left(z_t^- - \bar{\mu}\right)\left(z_t^- - \bar{\mu}\right)' \\ = &\sum_{j=1}^{J} \omega_j \hat{R}(j) + O_P\left(\sqrt{pT^{-1}}\right),\end{aligned} \tag{A.28}$$

since $(p_{z1}+1)T^{-1}\mathbb{E}\left\|\left(z_t^- - \bar{\mu}\right)\left(z_t^- - \bar{\mu}\right)'\right\|_2 \leq (p_{z1}+1)T^{-1}\mathbb{E}\left[\left\|\left(z_t^- - \bar{\mu}\right)\right\|_2^2\right] = O\left(pT^{-1}\right)$, where the last step follows by an argument similar to (A.26). Then by (A.27) and (A.28)

$$\left\|\hat{R} - R\right\|_2 \leq \sum_{j=1}^{J} \omega_j \left\|\hat{\Gamma}(j) - \Gamma(j)\right\|_2 + O_P\left(\sqrt{pT^{-1}}\right). \tag{A.29}$$

The same arguments as in the proofs of Theorems 2 and 4 show $\left\|\hat{\Gamma}(j) - \Gamma(j)\right\|_2 = o_P\left(p^{-1/2}\right)$ since these do not involve breaks. The condition $\mathbb{E}\hat{R}_T = O(1)$ follows from arguments analogous to those employed in the previous proofs above. This shows (A.20).

Next write

$$T^{-1}\sum_{t=p+1}^{T}\left(\mathbb{E}\left\|l'\left(z_t^- - \bar{\mu}\right)\right\|_2^2\right)^{1/2} \leq \sqrt{2}\sum_{j=1}^{J}T^{-1}\sum_{t\in S_j}\left(\mathbb{E}\left\|l'\left(z_t^- - \mu(j)\right)\right\|_2^2\right)^{1/2}$$
$$\sqrt{2}\sum_{j=1}^{J}T^{-1}\sum_{t\in S_j}\left(\left\|l'\left(\mu(j) - \bar{\mu}\right)\right\|_2^2\right)^{1/2} + \sum_{j=1}^{J}T^{-1}\sum_{t\in D_j}\left(\mathbb{E}\left\|l'\left(z_t^- - \bar{\mu}\right)\right\|_2^2\right)^{1/2}. \quad \text{(A.30)}$$

For the last term in (A.30) we have $\left(\mathbb{E}\left\|l'\left(z_t^- - \bar{\mu}\right)\right\|_2^2\right)^{1/2} = O_P(p)$ by arguments similar to those directly above (A.26). It follows that $\sum_{j=1}^{J}T^{-1}\sum_{t\in D_j}\left(\mathbb{E}\left\|l'\left(z_t^- - \bar{\mu}\right)\right\|_2^2\right)^{1/2} = O\left(pT^{-1}\right) = o\left(1\right)$. For the middle term in (A.30) we have $\sum_{j=1}^{J}T^{-1}\sum_{t\in S_j}\left(\left\|l'\left(\mu(j) - \bar{\mu}\right)\right\|_2^2\right)^{1/2} = \sum_{j=1}^{J}\lfloor(T\omega_j - p_{z1} - 1)\rfloor/T\left\|l'\left(\mu(j) - \bar{\mu}\right)\right\|_2 = O\left(1\right)$ by argument similar to (A.26). Finally the first term in (A.30) is also $O(1)$ since J is fixed and $T^{-1}\sum_{t\in S_j}\left(\mathbb{E}\left\|l'\left(z_t^- - \mu(j)\right)\right\|_2^2\right)^{1/2} = O\left(1\right)$ by the same arguments as in the proofs of theorems 2 and 4. This establishes (A.21).

As in the proof of Theorem 4, we will show that $T^{-1/2}\sum_{t=1}^{T}\varepsilon_{yt}\alpha_p'\left(z_t^- - \bar{\mu}\right)$ converges to the normal distribution given in (A.22) by verifying the three conditions of (Hall and Heyde, 1980, Theorem 3.2, p. 58) for $X_{Tt} := \varepsilon_{yt}\alpha_p'\left(z_t^- - \bar{\mu}\right)/\sqrt{T}$ in the scalar case. The multivariate case again follows from the Cramer-Wold device.

Condition (ii) of Hall and Heyde (1980, Theorem 3.2, p. 58) follows from

$$\mathbb{E}\max_{1\leq t\leq T}X_{Tt}^2 \leq \sum_{t=1}^{T}\mathbb{E}X_{Tt}^2 = \mathbb{E}\left[\varepsilon_{yt}^2\right]\alpha_p'\sum_{j=1}^{J}\omega_j\mathbb{E}\left[\left(z_t^- - \bar{\mu}\right)\left(z_t^- - \bar{\mu}\right)'\right]\alpha_p = \mathbb{E}\left[\varepsilon_{yt}^2\right]\alpha_p'R\alpha_p.$$

For condition (ii)

$$\sum_{t=1}^{T}X_{Tt}^2 = \mathbb{E}\left[\varepsilon_{ty}^2\right]\alpha_p'\hat{R}'\alpha_p + T^{-1}\sum_{t=1}^{T}\left(\varepsilon_{ty}^2 - \mathbb{E}\left[\varepsilon_{ty}^2\right]\right)\alpha_p'\left(z_t^- - \bar{\mu}\right)\left(z_t^- - \bar{\mu}\right)'\alpha_p \quad \text{(A.31)}$$

note that $\left\|\hat{\Gamma}\left(j\right) - \Gamma(j)\right\|_2 \to_p 0$ by the same arguments as in Theorems 2 and 4. Therefore by (A.29), this implies that $\left\|\hat{R} - R\right\|_2 \to_p 0$, so that the first term in (A.31) converges in probability to $\tilde{\eta}^2 = \mathbb{E}\left[\varepsilon_{ty}^2\right]\alpha_p'R\alpha_p$. The second term in (A.31) converges in probability to zero by the same arguments as in the proof of Theorem 4 (Lemma 2 implies that $\mathbb{E}\left[(\alpha_p'\left(z_t^- - \mu(j)\right))^4\right]$ and therefore $\mathbb{E}\left[(\alpha_p'\left(z_t^- - \mu\right))^4\right]$ is bounded). Noting that $\sum_{t=1}^{T} \mathbb{E}\left[X_{Tt}^2 \mathbf{I}\left(X_{Tt}^2 > \epsilon\right)\right] = \sum_{j=1}^{J} \lfloor T\omega_j \rfloor \mathbb{E}\left[X_{Tt}^2 \mathbf{I}\left(t \in S_j\right) \mathbf{I}\left(X_{Tt}^2 > \epsilon\right)\right]$, condition (i) also follows by the similar arguments as in Theorem 4. □