Persistence-Robust Granger Causality Testing

Dietmar Bauer
Arsenal Research
Vienna, Austria

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

October 15, 2008

Abstract

The observed persistence common in economic time series may arise from a variety of models that are not always distinguished with confidence in practice, yet play an important role in model specification and second stage inference procedures. Previous literature has introduced causality tests with conventional limiting distributions in I(0)/I(1)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. Building on this approach, but using an infinite order VARX framework, we 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. Since the distribution under the null hypothesis is the same in all cases, no prior knowledge, or first-stage testing or estimation is required and known lag orders are not assumed.

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

[*] Corresponding author. Previously entitled "Robust Granger Causality Tests in the VARX Framework." We thank 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 applications ranging from the causal relation between money and output (Friedman and Kuttner, 1992) to the export led growth hypothesis (Marin, 1992). This paper develops a simple but flexible augmented VARX approach to Granger causality testing that is highly robust to the degree and nature of the persistence in the causing variables. In particular, the same estimator and test statistic may be employed to test causality,[1] regardless of whether the true, but unknown, data generating process for the causal variable is characterized by stationarity, long-memory/fractional integration, a local-to-unity process, I(1) behavior, or breaks in the mean of the process. Consequently no prior knowledge, pre-estimation, or pre-test is required. Likewise, known lag orders are not assumed. Since the Granger causality test is based on the same Wald statistic and Chi-squared limiting distribution in all five cases, no prior knowledge, estimation, or testing is required to distinguish between these processes.

These are desirable characteristics for several reasons. Frequently it is difficult to determine the degree and nature of the persistence of the forcing variables with full confidence. This can often matter in both theory and practice for second stage model specification and inference. Likewise, recent developments in the cointegration literature have also stressed the importance of allowing for fractional integration[2] and near unit roots (Jansson and Moreira, 2006). In fact, two of the most recent studies (Phillips, 2005; Muller and Watson, 2007), emphasize agnostic approaches to the form of this persistence, motivated by robustness concerns similar in spirit to ours.

The difficulties associated with distinguishing $I(1)$ and $I(0)$ processes are well known. In short, unit root tests may have low power, confidence intervals on the largest root are often wide in practice, and problems of near observational equivalence put bounds on our ability to distinguish between true $I(1)$ series and persistent $I(0)$ series in finite sample (Faust, 1996; Faust, 1999). Moreover, for many purposes, processes with near unit roots may be better modelled as local-to-unity processes, which depend on a local-to-unity parameter, which cannot be consistently estimated in a time series context (Phillips, 1987; Chan, 1988; Nabeya and Sørensen, 1994).These problems may be further complicated with the allowance for structural breaks, in which numerous modelling possibilities arise, unit root tests become more complicated, and the similarity between break-stationary and unit root processes increases with the number of breaks. Recent literature has also highlighted the difficulties in distinguishing between structural breaks and long-memory/fractionally integrated processes (Diebold and Inoue, 2001; Gourieroux and Jasiak, 2001; Granger and Hyung, 2004). Thus while it is often easy to recognize that a given series is persistent, it can be far more difficult to determine with confidence the right approach for modelling this persistence. As

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

[2] (Jeganathan, 1999; Robinson and Marinucci, 2003; Robinson and Hualde, 2003; Kim and Phillips, 2004; Hualde, 2006; Hualde and Robinson, 2007, among others).

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. A typical VAR may be specified in levels if unit roots are rejected, in first-differences if the variables are individually integrated but not cointegrated, and in error-correction format if the variables are cointegrated. Likewise, structural breaks require explicit modelling and long-memory processes are not easily accommodated in a VAR setting. Such choices can have important practical implications. A recent controversial example involves the role of technology shocks in macroeconomic models, for which VARs provide evidence consistent with New Keynesian models if hours worked is treated as nonstationary and enters in differences (Gali, 1999) but supports instead the conclusions of standard Real Business Cycle models when the same variable is presumed stationary and enters in levels (Christiano *et al.*, 2003).

The persistence of the causing variable also matters for second-stage inference. Even in simple regression models, different critical values may apply depending on whether or not the regressor contains a unit root. Moreover, both stationary and unit root asymptotics can be misleading in the presence of near unit roots when roots are close, but not equal, to unity, as often modelled by the near unit root local-to-unity model mentioned above (Cavanagh *et al.*, 1995; Elliott, 1998). Such inference problems may be further complicated once one allows more realistically for the possibility of structural breaks or long-memory processes. They also matter in practical applications and have played a particularly important role in the predictive regression literature, including tests of stock return predictability (Stambaugh, 1999; Lewellen, 2004; Torous *et al.*, 2005), the expectations hypothesis of the term structure (Lanne, 2002), and uncovered interest rate parity (Baillie and Bollerslev, 2000; Maynard and Phillips, 2001), all of which constitute special cases of Granger causality testing.

Our approach builds on a rich literature, the original insights for which derive from the work of Park and Phillips (1989) and Sims *et al.* (1990). Their results imply that, despite the nonstandard asymptotics in I(1) /cointegrated systems, parameters that may be expressed as coefficients on stationary regressors retain a standard root-T normal asymptotic distribution. Similar results have also been found to 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 stationary and which have unit roots. Saikkonen and Lütkepohl (1996) extended these results to infinite order VARs.

This approach is very flexible with respect to inference in general I(0)/I(1) and cointegrated models. On the other hand, the pure VAR framework adopted in these surplus lag methods cannot accommodate long-memory, nor can it accommodate breaks unless they are explicitly modelled. Yet, since breaks are often tested for in conjunction with unit roots, the requirement that they be specifically modelled detracts somewhat from the advantageous features of the surplus lag approach. By incorporating an exogenously modelled component, we may accommodate a richer class of persistent processes for the forcing variable in the VARX framework, including those with long-memory/fractionally integration or unmodelled structural breaks. Moreover, we find that with the incorporation of the surplus lag, the null limit distribution continues to be unaffected by the particular form of this persistence. Likewise, these 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 our results also build on the literature on reasonable approximability (Berk, 1974; Lewis and Reinsel, 1985; Lütkepohl and Saikkonen, 1997) and provide some extensions to allow for exogenous regressors, including those with long-memory, and certain structural breaks. Some related extensions are provided by Poskitt (2007), who establishes autoregressive approximations to (univariate) non-invertible and stationary long-memory processes.

The simplicity and generality of the surplus lag approach does not come without cost. Naturally, the addition of an extra unnecessary lag reduces the efficiency of estimation, thereby leading to reduced power relative to a correctly specified model. However, as previous literature reports, the magnitude of this effects varies considerably. Power losses are greatest in unit root and cointegration tests, in which super-consistency and thus power against $O(T^{-1})$ alternatives is lost. Generally, the surplus lag is not recommended in this case, even by its proponents (Toda and Yamamoto, 1995; Saikkonen and Lütkepohl, 1996).[3] However, efficiency losses are often far more moderate in type of the Granger causality tests considered here, particularly when the baseline model already includes a number of lags, as in common macroeconomic applications. Therefore the excess lag approach provides a persistence-robust complement to, but not a substitute for, more efficient testing procedures.[4]

In the I(0)/I(1) context there arguably exist alternative methods that are as general, but more efficient, than existing results for the VAR-based surplus lag method. When the number of cointegrating vectors and orders of integration is known, error correction models provide a natural context for efficient causality testing, although in practice pre-tests are required (Toda and Phillips, 1993). The fully modified VAR estimation (Phillips, 1995; Kitamura and Phillips, 1997) is also efficient and shares the advantage of not requiring a priori knowledge on the number of $I(0)$ and unit root components. Fully modified regression can also be extended to cover fractional cointegration (Kim and Phillips, 2004). Nevertheless, most efficient tests designed for the $I(0)/I(1)$ case

[3] The surplus lag approach may also be unsuited to applications, such as forecasts for the persistent variable itself, in which explicit modelling of the low-frequency behavior is unavoidable.

[4] For the applied researcher, our approach should be particularly useful in confirming the robustness of rejections, whereas some caution should be applied in interpreting a failure to reject.

require adjustment in the presence of either near unit roots (Elliott, 1998) or fractional integration, whereas we show that, in the VARX context, the same surplus lag test continues to work without adjustment in both cases.

A second limitation of our approach is that we allow for long-memory in the forcing processes but not the error process for the dependent variables. The difficulty of weakening this assumption for time domain estimators is discussed in (Hidalgo, 2000; Hidalgo, 2005), who provides frequency based non-parametric causality tests, which allow for long-memory in both. On the other hand, these tests require covariance stationarity, ruling out many of the interesting cases considered here.

A second limitation of the surplus lag approach is that, like many econometric estimators, it provides only correct large sample, not finite sample size. In a simple bi-variate predictive regression context, sign and sign rank tests are both non-parametric and exact in finite samples, thus providing a very attractive alternative (Campbell and Dufour, 1997). Unfortunately, they do not easily generalize to more complicated models. As recent work has demonstrated (Dufour and Jouini, 2005), Monte-Carlo methods may also be employed to provide finite sample inference even in quite complicated parametric models. In principle, the generality of this approach is limited only by its computational complexity. On the other hand, the econometrician must simulate from all possible parametric models for the forcing variable, a set which may become increasingly large once we consider the possibility of long-memory and breaks. Moreover, the fact that exactly the same VARX based surplus lag test statistic works without modification in all the standard cases considered here hints at the possibility that the technique may work for a much wider class of processes. This may be an appealing aspect to those who suspect that the true mechanisms generating the persistence in the data are likely more sophisticated than the econometric models typically employed to capture them (Phillips, 2003, for example).

The remainder of the paper is organized as follows. Section 2 presents the model and explains the basic intuition behind our results. Section 3 presents the large sample results, showing that the VARX based excess lag Wald statistic has the same null limiting Chi-squared distribution for a variety of forcing processes. Section 4 provides some simulation results on the finite sample size and power of the test. Appendix A collects some technical results and the proofs of the main theorems are provided in Appendix B. The tables and figures are included at the back of the paper.

Finally, a word on notation. Lag orders are given by p and dimensions are given by k. Capital letters denote regression matrices. Variables with time subscripts are lower case. A variable with a minus sign in the superscript (e.g. x_t^-) includes all relevant lags. The dependent variable is denoted by y_t, the variables about which Granger causality is tested are often written in terms of x_t when referring to particular applications, but as z_{1t} for the theoretical results, and any additional control variables are included as z_{2t}. 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 $max\left\{\|Ax\|_2 : x(n \times 1), \|x\|_2 = 1\right\}$ when applied to the $m \times n$ matrix A.

2 The model

Throughout this paper we consider three basic variables: by y_t we denote a k_y vector of dependent variables, by z_{1t} we denote a k_{z1} vector of exogenously modelled forcing variables, and z_{2t} denotes an optional k_{z2} vector of endogenously modelled control variables.

We consider tests of the null hypothesis that z_{1t} does not Granger cause y_t after controlling for z_{2t}.[5] Let $\mathcal{F}_{t,y,z1,z2}$ define the information set generated by the past history of all three variables (i.e. by $\left\{(y'_{t-j}, z'_{1t-j}, z'_{2t-j})', j \geq 0\right\}$). Likewise, let $\mathcal{F}_{t,y,z2}$ denote the information set generated by the past history of the endogenous variables only, (i.e. by $\left\{(y'_{t-j}, z'_{2t-j})', j \geq 0\right\}$). Then we test the Granger noncausality condition

$$\mathbb{E}\left[y_t|\mathcal{F}_{t-1,y,z1,z2}\right] = \mathbb{E}\left[y_t|\mathcal{F}_{t-1,y,z2}\right]. \tag{1}$$

In practice this hypothesis is often tested by means of parameter restrictions on a joint VAR involving all three variables. However, our interest lies in cases in which the forcing variable z_{1t} displays persistent behavior, which may potentially be modelled in a variety of different ways. In this context, the pure VAR may be too restrictive. To maintain flexibility, the forcing variable z_{1t} is instead exogenously modelled in the sense that we do not explicitly model its possible dependence on past y_{t-j} and z_{2t-j}. This allows us to consider a number of alternative data generating processes for z_{1t} on a case by case basis, including $I(0)$, $I(1)$, local-to-unity, stationary and nonstationary long-memory processes, and processes with structural breaks. We will defer discussion of the exact modelling assumptions on z_{1t} to Section 3. However, it should be noted that although z_{1t} is exogenously modelled, it is not assumed to be strictly exogenous in a statistical sense. The innovations in the process for z_{1t} may correlate with the past innovations to y_t and z_{2t}.

We maintain more traditional assumptions regarding the behavior of the endogenously modelled variables. 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).

In practice, the infinite order VAR in (2) is approximated using a finite autoregression. Our primary interest lies in the process for y_t, which is approximated by

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

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

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[6]

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

and test the joint parameter 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. The role of the surplus lag 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. For instance, when z_{1t} is $I(1)$ it well known that they have a $\sqrt{T}$ convergence rate and joint normal limiting distribution (Park and Phillips, 1989; Sims *et al.*, 1990). If the integration order of z_{1t} were to exceed 1.5, e.g. $z_{1t} \sim I(2)$, a second surplus lag would be required. However, we do not consider this possibility.

The choice of p_{z1}, the lag order of z_{1t}, will generally influence the power of the test, but not its large sample size. In particular, the test has power against a more general set of alternatives for large p_{z1}, but has greater power against simpler alternatives, when the lag-length is small. Also, p_{z1} need not be set equal to p, the lag order of w_t. This is another way in which 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, in the pure VAR framework we require a surplus lag of all variables, whereas in the VARX we require only an extra lag of z_{1t}. This may improve efficiency, particularly when the number of lags 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

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

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

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 $Y = \left[\begin{array}{ccc} y_{p_{max}+1}^-, & \dots, & y_T^- \end{array} \right]'$, 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.

The null hypothesis of no-Granger causality is then $H_0 : \psi_{x1} = 0$ and the alternative hypothesis is $H_A : \psi_{x1} \neq 0$. Defining $X_{1.2} = X_1 - X_2(X_2'X_2)^{-1}X_2'X_1$, with rows denoted by $\left(x_{1.2t}^-\right)'$, as the residual from the projection of X_1 on X_2, we estimate the parameter of interest ψ_{x1} by $\hat{\psi}_{x1} = Y'X_{1.2}\left(X_{1.2}'X_{1.2}\right)^{-1}$ and the variance of $vec\left(\hat{\psi}_{x1}\right)$ is estimated in the standard way 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 $\hat{\varepsilon}_{yt,p} := y_t - \hat{\psi}_{x1}x_{1t}^- - \hat{\psi}_{x2}x_{2t}^-$.[7] The standard Wald test for $\psi_{x1} = 0$ then takes the form:

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

In the section below, we show that under suitable regularity conditions $\hat{W}$ has a χ^2 null limiting distribution for a wide variety of data generating processes under which z_{1t} may exhibit persistent behavior.

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 a variety of assumptions regarding the nature of the persistence in z_{1t}. We begin by stating the assumptions on the innovation process for the endogenous variables w_t specified in (2):[8]

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 $\varepsilon_t, \varepsilon_{t-1}, \dots$. Further assume that $\mathbb{E}\{\varepsilon_t\varepsilon_t'|\mathcal{F}_{t-1}\} = \mathbb{E}\varepsilon_t\varepsilon_t' = \Sigma > 0$ and that $\mathbb{E}\{\varepsilon_{t,a}\varepsilon_{t,b}\varepsilon_{t,c}|\mathcal{F}_{t-1}\} = \omega_{a,b,c}$ (constant) where $\varepsilon_{t,a}$ denotes the a-th coordinate of the vector ε_t. Finally finite fourth moments are assumed: $\mathbb{E}\{\varepsilon_{t,i}^4\} < \infty$.*

Many of the results presented below may be proved under more general assumptions on the innovations. In particular, finite fourth moments are often unnecessary. However, the above assumptions are standard in 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 is dropped the asymptotic

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

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

distributions change and the estimators have to be adapted to account for possible heteroskedasticity. We will not go into details in this respect.

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 $p = p(T)$ be a function of the sample size T and p_{z1} be a fixed integer. Then assume that the following conditions have been verified:*
(i) $\hat{\Sigma}_\varepsilon \to \Sigma$ in probability.
(ii) $\hat{\Gamma}_{1.2} \to \Gamma_{1.2}$ in probability 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 $T^{-1/2}vec(\sum_{t=p+1}^T \varepsilon_{yt,p}(x_{1.2t}^-)') \xrightarrow{d} Z_\varepsilon$ where $Z_\varepsilon \in \mathbb{R}^{k_{z1}k_y}$ is normally distributed with mean zero and variance $\Gamma_{1.2} \otimes \Sigma$.
From these high level assumptions the standard asymptotics for the Wald test are immediate from (8).

Theorem 1 *Let Assumption HL hold for $\varepsilon_{yt,p} = y_t - \psi_{x2}x_{2t}^- - \psi_{x1}x_{1t}^-$ where the integer p is chosen as in HL (iii). Then, under the null hypothesis $H_0 : \psi_{x1} = 0$, the Wald statistic $\hat{W}$ converges in distribution to a χ^2 distributed random variable with $k_y p_{z1} k_{z1}$ degrees of freedom where $p_{z1}k_{z1}$ equals the dimension of x_{1t}^-.*

Of course the high level assumptions are not directly applicable. In the following it will be shown that in a multitude of circumstances the high level conditions are fulfilled.

3.1 Infinite Order Stationary $VARX$

We first address the stationary case. Under the null hypothesis we will assume that the joint process $w_t := [y_t', z_{2t}']'$ admits a $VAR(\infty)$ representation of the form given in (2) where $\varepsilon_t = \begin{bmatrix} \varepsilon_{yt}' & \varepsilon_{z2t}' \end{bmatrix}'$ fulfills Assumption N. The noise assumptions are in line with Theorem 7.4.8. of Hannan and Deistler (1988) which extends Theorem 4 of Lewis and Reinsel (1985) where $(\varepsilon_t)_{t\in\mathbb{Z}}$ was assumed to be i.i.d. Many results building on Lewis and Reinsel (1985) also use independent noise (Lütkepohl and Saikkonen, 1999; Lütkepohl and Saikkonen, 1997; Dolado and Lütkepohl, 1996).

As described in Section 2, the test is based on the auxiliary model given in (4), which for $p = \infty$ nests the true model. Here $H_0 : \psi_{z1j} = 0, j = 1, \ldots, p_{z1}$ specifies the null hypothesis while a wide range of alternatives can be included for large p_{z1}. In the following we will only consider the case that p_{z1} is some pre-specified integer rather than the more general case, in which $p_{z1} \to \infty$ as a function of the sample size. It should be noted, however, that $p_{z1} \to \infty$ in some situations could also be dealt with leading to a more complicated asymptotic theory (Saikkonen and Lütkepohl, 1996). However, it is not clear how the integer p_{z1} in such a setting could be chosen. Additionally the assumptions on z_{1t} in order for such results to hold are more restrictive. We will not go into details in this respect.

We formalize the notion of reasonable approximability as is usually done in the literature (Lewis and Reinsel, 1985; Lütkepohl and Saikkonen, 1997):
Assumption P1:

(i) The noise $(\varepsilon_t)_{t\in\mathbb{Z}}$ fulfills Assumption N.
(ii) $\sum_{j=1}^{\infty}\|\pi_{w,j}\|_2 < \infty$. For $\pi_w(z) := I - \sum_{j=1}^{\infty}\pi_{w,j}z^j$ it holds that $\det\pi_w(z) \neq 0, |z| \leq 1$.
(iii) The integer p tends to infinity as a function of the sample size 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

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

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, unlike in (Lewis and Reinsel, 1985), the process $(z_{1t})_{t\in\mathbb{Z}}$ is not modelled endogenously. In other words the VAR framework of Lewis and Reinsel (1985) is exchanged with VARX modelling here. An important advantage of such an approach is that it allows us to vary the integer p_{z1} freely, i.e. the choice of p_{z1} is not tied to the approximation properties. Also we do not need to assume that $(z_{1t})_{t\in\mathbb{Z}}$ has a $VAR(\infty)$ representation. Hence the assumptions include overdifferenced processes which are excluded in Lewis and Reinsel (1985). Furthermore z_{1t} may be a function of lagged y_{t-j} and $z_{2t-j}, j \in \mathbb{N}$. The assumptions on $(z_{1t})_{t\in\mathbb{Z}}$ nevertheless exclude a number of interesting cases: Due to the summability assumptions on the coefficients θ_j and ϕ_j the process $(z_{1t})_{t\in\mathbb{Z}}$ is stationary. Integrated processes will be dealt with in later sections. Also certain long memory processes are excluded (see below). These are the topic of section 3.4.

The following result is a consequence of the proof of Theorem 3 of Lewis and Reinsel (1985), which it extends to the VARX framework:

Theorem 2 *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 P1 implies Assumption HL.*

The theorem shows that in situations where the true process follows a VARX(∞,p_{z1}) the Wald test statistic can be used as if the true process was a $VARX(p, p_{z1})$. The main advantage in comparison to previously obtained results, such as those in Dolado and Lütkepohl (1996), is that the regressors z_{1t} are not modelled endogenously. This leads to a more parsimonious model in the case where y_t can be modelled using only a few lags relative to a full VAR model also containing z_{1t}. From the proof of the theorem it is obvious that in this special case the result also holds if the surplus lag $z_{1t-p_{z1}-1}$ is not included in x_{2t}^-. The inclusion of this additional term may be expected to reduce the power of the tests where the decrease in power depends on the characteristics of z_{1t} and the lag length p_{z1}. Dolado and Lütkepohl (1996) contains a discussion on these issues. This power loss is the downside to the robustness properties of the test provided in the following sections.

3.2 Infinite Order Nonstationary $VARX$

In the last subsection we dealt with stationary processes. One of the main motivations behind the surplus lag approach was to obtain results without unit root and cointegration pre-testing in cases where components of y_t and/or z_t might be (co)integrated. This is allowed for in the following assumption.

Assumption P2:
(i) 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 (using a random value* w_0 *that has finite fourth moments and is independent of* $\varepsilon_t, \eta_t, t > 0$*)*

$$v_t := \left[\begin{array}{c} \gamma_\perp'(w_t - w_{t-1}) \\ \gamma' w_t \end{array} \right] \tag{10}$$

has an autoregressive 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 power series: $\sum_{j=1}^\infty j \|\pi_{v,j}\|_2 < \infty$.
(iv) The integer p is chosen as a function of the sample size such that $p^3/T \to 0$ *and* $T^{1/2} \sum_{j=p+1}^\infty \|\pi_{v,j}\|_2 \to 0$.
(v) The process $(z_{1t} - z_{1t-1})_{t\in\mathbb{N}}$ *for random* z_{10} *(finite fourth moment, independent of* $\varepsilon_t, \eta_t, t > 0$*) fulfills Assumption P1(iv) where additionally* $\sum_{j=0}^\infty j \|[\theta_j, \phi_j]\|_2 < \infty$ *holds.*

Here γ denotes the cointegrating relations as $(\gamma' w_t)_{t\in\mathbb{N}}$ is assumed to be stationary. Note that these assumptions also include many processes fulfilling Assumption P1. They lead to the following Theorem:

Theorem 3 *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 P2 implies Assumption HL.*

The theorem extends the results of Theorem 5 of Saikkonen and Lütkepohl (1996) from the VAR to the VARX framework with the same advantages as in the stationary case. The integer p_{z1} is not tied to the approximation quality and more general processes z_{1t} are allowed for. This is an appealing and useful feature of Granger-noncausality test in many different frameworks. All processes y_t, z_{1t} and z_{2t} can be stationary, integrated or cointegrated and we do not use any information on possible cointegrating relations. This robustness property follows from the fact that the properties of $\varepsilon_{t,p}$ are identical across all cases and that in all cases there exists a sequence of matrices τ_p such that $x_{1t}^- - \tau_p x_{2t}^-$ is stationary. The fact that this is sufficient for standard asymptotics to hold is also acknowledged in Theorem 1 of Toda and Phillips (1993) and footnote 3 of Sims *et al.* (1990).

Of course this robustness comes at a price: If z_{1t} is stationary then one lag of z_{1t} is unnecessarily omitted from the test restriction, leading to a loss of power with respect to tests under correct specification as noted by Dolado and Lütkepohl (1996). From the proof of Theorem 3 it is obvious that for stationary z_{1t} the variable $z_{1t-p_{z1}-1}$ can be omitted in the definition of x_{2t}^{-} and the asymptotic distribution of $\hat{W}$ is unchanged but the resulting test has higher power since it is based on a smaller model. If z_{1t} is in fact integrated but is considered to be stationary such that no lag of z_{1t} is included in x_{2t}^{-} then the asymptotic distribution is incorrect as is the size of the test. Hence, in situations where z_{1t} clearly is stationary rather than integrated the omission of $z_{1t-p_{z1}-1}$ seems to be the better procedure. If this decision cannot be made with certainty then there is an argument for using the robust version of the test given the fact that for large p_{z1} the expected loss in power is small whereas the effect of misspecification can be substantial.

If additionally the cointegrating rank of the joint process $[y_t', z_{2t}', z_{1t}']'$ is known then superior tests can exploit this knowledge (Toda and Phillips, 1993). In this situation the power loss can be substantial due to the T-consistency of the estimators of the cointegrating vectors as compared to the $\sqrt{T}$ consistency of the excess lag estimators. Note however, that in any case there is a risk of misspecification. The proposed tests given in this paper sacrifice power in special cases for obtaining robust inference under a wide variety of possible assumptions on the data generating process. The surplus lag Wald test $\hat{W}$ relies on neither pre-test nor pre-estimation to obtain a distribution that is invariant to the number of I(0) and I(1) components and the same invariance extends to the local-to-unity framework.

3.3 Local-to-unity processes

We next address the behavior of the surplus lag Wald test $\hat{W}$ under the local-to-unity framework, introduced by (Phillips, 1987; Chan, 1988), which bridges the gap between asymptotic theory for integrated and stationary processes. In this model, the largest root of z_{1t}, say $a_T = 1 + c/T$, is specified as a Pitman drift that approaches unity as $T \to \infty$. This is a modeling device, whose asymptotics have been found quite accurate in approximating small sample distributions when roots are slightly less than unity. It has played an important role in financial and macroeconomic applications in which a number of variables that should be stationary according to economic theory, are nonetheless found to be highly persistent in practice. It is also precisely the case in which the power of unit root tests is low and therefore uncertainty exists regarding the choice between level and difference specifications. This poses a challenge for inference since the asymptotic distribution of most estimators depends on the value of c, a parameter which cannot be consistently estimated using a single time series. For example, Elliott (1998) finds that the critical values of most common I(1)/cointegration methods, including those involving pre-test, may be inaccurate when the true process is local-to-unity. Likewise, stationary asymptotics are also not generally appropriate in this case. Thus, with a few recent exceptions (Jansson and Moreira, 2006; Maynard and Shimotsu, forthcoming) bounds procedures have been required due to critical values that depend on c (Cavanagh *et al.*, 1995, e.g.).

As we show below, the robustness advantage of the surplus lag method becomes particularly apparent in a local-to-unity context. Since the surplus lag Wald test $\hat{W}$ relies on neither pre-test nor pre-estimation to obtain a distribution that is invariant to the number of I(0) and I(1) components, it may be expected that the same invariance extends to the local-to-unity framework.

We will use the following assumptions:

Assumption P3 :
(i) Define $A_{T,w} := I + C_w/T, C_w = diag(c_1, c_2, \ldots, c_{k_y+k_{z2}-n})$ *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*)*

$$v_t := \begin{bmatrix} \gamma_\perp' w_t - A_{T,w}\gamma_\perp' w_{t-1} \\ \gamma' w_t \end{bmatrix} \tag{11}$$

has an autoregressive 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 P3 y_t, z_{1t} and z_{2t} are all defined as triangular arrays[9] 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 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 $c_i = c_{z,i} = 0$. It is obvious that Assumption P3 implies Assumption P2 with $c_i = c_{z,i} = 0$ for all i.

The theorem below shows that the same high level assumptions are implied under the local-to-unity model, which confirms that $\hat{W}$ maintains the same asymptotic normality irrespective of the value of c in the local-to-unity model. As in the I(1) case, the main technical reason is that there exists a sequence of matrices τ_p such that $x_{1t}^- - \tau_p x_{2t}^-$ is stationary.

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 P3 implies Assumption HL.*

The extension of this robustness result for the Wald test from the standard $I(1)$ framework to the local-to-unity framework is new to the best of our knowledge. From a

[9] For notational simplicity we follow common practice in suppressing the dependence on T due to the localization.

theoretical perspective, it is perhaps not surprising, but as discussed above, this is a rare property that has useful practical implications for the value of this method as a robust test.

3.4 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 important to modelling a number of phenomena in economics and finance, as well as in the natural sciences (Baillie, 1996). For values of $d < 0.5$, the process fits into a larger class of stationary long-memory models. Values of $d > 0.5$ correspond to nonstationary fractional integration. Because the degree of fractional integration d is consistently estimable, long-memory models do not pose quite the same difficulties for inference as the local-to-unity model. Nevertheless, specialized inference techniques are often required for addressing systems involving fractional (co)-integration.[10] Typically, these techniques depend on consistent estimates of the parameter d and are thus not the same as the techniques that would be applied under other models of persistent behavior, such as the near unit root model analyzed above. An advantage of the surplus lag Wald test statistic shown below is that it has a standard limit distribution for both stationary and nonstationary long-memory forcing variables.

3.4.1 Stationary long-memory

Assumption P1 for stationary infinite VARX processes imposed summability assumptions on the impulse response sequence of the exogenous process $(z_{1t})_{t\in\mathbb{N}}$. These assumptions exclude long-memory processes such as fractionally integrated processes with $0 < d < 0.5$. In this section we will provide less restrictive assumptions including such processes:

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 for this choice of p.

These assumptions on the exogenous inputs include many long-memory processes and in particular fractionally integrated processes. In fact the assumptions are much weaker than this including sums of fractionally integrated processes. Since the squared

[10] See footnote 2 for a partial list of references.

coefficients for $d \approx 0.5, d \leq 0.5$ are just summable, the conditions on the impulse response sequences are close to minimal. On the other hand, it has been found necessary to introduce 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 $p = T^c$ will fulfill the approximation restriction and choosing $c < 1 - 2d$ the condition on p can be easily met.

In this setting the advantage of the VARX framework is most clearly visible. If instead one modelled the process $[y_t', z_{1t}', z_{2t}']'$ using the VAR framework then overly large orders are needed in order to obtain a small approximation error $\varepsilon_{yt,p} - \varepsilon_{yt}$ due to the slow decay of the coefficients in the VAR(∞) representation of the true process. In the VARX framework this difficulty does not arise. Moreover, overdifferenced processes can also be used as regressors since they do not need to be approximated using autoregressive terms.

Again it can be shown that Assumption HL holds.

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 P4 implies Assumption HL.*

This result for stationary long-memory does not depend on the surplus lag and also holds if $z_{1t-p_{z1}-1}$ is omitted in the definition of x_{2t}^-. Again this results in a power-robustness tradeoff.

3.4.2 Nonstationary long-memory

It has been observed in the literature that the estimates of d for fractionally integrated processes often are close to $d = 0.5$ (cf. the references given in Baillie, 1996, section 6, p. 43). The last theorem showed that the Wald test is robust with respect to fractional integration for $-0.5 < d < 0.5$. Previously, robustness with respect to integration has been given in Theorem 3. We next combine these two results to conclude 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* $\beta_\perp' z_{10} = 0$

$$\begin{bmatrix} \beta_\perp'(z_{1t} - z_{1t-1}) \\ \beta' z_{1t} \end{bmatrix} = v_t, t \in \mathbb{N} \tag{12}$$

where

$$v_{i,t} = \sum_{j=0}^{\infty} L_i(j) \frac{\Gamma(j+d_i)}{\Gamma(d_i)\Gamma(j+1)} \alpha_i' \begin{pmatrix} \nu_{t-j} \\ \varepsilon_{t-j} \end{pmatrix}, \tag{13}$$

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* $p = o_p\left(T^{\min\{1/3, 1-2d_{\max}, 1/3(1+2d_{\min})\}}\right)$ *and* $T^{1/2}\sum_{j=p+1}^{\infty} \|\pi_{w,j}\|_2 \to 0$ *for this choice of* p*.*

Nonstationary fractional integration in the forcing variable is allowed for through the hyperbolic rates of decay on $\beta_\perp'(z_{1t} - z_{1t-1})$, through (13), which allows for different values of d in each element of $\beta_\perp' z_{1t}$.[11] The cointegrating residuals, given by $\beta' z_{1t}$, 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), as discussed in Davidson and Hashimzade (2007). In comparison to our previous assumptions, the required restrictions on the increase of p as a function of the sample size are striking. Assumption P4 showed problems for the fractional integration parameter 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. The reason for problems in that case is 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. On the other hand, the remaining assumptions are far from minimal. In particular the assumptions on v_t appear to be overly strong. Again we stress that we are not interested in the most general setup but only in providing cases in which the standard asymptotics for the Wald test hold:

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 P5 implies Assumption HL.*

The theorem shows that in the case that the exogenous regressors are fractionally integrated of order $0.5 < d < 1.5$ the null asymptotics of the surplus-lag Wald test for Granger-noncausality remain standard. 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.

[11] Marinucci and Robinson (1999) distinguish between two types of fractional integration, Type I and Type II, depending on the treatment of the initial conditions. Our assumptions allow for Type I fractional integration.

3.5 Stationary processes with 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 $j = 1, \dots, J$ denote the fraction of the sample spent in regime j between the $(j-1)^{th}$ break and the j^{th} break, with $\sum_{j=1}^{J} \omega_j = 1$, the process $(z_{1t})_{t\in\mathbb{Z}}$ is generated according to the equation

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

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 $\sum_{j=0}^{\infty} \|[\theta_j, \phi_j]\| < \infty$ is assumed.

While we do not assume that z_{1t} is explicitly modelled in the empirical analysis, the estimated VARX must either include an intercept, or at a minimum, z_{1t} must be demeaned prior to estimation. It will be convenient to work with deviations from means. Let $x_t^- := \begin{bmatrix} (x_{1t}^-)' & (x_{2t}^-)' \end{bmatrix}'$ denote the full set of regressors and define

$$\mu(j) := \mathbb{E}\left[x_t^- \mathbf{I}\left(t \in S_j\right)\right] \text{ and } \bar{\mu} := \sum_{j=1}^{J} \omega_j \mu(j)$$

as the mean within regime j and the average mean across regimes, respectively. Here we define $S_j = \left\{p_{z1} + 1 + \lfloor \sum_{k=0}^{j-1} \omega_k T \rfloor, \dots, \lfloor \sum_{k=1}^{j} \omega_k T \rfloor \right\}$ as the data range that would result if restricted to regime j only.

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 $x_t^- - \bar{x}^-$. To keep the proofs simple, we first show that the infeasible estimator $\widehat{W}\left(\bar{\mu}\right)$ has the correct large sample distribution.

Theorem 7 *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 P6 implies Assumption HL is satisfied for the infeasible Wald statistic $\widehat{W}\left(\bar{\mu}\right)$.*

The result is easily extended to the feasible statistic $\widehat{W}\left(\bar{x}^{-}\right)$ in the following corollary, the proof of which is omitted.

Corollary Assume that P6 holds. Then under the null hypothesis $H_0 : \psi_{x1} = 0$ the Wald statistic $\widehat{W}\left(\hat{x}^{-}\right)$ converges in distribution to a χ^2 distributed random variable with $k_y p_{z1} k_{z1}$ degrees of freedom where $p_{z1} k_{z1}$ equals the dimension of x_{1t}^{-}.

4 Simulation results

In this section we conduct a small Monte Carlo study to investigate the finite sample performance of the test. Following the notation above, we refer to y_t as the dependent variable and z_{1t} as the forcing variable. The primary motivation for the inclusion of the excess lag is to enhance the robusness of the resulting causality test to both the degree and nature of the persistence in the forcing variable. Therefore we consider several different models for z_{1t}, including I(0), I(1), near-unit root, and fractionally integrated cases. We also consider both $I(0)$ and $I(1)$ models of the dependent variable and allow both for cases in which y_t and z_{1t} are cointegrated and for cases in which y_t and z_{1t} are nonstationary but do not cointegrate. Some related simulation results for the pure VAR based surplus-lag tests in I(1) and cointegrating settings are provided by Dolado and Lütkepohl (1996) and, more extensively, by Swanson *et al.* (2003). However, we are not aware of any results on their performance in near unit root or long-memory settings. We also present some new small sample power comparisons.

4.1 Simulation Models

In the simulation models presented below, we denote by δ the parameter that is used to measure the distance of the true model from the null hypothesis. With a few exceptions, we test $H_0 : \delta = 0$ against $H_A : \delta \neq 0$ and thus we provide results for both finite sample size ($\delta = 0$) and power ($\delta \neq 0$). We note, however, that the meaning of δ differs in each specification and therefore test power is not directly comparable across models. The innovation process is specified as[12]

$$\varepsilon_t' = (\varepsilon_{1t}, \varepsilon_{2t}) \sim \text{i.i.d. } N(0, \Sigma) \text{ for } \Sigma = \begin{bmatrix} 1 & \sigma_{12} \\ \sigma_{12} & 1 \end{bmatrix}, \text{ and } \sigma_{12} = -0.8. \qquad (14)$$

4.1.1 I(0), I(1), and cointegration cases

We first consider the behavior of the test under three standard models involving I(0), I(1), and cointegrated variables. The first is a stationary levels VAR (levels-VAR) in

[12] Granger noncausality (condition 1) has no implication for the residual cross-correlation, σ_{12}.

which both y_t and z_{1t} are I(0):

$$\begin{bmatrix} y_t \\ z_{1t} \end{bmatrix} = \begin{bmatrix} 0.5 & \delta \\ 0.3 & 0.5 \end{bmatrix} \begin{bmatrix} y_{t-1} \\ z_{1t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{bmatrix}. \tag{15}$$

The second is a difference VAR (difference-VAR) in which both y_t and z_{1t} are I(1) and there are no cointegrating vectors:[13]

$$\begin{bmatrix} \Delta y_t \\ \Delta z_{1t} \end{bmatrix} = \begin{bmatrix} 0.5 & \delta \\ 0.3 & 0.5 \end{bmatrix} \begin{bmatrix} \Delta y_{t-1} \\ \Delta z_{1t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{bmatrix}. \tag{16}$$

The third type model we consider is a vector error correction model (VECM), in which y_t and z_{1t} are I(1) and cointegrated:

$$\begin{bmatrix} \Delta y_t \\ \Delta z_{1t} \end{bmatrix} = \begin{bmatrix} -\delta_1 \\ 1 \end{bmatrix} \begin{bmatrix} 1 & -\delta_2 \end{bmatrix} \begin{bmatrix} y_{t-1} \\ z_{1,t-1} \end{bmatrix} + \begin{bmatrix} 0.5 & \delta_3 \\ 0.3 & 0.5 \end{bmatrix} \begin{bmatrix} \Delta y_{t-1} \\ \Delta z_{1t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{bmatrix} \tag{17}$$

Re-expressing the top equation as $\Delta y_t = -\delta_1 y_{t-1} + \delta_1\delta_2 z_{1t-1} + 0.5\Delta y_{t-1} + \delta_3 \Delta z_{t-1} + \varepsilon_{1t}$ illustrates three possible sources of Granger-causality running from z_{1t} to y_t: the speed-of-adjustment term (δ_1), the cointegrating coefficient on z_{1t} (δ_2), and the lagged first differences (δ_3), the last of which is also examined in the simulations of Dolado and Lütkepohl (1996). Due to the non-linearity, in this case the null hypothesis is given by $H_0 : \delta_1\delta_2 = \delta_3 = 0$. Because we expect this distinction to matter, we consider test power along all three dimensions.

4.1.2 Models with near-unit-root/local-to-unity

We define $c \leq 0$ as the local-to-unity coefficient and $a_T = 1 + c/T$. In this case there is no equivalent to the stationary model in levels (levels-VAR) considered above. However, similar to the difference VAR above, we include a model with non-cointegrated near unit roots (no-cointegration):

$$\begin{bmatrix} \Delta y_t \\ \Delta z_{1t} \end{bmatrix} = \begin{bmatrix} a_T - 1 & 0 \\ 0 & a_T - 1 \end{bmatrix} \begin{bmatrix} y_{t-1} \\ z_{1t-1} \end{bmatrix} + \begin{bmatrix} 0.5 & \delta \\ 0.3 & 0.5 \end{bmatrix} \begin{bmatrix} \Delta y_{t-1} \\ \Delta z_{1t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{bmatrix}. \tag{18}$$

We also consider two models that allow for cointegration between near unit roots. In the first model (z-adjusts):

$$\begin{bmatrix} \Delta y_t \\ \Delta z_{1t} \end{bmatrix} = \begin{bmatrix} a_T - 1 & 0 \\ a_T & -1 \end{bmatrix} \begin{bmatrix} y_{t-1} \\ z_{1t-1} \end{bmatrix} + \begin{bmatrix} 0.5 & \delta \\ 0.3 & 0.5 \end{bmatrix} \begin{bmatrix} \Delta y_{t-1} \\ \Delta z_{1t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{bmatrix} \tag{19}$$

(y_t, z_{1t}) have cointegrating vector $(1, -1)$ and z_{1t} may be seen as the variable that adjusts to restore long-run equilibrium. In the second model (y-adjusts), specified by,

$$\begin{bmatrix} \Delta y_t \\ \Delta z_{1t} \end{bmatrix} = \begin{bmatrix} -1 & a_T\delta_1 \\ 0 & c/T \end{bmatrix} \begin{bmatrix} y_{t-1} \\ z_{1t-1} \end{bmatrix} + \begin{bmatrix} 0.5 & \delta_2 \\ 0.3 & 0.5 \end{bmatrix} \begin{bmatrix} \Delta y_{t-1} \\ \Delta z_{1t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{bmatrix}, \tag{20}$$

[13] 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).

it is y_t that performs this adjustment and therefore the cointegrating vector $(-1, \delta_1)$ specifies an alternative to the null model of Granger noncausality. Both cointegrating specifications are specializations of (Elliott, 1998, eq. 2), who considers inference on cointegrating parameters, but not causality testing, in near-integrated models.

4.1.3 Models involving fractional integration in z_{1t}

In this case, we assume that z_{1t} is fractionally integrated of order d and model it as:

$$z_{1t} = (1-L)^{-d} \varepsilon_{2t} \tag{21}$$

for $0 < d < 1$. Under the null hypothesis we consider cases in which y_t is I(0) and cases in which y_t is I(1). We do not include cases in which y_t is I(d) and fractionally cointegrated with z_{1t} because this is not a case in which our surplus lag method, which relies on finite order approximate models under the null hypothesis, can be expected to work. However, under the alternative we may allow for all three cases: y_t I(0), y_t I(1), and y_t I(d) and fractionally cointegrated with z_{1t}.

To be concrete, we consider two models. In the first model:

$$\Delta y_t \;=\; 0.5\Delta y_{t-1} + \delta \Delta z_{1t-1} + \varepsilon_{1t}, \tag{22}$$

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. In the second model we consider

$$\Delta y_t = -0.5(y_{t-1} - \delta_1 z_{1t-1}) + \delta_2 \Delta z_{t-1} + \varepsilon_{1t}, \tag{23}$$

in which y_t is I(0) under the null hypothesis (H_0: $\delta_1 = \delta_2 = 0$). Under the alternative, y_t may be either I(0) $(\delta_1 = 0, \delta_2 \neq 0)$ or I(d) and cointegrated with z_{1t} (with cointegrating vector $(1, -\delta_1)$).

4.2 Test procedures

Under each of these models we compare four simple methods of testing Granger noncausality. The first two tests, based on a VAR(2) in levels (Levels-VAR) and a VAR(1) in differences (Dif-VAR), are included as a basis of comparison. They are both based on standard Wald tests for Granger causality that do not employ the surplus lag methodology. The levels VAR tests the null restriction $H_0 : A_{12}(1) = A_{12}(2) = 0$ using the fitted AR(2) model $(\hat{y}_t, \hat{z}_{1t})' = \hat{A}(0) + \sum_{i=1}^{2} \hat{A}(i)(y_{t-i}, z_{1t-i})'$. The difference VAR tests the restriction $H_0 : A_{12}(1) = 0$ in the fitted model $(\widehat{\Delta y}_t, \widehat{\Delta z}_{1t})' = \hat{A}(0) + \hat{A}(1)(\Delta y_{t-1}, \Delta z_{1t-1})$. In both cases all relevant lags are tested. A levels VAR is an appropriate specification under (15), while the difference VAR is correct under (16). However, as we do not assume a priori knowledge of integration orders, we consider the behavior of both tests under both data generating processes (hereafter DGPs).[14]

[14] The lag order of the levels-VAR ($p = 2$) is chosen to be large enough to accommodate (16), when rewritten as a nonstationary VAR(2) in levels. When applying the levels-VAR(2) to (16), this choice of $p = 2$ allows us to observe the effect of misspecifying the order of integration without misspecifying the lag order.

The third test (Toda-Phillips), an implementation of the method suggested by Toda and Phillips (1993), is a causality test based on a vector error correction model, with pre-tests for unit roots and cointegration rank. Unlike the two previous tests, this test does not rely on prior information regarding integration orders and cointegration rank and thus provides a far more serious and difficult benchmark than the two previous tests.

The final test (Surplus-VARX), based on an ARX(2,3), is an example of the surplus lag VARX causality test analyzed in theoretical sections above. In contrast to the previous three approaches, the forcing variable z_{1t} is not explicitly modelled. Instead, we estimate the autoregressive distributive lag or ARX model $\hat{y}_{1t} = \hat{a}(0) + \sum_{i=1}^{2} \hat{a}(i)y_{1t-i} + \sum_{i=1}^{3} \hat{b}(i)z_{1t-i}$, in which z_{1t} is an unmodelled forcing process, and test $H_0 : b(1) = b(2) = 0$. The implication of Granger noncausality for the surplus lag $b(3)$ is not tested. Because we do not require any extra surplus lags of y_{1t} we base the test on a surplus lag ARX(2,3) rather than a surplus-lag ARX(3,3).

4.3 Rejection rates under the null

Since we are interested in the robustness of these procedures to possible misspecification of the order of integration, we consider the application of each of these four methods under the null hypothesis corresponding to each of the data generating processes considered in Section 4.1 above. The results are divided into three tables. Table 1 provides null rejection rates for the I(0), I(1) and cointegrated models of Section 4.1.1, Table 2 presents results for the local-to-unity specifications of 4.1.2, and Table 3 gives the rejection rates for the fractionally integrated model of Section 4.1.3. In all cases the error processes are described by (14), with $\sigma_{12} = -0.8$. Within each panel we present results for sample sizes of $T = 50$, 100, 200, and 500. The table entries show finite sample rejection rates under the null hypotheses for a five-percent nominal test based on one thousand simulations.

We turn first to Table 1. Columns 3-4 provides null rejection rates for data simulated under the stationary levels VAR (eq. 15, $\delta = 0$) and the difference VAR (eq. 16, $\delta = 0$), respectively. Columns 5-6 are both generated under the error correction model (eq. 17, $\delta_3 = 0$), which has cointegrating vector $(1, -\delta_2)'$. In Column 5, $(\delta_1, \delta_2) = (1, 0)$, a null hypothesis under which there is no cointegration between y_t and z_{1t}. In Column 6, $(\delta_1, \delta_2) = (0, 1)$, implying cointegration but not Granger causality from z_{1t} to y_t.

As expected, the levels VAR (top panel) has good size when the DGP is given by the stationary VAR (Column 3). It also gives reasonable size under the two error correction model specifications (Columns 5-6). However, it suffers from non-trivial size distortion when the true model is a VAR in first differences (Column 4). Similarly, the difference VAR (Panel 2) has good size when the true model is a difference VAR in levels (Column 4) but suffers serious size-distortion when the true model is stationary in levels (Column 5). This is due to over-differencing and also results in size distortion in the VECM specification of Column 5, in which Δy_t is again over-differenced. By contrast, both the Toda-Phillips and surplus-lag procedures show reasonable rejection rates across all four specifications for sample sizes of $T = 200$ or larger. For both tests some size distortion is nonetheless observed in smaller samples, particularly in Columns 2 and 3.

Table 2 shows simulation results under the local-to-unity DGPs of Section 4.1.2, using a local-to-unity parameter of $c = -5.0$. The results in Columns 3-5, correspond to simulation models (18), (19), and model (20), respectively, with $\delta = 0$. The levels VAR is moderately over-sized in model (18), in which both variables are quasi-differenced, a DGP that specializes to the difference VAR for $c = 0$ ($a_T = 1$). On the other hand, it shows reasonable accurate size for the specifications of Columns 4-5, based on models that allow for cointegration between near unit roots. The difference specification gives good results under the DGPs of Columns 3-4, but as shown in Column 5 is terribly over-sized in model (20). Although the Toda-Phillips procedure was not specifically designed to work in a local-to-unity setting, it nonetheless provides quite good size under model (19), as seen in Column 4. On the other hand, it is subject to size distortion under models (18), and (20) respectively. This may be due in part to difficulties associated with unit root and cointegration pre-tests in local-to-unity models and we found that the distortion appears only to occur in the case of residual cross-correlation (i.e. $\sigma_{12} \neq 0$). The surplus-lag VARX also suffers from some small sample size distortion, particularly for $T = 50$, but provides quite accurate size across all three models in larger sample sizes.

Finally, in Table 3, we consider the empirical size of the causality tests under the fractionally integrated models of Section 4.1.3, in which z_{1t} is generated by (21) using values of $d = 0.4$ (stationary long-memory) in Columns 3 and 5 and $d = 0.8$ (nonstationary fractional integration) in Columns 4 and 6. In Columns 3-4, the dependent variable, y_t, is generated by (22), with $\delta = 0$, a model without cointegration. In Columns 5-6 y_t is given by (23), with $\delta_1 = \delta_2 = 0$, a model allowing fractional cointegration under the alternative.

All four methods show only moderate size distortion in larger samples when z_{1t} has stationary long-memory (Columns 3 and 5). Recall that our theoretical results did not in this case, require the addition of a surplus lag. On the other hand, some difficulties are again encountered when z_{1t} is modeled as a nonstationary fractionally integrated process in Columns 4 and 6. The levels VAR shows size distortion in both models for $d = 0.8$. While the difference VAR works fine in the model without cointegration (Column 3), it again shows severe distortion in model (23) (Column 5). The Toda-Phillips is slightly over-sized in both models. The surplus-lag VARX shows size distortion in small samples, but is again the only test to show reasonable size across all models in the larger samples.

To summarize briefly, both the levels and difference VAR specifications can show substantial size distortion when misspecified. In the case of the levels VAR this is likely due to the use of standard critical values when nonstandard asymptotics apply. The difference VAR is misspecified in cases of over-differencing and these cases result in particularly large size distortions. While these results are not surprising, they underline the importance of inference methods that are more robust to differing orders of integrations. Both the Toda-Phillips and surplus-VARX model fit this description and both tests have good size against all I(0), I(1) and cointegrated models. Despite the fact that the Toda-Phillips procedure is not designed for the local-to-unity or fractionally integrated models, its size was still quite accurate across a number of these specifications. Nevertheless, some size distortions were detected under certain local-unity and nonstationary fractionally integrated models. Of the four models, only the surplus-VARX based test provided reasonably accurate rejection rates in moderate sample sizes across all specifications considered. This underlines its value as a causality test that is particularly robust to misspecification of integration orders. Of course this robustness does not come without cost. As we discussed earlier, the addition of an unnecessary lag may be expected to reduce test power. The magnitude of this power loss under various specifications is examined below.

4.4 Size-adjusted power

We next consider the behavior of the tests under the alternative hypothesis of Granger causality ($\delta \neq 0$ in most cases). Due to the large number of models and alternative specifications in 4.1 we present only the results for the alternatives discussed in Section 4.1.1 to save space. The full set of results are available upon request.

Figures 1-5 display the size-adjusted power curves for all four tests, under the alternative specifications of Section 4.1.1. Due to the large size distortions observed in the level and difference VARs, when misspecified, we compare size-adjusted power (defined here as power - actual size + nominal size) rather than power itself. In a few cases we also omitted the difference VAR test due to its extreme distortion.

Figure 1 is generated under the levels VAR in (15). The misspecified difference VAR appears biased and sized distorted. However, the size-adjusted power of the other three tests are quite similar. In this case there appears to have been relatively little power loss from use of the surplus lag. In Figure 2, the data is simulated from the difference VAR specification given in (16). Here, the now correctly specified difference VAR has substantially higher power than the other three methods. However, the size-adjusted power of the surplus VARX is sometimes better and never much worse than that of the other two tests.

The power loss due to the surplus lag becomes more evident in the VECM model, which allows for the possibility of cointegration between y_t and z_{1t}. Figures 3, 4, and 5 show results for three different alternative specifications in (17). In figure 3, we set $(\delta_1, \delta_2) = (0, 1)$ and vary δ_3 across the horizontal axis. In this model y_t and z_{1t} are cointegrated under both the null and alternative and Granger causality is controlled through the coefficients on the lagged first differences. In figure 4, $(\delta_1, \delta_3) = (1, 0)$ and we vary δ_2. In this case, there is no cointegration between y_t and z_{1t} under the null hypothesis $\delta_2 = 0$, but there is cointegration under the alternative. Finally, in Figure 5, we set $(\delta_2, \delta_3) = (1, 0)$ and trace the power curve by varying δ_1, the speed-of-adjustment parameter for y_t. In all three cases the size-adjusted power of the surplus-lag causality test lies substantially below that of both the levels VAR and Toda-Phillips test, at least until power reaches close to one.

4.5 A brief comparison of two surplus lag approaches

We next provide a few comparisons of the two variants of the surplus-lag causality tests: the VAR based approach considered in the previous literature (Dolado and Lütkepohl, 1996; Toda and Yamamoto, 1995; Saikkonen and Lütkepohl, 1996) and the VARX based approach studied above. The first involves an addition of one extra, untested, lag to each variable for a joint VAR for y_t and z_{1t}, whereas in the second approach only y_t is modelled by means of a VAR and only z_{1t} requires an extra surplus lag. As seen in the theoretical section above, it is not in fact necessary to explicitly model z_{1t} in order to conduct a test of causality. Therefore, although the two approaches are similar in spirit, one may expect certain advantages from the parsimony of VARX based approach. More concretely, this may be expected to give rise to the following two benefits:

1. The surplus-lag VARX requires only a surplus lag of z_{1t}, whereas the surplus lag VAR employs a surplus lag of both y_t and z_{1t}.

2. In the surplus lag VAR, the lag length must be chosen large enough to approximate the dynamics of both y_t and z_{1t}, whereas in the surplus lag VARX it needs only be chosen large enough to accommodate the dynamics of y_t (including its possible dependence on past z_{1t}).

In our Monte-Carlo experiments, we have generally found that the first of these two advantages leads to relatively minor gains, whereas as the second advantage listed above can be quite important, particularly when the dynamics of z_{1t} are complicated relative to those of y_t. This is not an uncommon circumstance in practice. Stock returns, for example, have simple mean-dynamics, whereas the variables often used to predict them, such as earnings and dividend price ratios, are highly persistent with strong seasonal dynamics. Likewise, it is common to attempt prediction of exchange rate returns using the forward premium, which shows long-memory characteristics.

To demonstrate this point as simply as possible, consider a stationary seasonal VAR model, in which z_{1t} has a seasonal autoregressive component, but y_t does not:

$$\begin{bmatrix} y_t \\ z_{1t} \end{bmatrix} = \begin{bmatrix} 0.5 & \delta \\ 0.3 & 0.5 \end{bmatrix} \begin{bmatrix} y_{t-1} \\ z_{1t-1} \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & a_{22}(s) \end{bmatrix} \begin{bmatrix} y_{t-s} \\ z_{1t-s} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{bmatrix}. \qquad (24)$$

In this case, the surplus-lag VAR would be implemented by estimating the VAR$(s+1)$ model, $(\hat{y}_t, \hat{z}_{1t})' = \hat{A}(0) + \sum_{i=1}^{s+1} \hat{A}(i)(y_{t-i}, z_{1t-i})'$, and testing the relevant elements of the first s autoregressive coefficients, i.e. by testing $H_0 : A_{12}(i) = 0$ for $i = 1, \dots s$. On the other hand, when using the surplus lag VARX we need only estimate the ARX(1,2) given by $\hat{y}_{1t} = \hat{a}(0) + \hat{a}(1)y_{1t-1} + \sum_{i=1}^{2} \hat{b}(i)z_{1t-i}$ and test $H_0 : b(1) = 0$.

Figures 6 and 7 show the corresponding size-adjusted power curves for $s = 4$ (corresponding to seasonality in quarterly data) and $s = 12$ (corresponding to monthly data), respectively. In both cases we set $a_{22}(s) = 0.3$. These comparisons demonstrate well the potential gains that can be achieved from employing the more parsimonious surplus-VARX in place of the surplus-lag VAR. The size-adjusted power of the surplus-lag VARX (dashed green line) is considerably higher than that of the larger surplus-lag VAR (solid red line). This is particularly apparent at the monthly frequency ($s = 12$) in Figure 7.

While the case discussed above illustrates the relative advantages of the VARX version of the surplus-lag test, these gains are far more modest when the dynamics of both y_t and z_{1t} are simple. For example, if we eliminate the seasonal root in the above model, the surplus-lag VAR would simply be based on a test of $H_0 : a_{12}(1) = 0$ in an estimated VAR(2) model. This still requires the estimation of a slightly larger model than the ARX(1,2) in the surplus-lag VARX, but the difference in model sizes is now far smaller. The size-adjusted power curve corresponding to this experiment is shown in Figure 8. The surplus-lag VARX still shows marginally better power, but its advantage in this case is quite slight. Likewise, in most of the simple models considered in section 4.1 we found the small sample behavior of the surplus-lag VAR and VARX to be quite similar.

Although the simulations above are somewhat unrealistic in assuming known lag lengths, one may expect to find similar comparisons when employing model selection methods. For example, the model selected for the VAR is likely to be considerably larger when the true model is given by (24) than when it is given by (15), whereas the orders selected for the ARX would likely be small in both cases. A similar model selection outcome would be expected for the case in which z_{1t} has long-memory, but y_{1t} does not.

5 Conclusion

Employing a surplus lag in VAR based tests has been known to provide for inference which is invariant to possible $I(1)$ nonstationarity without necessitating unit root or cointegration pre-tests (Toda and Yamamoto, 1995; Dolado and Lütkepohl, 1996; Saikkonen and Lütkepohl, 1996). This provides for robust inference at some cost in terms of efficiency. On the other hand, there are arguably more efficient competing methods, which make fuller use of the I(0)/I(1) framework, without requiring knowledge on cointegration orders (Toda and Phillips, 1993; Kitamura and Phillips, 1997).

As our results demonstrate, the full advantage of the surplus approach becomes more apparent once one departs from both the pure VAR model and the I(0)/I(1) framework, of which it makes little explicit use, in order to allow for more general models of persistence. In particular, by applying the surplus lag to a VARX, in which the causing variables are exogenously modelled, we have shown that the same Chi-squared test statistic and critical values can be used to test Granger causality under a variety of possible data generating processes that may characterize the persistence in the forcing variable. These include the $I(0)$, $I(1)$ and cointegrated models considered earlier in the VAR context, as well as stationary and nonstationary long-memory, local-to-unity, and certain structural break models. In keeping with the earlier literature, no estimates of the long-memory parameter or first-stage confidence intervals on the local-to-unity parameter are required and the structural breaks need not be tested for or explicitly modelled. The VARX framework turns out to be particularly useful in allowing for long-memory and unmodelled structural breaks, which are not easily incorporated into a pure VAR. However, it is only in the context of the surplus lag that nonstationary processes, such as non-stationary fractionally integrated processes, can be accommodated without altering the limit distribution of the test statistic. Our simulation results suggest that this method works well in moderate sample sizes and can, in some circumstances, provide substantial power improvements over the use of the surplus lag test in a pure VAR model.

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

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

Campbell, B and J.-M Dufour (1997). Exact nonparametric tests of orthogonality and random walk in the presence of a drift parameter. *International Economic Review* **38**(1), 151–173.

Cavanagh, C. L, G Elliott and J. H Stock (1995). Inference in models with nearly integrated regressors. *Econometric Theory* **11**, 1131–1147.

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

Chan, N. H and W Palma (1998). State space modeling of long-memory processes. *The Annals of Statistics* **26**, 719–740.

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

Christiano, L, M Eichenbaum and R Vigfusson (2003). What happens after a technology shock. Mimeo, Northwestern University.

Davidson, J (1994). *Stochastic Limit Theory.* Oxford University Press.

Davidson, J and N Hashimzade (2007). Convergence to stochastic integrals with fractionally integrated processes: Theory, and applications to cointegrating regression. 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 and T Jouini (2005). Finite-sample simulation-based inference in VAR models with applications to order selection and causality testing. *Journal of Econometrics*. forthcoming.

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

Faust, J (1996). Near observational equivalence and theoretical size problems with unit root tests. *Econometric Theory* **12**, 724–731.

Faust, J (1999). Conventional confidence intervals for points on the spectrum have confidence level zero. *Econometrica* **67**, 629–37.

Friedman, B and K Kuttner (1992). Money, income, prices, and interest rates. *American Economic Reivew* **82**, 472–92.

Gali, J (1999). Technology, employment, and the business cycle: Do technology shocks explain aggregrate fluctuations?. *American Economic Review* **89**(1), 249–271.

Gourieroux, C and J Jasiak (2001). Memory and infrequent breaks. *Economics Letters* **70**, 29–41.

Granger, C. W and N Hyung (2004). Occasional structural breaks and long memory with an application to the S&P 500 absolute stock returns. *Journal of Empirical Finance* **11**, 399–421.

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.

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

Hannan, E. J (1976). The asymptotic dibribution of serial covariances. *Annals of Statistics* **4**(2), 396–399.

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

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.

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

Hosking, J. R. M (1996). Asymptotic distributions of the sample mean, autocovariances, and autocorrelations of long-memory time series. *Journal of Econometrics* **73**, 261–284.

Hualde, J (2006). Unbalanced cointegration. *Econometric Theory* **22**, 765–814.

Hualde, J and P. M Robinson (2007). Root-n-consistent estimation of weak fractional cointegration. *Journal of Econometrics* **140**, 450–484.

Jansson, M and M. J Moreira (2006). Optimal inference in regression models with nearly integrated regressors. *Econometrica* **74**, 681–714.

Jeganathan, P (1999). On asymptotic inference in cointegrated time series with fractionally integrated errors. *Econometric Theory* **18**, 1309–1335.

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.

Lanne, M (2002). Testing the predictability of stock returns. *The Review of Economics and Statistics* **84**(3), 407–415.

Lewellen, J (2004). Predicting returns with financial ratios. *Journal of Financial Economics* **74**(2), 209–235.

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 (1996). *Handbook of Matrices*. John Wiley and Sons.

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

Lütkepohl, H and P Saikkonen (1999). *Order Selection in Testing for the Cointegrating Rank of a VAR Process.* Oxford University Press. Engle, White (eds.): Cointegration, Causality and Forecasting - A Festschrift in honour of Clive W. J. Granger.

Marin, D (1992). Is the export-led growth hypothesis valid for industrialized countries?. *The Review of Economics and Statistics* **74**, 678–688.

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

Maynard, A and K Shimotsu (forthcoming). Covariance-based orthogonality tests for regressors with unknown persistence. *Econometric Theory.*

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.

Muller, U. K and M. W Watson (2007). Low-frequency robust cointegration testing. Mimeo, Princeton University.

Nabeya, S and B. E Sørensen (1994). Asymptotic distributions of the least-squares estimators and test statistics in the near unit root model with non-zero initial value and local drift and trend. *Econometric Theory* **10**, 937–966.

Palma, W and M Zevallos (2004). Analysis of the correlation structure of square time series. *Journal of Time Series Analysis* **25**, 529 – 550.

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 (1995). Fully modified least squares and vector autoregression. *Econometrica* **63**, 1023–1078.

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

Phillips, P. C. B (2005). Challenges of trending time series econometrics. *Mathematics and Computers in Simulation* **86**, 401–416.

Phillips, P. C. B and V Solo (1992). Asymptotics for linear processes. *Annals of Statistics* **20**, 971–1001.

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.

Robinson, P. M and D Marinucci (2003). Semiparametric frequency domain analysis of fractional cointegration. In: *Time Series With Long Memory* (P.M. Robinson, Ed.). pp. 334–373. Oxford University Press. Oxford.

Robinson, P. M and J Hualde (2003). Cointegration in fractional systems with unknown integration orders. *Econometrica* **71**, 1727–1766.

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 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.

Torous, W, R Valkanov and S Yan (2005). On predicting stock returns with nearly integrated explanatory variables. *Journal of Business* **78**(1), 937–966.

A 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:*

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

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

Proof: The proof uses the results of Theorem 1, 3 and 5 of Hosking (1996). The main difference is that we are interested in expressions uniformly in the lag whereas Hosking (1996) deals with fixed lags.
First note that using $\Omega := \mathbb{E}\varepsilon_t\varepsilon_t'$ we have for some constant $0 < K < \infty$ not depending on $j \in \mathbb{Z}$

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

since $\|\Omega\|_2 < C, \|\phi_{w,i}\|_2 \leq C_k i^{d-1}$ for some $K < \infty$ (see Lemma 2, (Palma and Zevallos, 2004)). 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

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

Note that $\mathbb{E}w_tw_sw_rw_0 = \gamma_{t-s}\gamma_r + \gamma_{t-r}\gamma_s + \gamma_t\gamma_{s-r} + \kappa_4(t,s,r)$ 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 $\mathbb{E}w_0^4 \leq M_4 < \infty$ since $\|\phi_{w,a}^4\|_2 = O(a^{4d-4}) = o(a^{-2})$. Next

$$T^{-2}\sum_{t,s=1+p}^{T}\mathbb{E}w_{t+j}w_tw_sw_{s+k} = T^{-2}\sum_{t,s=1+p}^{T}\gamma_j\gamma_k+\gamma_{t-s+j}\gamma_{t-s-k}+\gamma_{t+j-s-k}\gamma_{t-s}+\kappa_4(t-s,t-s+j,k). \tag{25}$$

The first term here is equal to $(T-p)^2T^{-2}\gamma_j\gamma_k = \mathbb{E}\hat{\gamma}_j\mathbb{E}\hat{\gamma}_k$ independent of the value of d.
The derivation of the bounds for the remaining terms in (25) will be done separately for the different cases for d. Thus let $0.25 < d < 0.5$ for the moment. The last term in (25) 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 $\|\gamma_l\|_2 \leq Kl^{2d-1}$ as shown above:

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

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

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

This holds for $d \neq 0.25$. For $d = 0.25$ the same argument shows the bound $O(\log T)$ (cf. Hosking, 1996, top of p. 278). For $j \leq 0$ 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 $\mathbb{E}|\Delta_{j,k}| \leq MT^{4d-2}$ for $0.25 < d < 0.5$.
For $d = 0.25$ the bound on the last term in (25) is identical to the case $0.25 < d < 0.5$. Further $\mathbb{E}|\Delta_{j,k}| \leq M(\log T)/T$ for $d = 0.25$ according to 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 $\mathbb{E}|\Delta_{j,k}| \leq M/T$ for $d < 0.25$. All bounds hold uniformly in $1 \leq j, k \leq p$ and $1 \leq p \leq T$. □
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 $\sup_p \mathbb{E}\kappa_{4,p} < \infty$. □

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 $p = o(T^{1/3})$)*

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

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

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

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

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 which can be dealt with using 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 $0 \leq j \leq T^{1/3}$.□

Lemma 4 *Let $v_{t,T} - A_Tv_{t-1,T} = u_t, t \in \mathbb{N}, A_T = I - diag(c_1,\dots,c_k)/T, c_i \geq 0$ 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 $\det\pi_u(z) \neq 0, |z| \leq 1$ 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 $w_t = \sum_{j=0}^{\infty}\phi^{\varepsilon}_{w,j}\varepsilon_{t-j} + \phi^{\eta}_{w,j}\eta_{t-j}$ 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 $\sum_{j=-\infty}^\infty \|\mathbb{E}u_0u_j'\|_2 < \infty$ 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 $i = 1, \dots k$, 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

$$|\sum_{i,j=0}^{t-1} \mathbb{E}u_{t-i}'(A_T^i)'A_T^j u_{t-j}| \leq \sum_{i,j=0}^{t-1} \|\mathbb{E}u_{t-i}u_{t-j}'\|_2 \leq t \sum_{j=-\infty}^{\infty} \|\mathbb{E}u_0u_j'\|_2 = O(t).$$

(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

$$\mathbb{E}v_{t,T}v_{s,T}w_tw_s = \mathbb{E}v_{t,T}v_{s,T}w_t^\varepsilon w_s^\varepsilon + \mathbb{E}v_{t,T}v_{s,T}\mathbb{E}w_t^\eta w_s^\eta \tag{26}$$

because $\mathbb{E}v_{t,T}v_{s,T}w_t^\varepsilon w_s^\eta = \mathbb{E}v_{t,T}v_{s,T}w_t^\varepsilon \mathbb{E}w_s^\eta = 0$. Therefore the evaluations can be given for w_t^ε and w_t^η separately. All expectations exist due to assumed finite fourth moments. The contribution to $\mathbb{E}\|T^{-3/2}\sum_{t=p+1}^T v_{t,T}w_t\|_2^2$ of the second term involving w_t^η can be bounded as

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

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

For the first term in (26), we use the Beveridge-Nelson decomposition (Phillips and Solo, 1992) $w_t^\varepsilon = \phi_w(1)\varepsilon_t + w_t^* - w_{t-1}^*$. The main strategy is to rewrite $\sum_{j=p+1}^T v_{t,T}w_t^\varepsilon$ as a sum of several terms and then 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, as is straightforward to verify. It follows that

$$\begin{aligned} T^{-3/2}\sum_{t=1+p}^T v_{t,T}w_t^\varepsilon &= T^{-3/2}\sum_{t=1+p}^T v_{t,T}\varepsilon_t\phi_w(1) + T^{-3/2}\sum_{t=1+p}^T v_{t,T}(w_t^* - w_{t-1}^*) \\ &= T^{-3/2}\sum_{t=1+p}^T v_{t,T}\varepsilon_t\phi_w(1) - T^{-3/2}\sum_{t=p}^{T-1}(v_{t+1,T} - v_{t,T})w_t^* \\ &\quad + T^{-3/2}v_{T,T}w_T^* - T^{-3/2}v_{p,T}w_p^* \;.. \end{aligned} \tag{27}$$

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 (27) 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 P3 (i)-(iv) 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,*

$$\mathbb{E}(\|\varepsilon_{yt,p} - \varepsilon_{yt}\|_2^2)^{1/2} \leq c\sum_{j=p}^{\infty}\|\pi_{v,j}\|_2 \tag{28}$$

Proof: Using (11) and the definition of $\pi_{w,j,T}$ above 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 (28) 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 (see Assumption P2) 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) matrices whose dimension g_T depends on the sample size T. Let $\hat{R}_T$ denote a sequence of random matrices such that $\|\hat{R}_T - R_T\|_2 = O_P(f_T)$ where $f_T \to 0$. Then if* $\sup_{T\in\mathbb{N}} \|R_T^{-1}\|_2 < \infty$ *a.s. it follows that $\|\hat{R}_T^{-1} - R_T^{-1}\|_2 = O_P(f_T)$.*

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

Lemma 7

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} A^{-1} & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} -A^{-1}B \\ I \end{bmatrix} \left[D - CA^{-1}B\right]^{-1} \begin{bmatrix} -CA^{-1} & I \end{bmatrix} \tag{29}$$

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: Note that $z_{1t} = z_{1t}^\nu + z_{1t}^\varepsilon$ where $z_{1t}^\nu = \nu_t + \sum_{j=1}^\infty \theta_j \nu_{t-j}$ and $z_{1t}^\varepsilon = \sum_{j=1}^\infty \phi_j \varepsilon_{t-j}$ are mutually independent. Consequently $\mathbb{E} z_{1t-i} z_{1t-j}' = \mathbb{E} z_{1t-i}^\nu (z_{1t-j}^\nu)' + \mathbb{E} z_{1t-i}^\varepsilon (z_{1t-j}^\varepsilon)'$. Letting x_{1t}^ε and x_{1t}^ν denote the components of x_{1t}^- generated from ε_t and ν_t respectively , it follows that

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

Clearly $0 \le \Gamma_p^\varepsilon, 0 \le \Gamma_p^\nu$. 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 $\mathbb{E}(x_{p,3}' z^\nu_{1t-p_{z1}-1} + x_{p,4}' x^\nu_{1t})(x_{p,3}' z^\nu_{1t-p_{z1}-1} + x_{p,4}' x^\nu_{1t})' \to 0$. 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. □

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 the noncausality hypothesis, $H_0:\gamma_{z1j}=0$ for all j, and for $T>max(c_i)$, (4) can be reformulated as*

$$\Delta y_t = \Psi_{0,p,T}(\gamma_\perp' w_{t-1}) + \sum_{j=1}^{p}\Xi_{j,p,T}v_{t-j} + \left(\sum_{j=1}^{p_{z1}+1}\psi_{z1j}\right)z_{1t-p_{z1}-1} + \sum_{j=1}^{p_{z1}}\psi_{z1j}(z_{1t-j}-z_{1t-p_{z1}-1}) + \varepsilon_{yt,p}, \tag{30}$$

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 $A_{T,w}=I$. pure unit roots ($A_{T,w}=I$). However, the derivations and notation differ.*

Proof: Using $[\psi_{yj},\psi_{z2j}]=-[I,0]\pi_{w,j,T}, j=1,\ldots,p-1, [\psi_{yp},\psi_{z2p}]=[I_s,0](\Gamma')^{-1}\pi_{v,p-1}\left(\gamma_\perp A'_{T,w},\right)'$ (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

$$\Delta y_t = [I:0]\left[-(\Gamma')^{-1}\begin{bmatrix}\gamma_\perp' w_{t-1}\\ v_{2,t-1}\end{bmatrix} - \sum_{j=1}^{p}\pi_{w,j,T}(\Gamma')^{-1}\begin{bmatrix}\gamma_\perp' w_{t-j}\\ v_{2,t-j}\end{bmatrix}\right] + \sum_{j=1}^{p_{z1}+1}\psi_{z1j}z_{1t-j} + \varepsilon_{yt,p}. \tag{31}$$

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 (31) 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 (30). Since, by using (11) 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). $\sup_{p,T}(\sum_{j=1}^{\infty}\|\Xi_{1,j,p,T}\|_2)\leq[I:0]\sum_{j=1}^{\infty}\sum_{h=j+1}^{\infty}\|\pi_{w,h,T}\|_2(\Gamma')^{-1}[I:0]'<\infty$ and absolute summability of $\Xi_{2,j}$ both follow. □

B 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 $y_t = A(p)z_t^- + \varepsilon_{yt,p} = [A_1(p), A_2(p), A_3(p)][\left(z_{t,1}^-\right)', \left(z_{t,2,p}^-\right)', z_{3,t}']' + \varepsilon_{yt,p}$ where $z_t^- \in \mathbb{R}^{k_{zp}}$ is partitioned as $z_t^- = [\left(z_{1,t}^-\right)', \left(z_{2,t,p}^-\right)', z_{3,t}']'$ such that $z_{t,1}^- = \left[z_{t-1,1}', \dots, z_{t-p_1,1}'\right]' \in \mathbb{R}^{k_{z1}}$ (where p_1 is fixed) and $z_{3,t} \in \mathbb{R}^{k_{z3}}$ do not depend on p and $z_{2,t,p} = \left[z_{2t-1}', \dots, z_{2t-p}'\right]'$ 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)

$$\max_{T\in\mathbb{N}} \lambda_{max}(\mathbb{E}R_T) = O(1) \quad , \quad \lambda_{max}(R_T) = O_P(1), \lambda_{max}(R_T^{-1}) = O_P(1), \qquad (32)$$

$$R_T = \begin{bmatrix} R_{1,1} & R_{T,1,2} & 0 \\ R_{T,2,1} & R_{T,2,2} & 0 \\ 0 & 0 & R_{T,3,3} \end{bmatrix}, \qquad (33)$$

$$\hat{R}_T := D_T \sum_{t=p+1}^{T} z_t^-(z_t^-)'D_T, \text{ such that } \|\hat{R}_T - R_T\|_2 = o_P(p^{-1/2}), \text{ and } \mathbb{E}\hat{R}_T = O(1) \text{ elementwise} \qquad (34)$$

$$\sup_{l\in\mathbb{R}^{k_{zp}}, \|l\|_2=1} T^{-1/2}\sum_{t=p+1}^{T} (\mathbb{E}\|l'D_T z_t^-\|_2^2)^{1/2} = O(1), \qquad (35)$$

$$vec\left[\sum_{t=p+1}^{T} \varepsilon_{yt}(z_t^-)'D_T R_T^{-1} \begin{pmatrix} I & 0 & 0 \end{pmatrix}'\right] \xrightarrow{d} Z, \qquad (36)$$

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[15]

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

Now

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

[15] 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

$$\begin{aligned}\mathbb{E}\|\sum_{t=p+1}^{T}(\varepsilon_{yt,p}-\varepsilon_{yt})(z_t^-)'D_T\|_2 &\leq \sum_{t=p+1}^{T}(\mathbb{E}\|\varepsilon_{yt,p}-\varepsilon_{yt}\|_2^2)^{1/2}(\mathbb{E}\|D_T(z_t^-)\|_2^2)^{1/2} \\ &= (T^{1/2}\left(\mathbb{E}\|\varepsilon_{y1,p}-\varepsilon_{y1}\|_2^2\right)^{1/2})\left(T^{-1/2}\sum_{t=p+1}^{T}(\mathbb{E}\|D_T(z_t^-)\|_2^2)^{1/2}\right) = o(p^{1/2}). \quad (37)\end{aligned}$$

Here (35) and Lemma 5 are used. Moreover letting $\varepsilon_{yt(i)}, i=1,\dots,k_y$, 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, we obtain $\|(\hat{A}(p)-A(p))D_T^{-1}\|_2 = O_P(p^{1/2})$ using (35) and (32, 34) in combination with Lemma 6.

Then consider $\hat{\Sigma}_\varepsilon := T^{-1}\sum_{t=p+1}^{T}\hat{\varepsilon}_t\hat{\varepsilon}_t'$: We obtain

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

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).

With respect to HL (ii) note that $\hat{\Gamma}_{1.2}^{-1}$ equals the (1,1) block of $\hat{R}_T^{-1}$. Then (34) in combination with the bound on the norm of R_T and R_T^{-1} given in (32) 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 (37) and $\sum_{t=p+1}^{T} \varepsilon_{yt}(z_t^-)' D_T = O_P(p^{1/2})$ as used above. Then (36) and HL (ii) imply HL (iii). □

B.1 Proof of Theorem 2

Proof: The proof uses a number of results of Lewis and Reinsel (1985), henceforth called LR. 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 $k_{zp} = k_{z1}p_{z1} + p(k_y + k_{z2}) + k_{z1}$. Consequently $D_T = T^{-1/2}I$. Also, the assumptions of the theorem imply that all occurring variables are stationary with bounded variance. Then $\mathbb{E}\hat{R}_T = (T-p)/T R_T$. 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 (32), (33) and (35).

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 $O(T^{-2})$. This shows that 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 (34).

Finally (36) 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. □

B.2 Proof of Theorem 3

The proof is omitted because it is a special case of Theorem 4 provided below in which $A_{T,w} = I$ and $A_{T,z} = I$.

B.3 Proof of Theorem 4

Proof: The proof builds on (and generalizes) the results of Saikkonen and Lütkepohl (1996), henceforth denoted as SP96. Essential in the developments is the reparameterization of the auxiliary model (4) using (30). Since we are interested in testing the significance of $\psi_{z1j}, j = 0, \ldots, p_{z1}$ and not in the other parameters the reparameterization is immaterial to our purposes since the estimates of ψ_{z1j} in both, the original model and equation (30), coincide.

Note that in (30) there are two variables containing nonstationary regressors: $(\gamma_\perp' w_{t-1})$

and $z_{1t-p_{z1}-1}$. Assumption P3 allows for the existence of 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 $\beta'\beta_\perp = 0$[16] 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 allowing for no cointegrating relation. Thus instead of the original regression, we can consider the regression

$$\Delta y_t = [\psi_{x1}, \tilde{\psi}_{x2}, \tilde{\Psi}_0] \left[\begin{array}{c} z_{1,t}^- \\ z_{2,t,p}^- \\ z_{3,t} \end{array} \right] + \varepsilon_{yt,p} = A(p,T) z_t^- + \varepsilon_{yt,p},$$

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 $A_{T,z}^{p_{z1}-j+1} - 1 = O(T^{-1})$. Therefore it follows from Lemma 4 that the second term does not influence any of the results results. Thus, it is sufficient to verify the conditions of Lemma 10.

Define $D_T := \text{diag}(T^{-1/2}I, T^{-1/2}I, T^{-1}I)$, with partitioning corresponding to the partitioning of z_t^- into $z_{1,t}^-, z_{2,t,p}^-$ and $z_{3,t}$ and let $\hat{R}_T := D_T(\sum_{t=p+1}^T z_t^-(z_t^-)')D_T$. Note that in z_t^- the last k_{z3} coordinates are integrated, whereas the rest are stationary, apart from lower order remainders. Further let

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

such that obviously (33) 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 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), $c = 0$) 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 (32). $\mathbb{E}\hat{R}_T = O(1)$ 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.

Lemma 1 for $d = 0$ and Lemma 4 (ii) imply that each entry in $\hat{R}_T - R_T$ has variance uniformly of order $O(T^{-1})$. Accordingly $\|\hat{R}_T - R_T\|_2 = O_P(p/T^{-1/2})$ establishing (34) for $p = o(T^{1/3})$.

Next 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 according to the partitioning of z_t^-. According to Lemma 4

[16] 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 $n_z = 0$.

(i), we have $\mathbb{E}\|z_{3,t}\|_2^2 = O(t)$. Due to stationarity of the remaining terms we have $\mathbb{E}\|l' D_T z_t^-\|_2^2 = O(T^{-1})$, analogously to the proof in Theorem 2. Therefore (35) 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 (36) holds. This concludes the proof. □

B.4 Proof of Theorem 5

Proof: The proof follows the same route as the proof of Theorem 2. The main difference lies in the fact that the impulse response sequence corresponding to z_{1t} is not summable. Note, however, that only the process (z_{1t}) has long-memory whereas y_t and z_{2t} remain short-memory processes.

Hence let $D_T = T^{-1/2} I$ and $R_T = \mathbb{E} z_t^- (z_t^-)'$, where z_t^- is defined as in the proof of Theorem 2. Accordingly $\hat{R}_T := T^{-1} \sum_{t=p+1}^T z_t^- (z_t^-)'$. In order 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, according to 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 $\mathbb{E}\hat{R}_T = (T-p)/T R_T$. Therefore the expectation of the sum of squared entries of $\hat{R}_T - R_T$ is of order $O(T^{4d-2}p + p^2 T^{-1})$ for $0.25 < d < 0.5$, of order $O(pT^{-1}\log T + p^2 T^{-1})$ for $d = 0.25$ and of order $O(pT^{-1} + p^2 T^{-1})$ for $d < 0.25$. Here we use the fact that there are only $O(p)$ terms involving the long-memory processes since y_t and z_{2t} are short memory processes contributing p^2 terms of order $O(T^{-1})$. Therefore, in order for $\|\hat{R}_T - R_T\|_2 = o_P(p^{-1/2})$ it is sufficient that $p^2 T^{4d-2} + p^3 T^{-1} \to 0$ for $0.25 < d < 0.5$, $(p^2 \log T + p^3)/T \to 0$ for $d = 0.25$, and in all other cases $p^3 T^{-1} \to 0$. This shows (34). The bounds (32) on the eigenvalues of R_T 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 (33) follows trivially. Stationarity and finite variances of $(z_{1t})_{t\in\mathbb{N}}$ implies (35) as in the proof of Theorem 2.

It remains to verify (36). 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 $0 < c < \inf_{p\in\mathbb{N}} \|\alpha_p\|_2 \le \sup_{p\in\mathbb{N}} \|\alpha_p\|_2 \le C$ 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 $(X_{Tt})_{1\le t\le T}$ is a martingale difference sequence with respect to the sigma field generated by $\varepsilon_s, \nu_s, s \le t$. In the following we will only deal 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

$$(i) \qquad \max_{1\le t\le T} |X_{Tt}| \xrightarrow{p} 0,$$

$$
\begin{aligned}
&(ii) \qquad \sum_{t=1}^{T} X_{Tt}^2 \xrightarrow{p} \eta^2 (\text{a constant}),\\
&(iii) \qquad \mathbb{E} \max_{1\le t\le T} X_{Tt}^2 \quad \text{is bounded in} \quad T.
\end{aligned}
$$

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

$$\mathbb{E} \max_{1\le t\le T} X_{Tt}^2 \le \sum_{t=1}^{T} \mathbb{E}X_{Tt}^2 \le M$$

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 $\alpha_p' z_t^- = \sum_{j=0}^{\infty}\phi_{p,j}^{\nu}\nu_{t-j} + \phi_{p,j}^{\varepsilon}\varepsilon_{t-j}$. Therefore it is sufficient to show that $\sup_p\sum_{j=0}^{\infty}\|[\phi_{p,j}^{\nu}, \phi_{p,j}^{\varepsilon}]\|_2^2 < \infty$. But this follows since since $\sup_p\|\alpha_p\|_2$ is bounded by assumption and for each of y_t, z_{1t} and z_{2t} the summability assumption holds, which is straightforward to verify. Uniform integrability then follows from Davidson (1994, Theorem 12.10.). Hence 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. □

B.5 Proof of Theorem 6

Proof: The proof of Theorem 6 combines the arguments from the proof of Theorems 3 and 5. Analogously to equation (30) we obtain

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

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 $1 \leq j < p_{z1}$. 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 remaining coordinates are stationary. Define $D_T := \text{diag}\left(T^{-1/2}I, T^{-(d_1+1)}, \ldots, T^{-(d_{c_{z1}}+1)}\right)$, $\hat{R}_T := D_T \sum_{t=p+1}^T z_t^-(z_t^-)' D_T$, and

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

Obviously (33) holds with this choice. The uniform bound on the eigenvalues of R_T follows as in the proof of Theorem 5 and from

$$\text{diag}\left(T^{-(d_1+1)}, \ldots, T^{-(d_{c_{z1}}+1)}\right) \sum_{t=p+1}^T z_{3,t} z_{3,t}' \text{diag}\left(T^{-(d_1+1)}, \ldots, T^{-(d_{c_{z1}}+1)}\right) \xrightarrow{d} \Xi \qquad (38)$$

where Ξ is an a.s. positive definite random variable by Lemma 3 (i). Consequently (32) holds.

Next we show that (34) 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 5, with $d_{\max} := \max(d_1, \ldots, d_{k_{z1}})$ replacing d: The upper bound on the increase of p as a function of the sample size 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 $T^{-(d_r+3/2)} \sum_{t=p+1}^T [z_{3,t}]_r [(\beta' z_{1t-j})']_s = O_p(T^{\max(d_r+d_s,0)-dr-1/2})$ $T^{-(d_r+3/2)} \sum_{t=p+1}^T [z_{3,t}]_r [\Delta z_{1t-j}']_s = O_p(T^{\max(d_r+d_1,\ldots,d_r+d_{cz},0)-d_r-1/2})$ 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[17]

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

Therefore the sum over these terms is of order $O_P(pT^{d_{0,r}-d_r-1/2}) = o_p(p^{-1/2})$ since, by Assumption P5 (iii), we have both $p < T^{1/3}$, as needed for $0 \leq d_r < 1/2$, and $p < T^{2/3(1/2+d_r)}$ as needed for $-1/2 < d_r < 0$.

Further $\mathbb{E}[z_{3,t}]_r^2 = O(T^{2d_r+1})$ follows from Davidson and Hashimzade (2007). Thus (34) holds under the restrictions on p imposed in Assumption P5.

From (38) it also follows that the contribution of this block to $\mathbb{E}\|l' D_T z_t^-\|_2^2$ is $O(1)$, showing (35). Finally the arguments to show (36) are analogous to those used in the proof of Theorem 5 since the nonstationary components are not involved. This concludes the proof. □

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

B.6 Proof of Theorem 7

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 $D_T = T^{-1/2}I$. Define $S_j = \left\{p_{z1} + 2 + \lfloor \sum_{k=0}^{j-1} \omega_k T \rfloor, \ldots, \lfloor \sum_{k=1}^{j} \omega_k T \rfloor \right\}$ as the data range that would result if restricted to regime j only. S_j omits $p_{z1}+1$ discarded lags, denoted by $D_j := \left\{ \lfloor \sum_{k=0}^{j-1} \omega_k T \rfloor + 1, \ldots, \; p_{z1} + 1 + \lfloor \sum_{k=0}^{j-1} \omega_k T \rfloor \right\}$. 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 5 for $\Gamma(i)$ we have $\lambda_{\max}(\Gamma(i)), \lambda_{\max}\left(\Gamma(i)^{-1}\right) = O(1)$. 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

$$\begin{aligned}
\lambda_{max}(R) &\leq \sum_{j=1}^{J} \omega_j \lambda_{max}(\Gamma(j)) + \sum_{j=1}^{J} \omega_j \lambda_{max}\left((\mu(j) - \bar{\mu})(\mu(j) - \bar{\mu})' \right) = O(1) \text{ and} \\
\lambda_{max}\left(R^{-1}\right) &\leq \left(\sum_{j=1}^{J} \omega_j \lambda_{min}(\Gamma(j)) \right)^{-1} = O(1)
\end{aligned}$$

where J is fixed. This shows (32).

Next define sample counterparts:

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

and note that

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

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

$$\mathbb{E}\left\| (\bar{z}(j) - \mu(j)) \mathbf{I}(t \in S_j) \right\|_2^2 = (p_{z1} + 1) \sum_{i=1}^{k_{z1}} \mathbb{E}\left[\left(\bar{x}_{1ti}(j) - \mathbb{E}\left[x_{1ti} \mathbf{I}(t \in S_j) \right] \right)^2 \mathbf{I}(t \in S_j) \right]$$

$$+ \quad p \sum_{i=1}^{k_y+k_{z2}} \mathbb{E}\left[\left(\bar{x}_{2ti}(j) - \mathbb{E}\left[x_{2ti} \mathbf{I}\left(t \in S_j \right) \right] \right)^2 \mathbf{I}\left(t \in S_j \right) \right] = O\left(pT^{-1} \right). \tag{39}$$

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

$$\left\| \hat{R}(j) - R(j) \right\|_2 \quad \leq \quad \left\| \hat{\Gamma}(j) - \Gamma(j) \right\|_2 + O_P\left(pT^{-1} \right). \tag{40}$$

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

$$\hat{R} = \sum_{j=1}^{J} \omega_j \hat{R}(j) + \sum_{j=1}^{J} T^{-1} \sum_{t \in D_j} \left(z_t^- - \bar{z} \right) \left(z_t^- - \bar{z} \right)' = \sum_{j=1}^{J} \omega_j \hat{R}(j) + O_P\left(\sqrt{pT^{-1}} \right), \tag{41}$$

since $(p_{z1}+1)T^{-1}\mathbb{E}\left\| \left(z_t^- - \bar{z}\right)\left(z_t^- - \bar{z}\right)' \right\|_2 \leq (p_{z1}+1)T^{-1}\mathbb{E}\left[\left\| \left(z_t^- - \bar{z}\right) \right\|_2^2 \right] = O\left(pT^{-1}\right)$, where the last step follows by an argument similar to (39). Then by (40) and (41)

$$\left\| \hat{R} - R \right\|_2 \quad \leq \quad \sum_{j=1}^{J} \omega_j \left\| \hat{\Gamma}(j) - \Gamma(j) \right\|_2 + O_P\left(\sqrt{pT^{-1}} \right). \tag{42}$$

The same arguments as in the proofs of Theorems 2 and 5 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 (34). Next write

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

For the last term in (43) 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 (39). 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 (43) 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 (39). Finally the first term in (43) 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 5. This establishes (35).

As in the proof of Theorem 5, 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 (36) 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)

$$\sum_{t=1}^{T} X_{Tt}^2 = \mathbb{E}\left[\varepsilon_{ty}^2\right]\alpha_p'\hat{R}'\alpha_p + T^{-1}\sum_{t=1}^{T}\left(\varepsilon_{ty}^2 - \mathbb{E}\left[\varepsilon_{ty}^2\right]\right)\alpha_p'\left(z_t^- - \bar{\mu}\right)\left(z_t^- - \bar{\mu}\right)'\alpha_p \quad (43)$$

Note that $\left\|\hat{\Gamma}\left(j\right) - \Gamma(j)\right\|_2 \rightarrow_p 0$ by the same arguments as in Theorems 2 and 5 and therefore by (42), this implies that $\left\|\hat{R} - R\right\|_2 \rightarrow_p 0$, so that the first term in (43) converges in probability to $\tilde{\eta}^2 = \mathbb{E}\left[\varepsilon_{ty}^2\right]\alpha_p' R\alpha_p$. The second term in (43) converges in probability to zero by the same arguments as in the proof of Theorem 5 (Lemma 2 implies that $\mathbb{E}\left[\left(\alpha_p'\left(z_t^- - \mu(j)\right)\right)^4\right]$ and therefore $\mathbb{E}\left[\left(\alpha_p'\left(z_t^- - \mu\right)\right)^4\right]$ 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 5. □

Table 1: Rejection rates under the null hypothesis of Granger noncausality: stationary, nonstationary, and cointegrated models.

		Simulation DGP			
Method	sample size	Levels VAR(1) $(\delta = 0)$	Difference VAR(1) $(\delta = 0)$	VECM $\delta' = (1, 0, 0)$	VECM $\delta' = (0, 1, 0)$
Levels-VAR	50	0.089	0.211	0.124	0.069
	100	0.061	0.175	0.081	0.062
	200	0.057	0.140	0.069	0.045
	500	0.053	0.139	0.068	0.041
Dif-VAR	50	0.153	0.064	0.895	0.057
	100	0.227	0.058	0.999	0.064
	200	0.432	0.050	1.000	0.061
	500	0.735	0.064	1.000	0.043
Toda Phillips	50	0.104	0.110	0.152	0.067
	100	0.059	0.080	0.113	0.065
	200	0.047	0.071	0.086	0.046
	500	0.051	0.075	0.089	0.044
Surplus-VARX	50	0.097	0.137	0.099	0.068
	100	0.055	0.086	0.055	0.068
	200	0.058	0.060	0.047	0.044
	500	0.052	0.060	0.057	0.060

Table entries show empirical rejection rates under the null hypothesis for a nominal five percent test. In Columns 3-4, the data is generated by (eq. 15, with $\delta = 0$) and (eq. 16, with $\delta = 0$), respectively. In Columns 5 and 6 the data is generated by (eq. 17), with $(\delta_1, \delta_2, \delta_3) = (1, 0, 0)$ and $(\delta_1, \delta_2, \delta_3) = (0, 1, 0)$, respectively. In all cases the residuals are generated by (14) with $\sigma_{12} = -0.8$. In Panel 1 (levels-VAR), Granger-causality is tested using a VAR(2) in levels. Panels 2-4 employ tests based on a VAR(1) in first-differences, Toda and Phillips (1993), and the surplus-lag ARX(2,3), respectively.

Table 2: Rejection rates under the null hypothesis of Granger noncausality: local-to-unity models.

Method	sample size	Simulation DGP no-cointegration (18, $\delta = 0$)	 $z_{1t}-$adjusts; (19, $\delta = 0$)	 y_t-adjusts; (20, $\delta' = (0, 0)$)
	50	0.126	0.072	0.084
Levels-VAR	100	0.091	0.059	0.069
	200	0.079	0.050	0.074
	500	0.096	0.042	0.062
	50	0.103	0.074	0.336
Dif-VAR	100	0.059	0.069	0.692
	200	0.051	0.058	0.969
	500	0.055	0.047	1.000
	50	0.143	0.083	0.390
Toda-Phillips	100	0.130	0.072	0.326
	200	0.126	0.052	0.344
	500	0.130	0.049	0.332
	50	0.099	0.076	0.090
Surplus-VARX	100	0.058	0.057	0.052
	200	0.042	0.041	0.060
	500	0.066	0.057	0.059

Table entries show empirical rejection rates under the null hypothesis for a nominal five percent test. In Column 3, the data is generated by (eq. 18, with $\delta = 0$). In Columns 4-5 the data is generated from local-to-unity models with cointegration. In Column 4, it is generated by (eq. 19, with $\delta = 0$), in Column 5 it is simulated from (eq. 20, with $\delta_1 = \delta_2 = 0$). In all cases the local-to-unity (LTU) parameter is set to $c = -5.0$ and the residuals are generated by (14) with $\sigma_{12} = -0.8$. In Panel 1 (levels-VAR), Granger-causality is tested using a VAR(2) in levels. Panels 2-4 employ tests based on a VAR(1) in first-differences, Toda and Phillips (1993), and the surplus-lag ARX(2,3), respectively.

Table 3: Rejection rates under the null hypothesis of Granger noncausality: fractionally integrated models

		Simulation DGP			
Method		no cointegration		test cointegration	
		(eq. 22, $\delta = 0$)		(eq. 23, $\delta' = (0, 0)$)	
		$d = 0.4$	$d = 0.8$	$d = 0.4$	$d = 0.8$
Levels-VAR	50	0.100	0.155	0.097	0.164
	100	0.091	0.129	0.086	0.124
	200	0.083	0.111	0.061	0.136
	500	0.070	0.111	0.055	0.116
Dif-VAR	50	0.037	0.059	0.065	0.139
	100	0.045	0.052	0.061	0.244
	200	0.064	0.055	0.064	0.504
	500	0.049	0.054	0.058	0.920
Toda-Phillips	50	0.128	0.106	0.126	0.120
	100	0.108	0.073	0.072	0.100
	200	0.082	0.082	0.042	0.100
	500	0.067	0.090	0.048	0.129
Surplus-VARX	50	0.098	0.125	0.103	0.126
	100	0.092	0.093	0.087	0.080
	200	0.078	0.064	0.061	0.069
	500	0.070	0.068	0.057	0.052

Table entries show empirical rejection rates under the null hypothesis for a nominal five percent test. In Columns 3 and 4 the data is simulated from (eqs. 21 & 22, with $\delta = 0$), for $d = 0.4$ and $d = 0.8$, respectively. In Columns 5 and 6, the data is generated by (eqs. 21 & 23, with $\delta_1 = \delta_2 = 0$), for $d = 0.4$ and $d = 0.8$, respectively. In all cases the residuals are generated by (14) with $\sigma_{12} = -0.8$. In Panel 1 (levels-VAR), Granger-causality is tested using a VAR(2) in levels. Panels 2-4 employ tests based on a VAR(1) in first-differences, Toda and Phillips (1993), and the surplus-lag ARX(2,3), respectively.

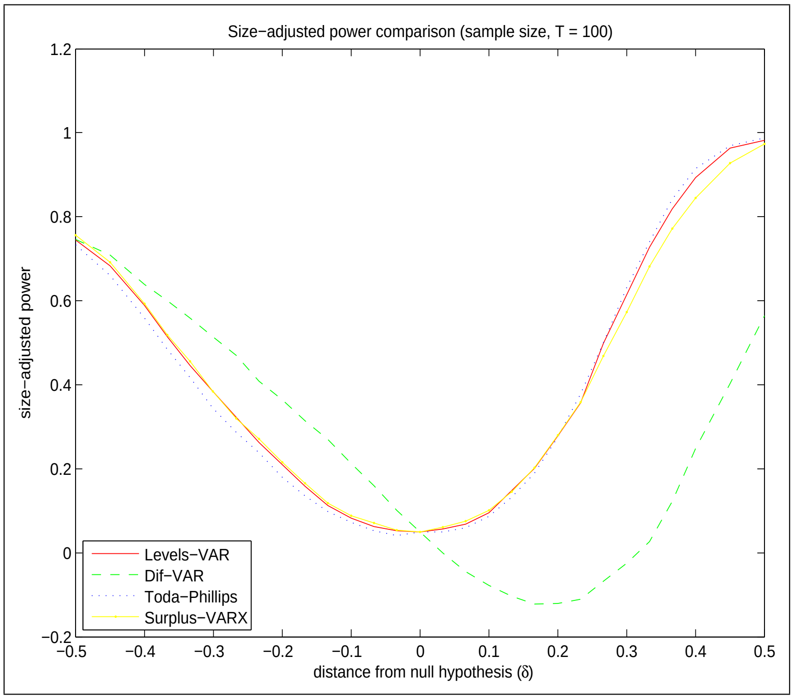

Figure 1: Size-Adjusted Power (power - actual size + nominal size): DGP is the stationary levels VAR in (15): I(0) series; $T = 100$.

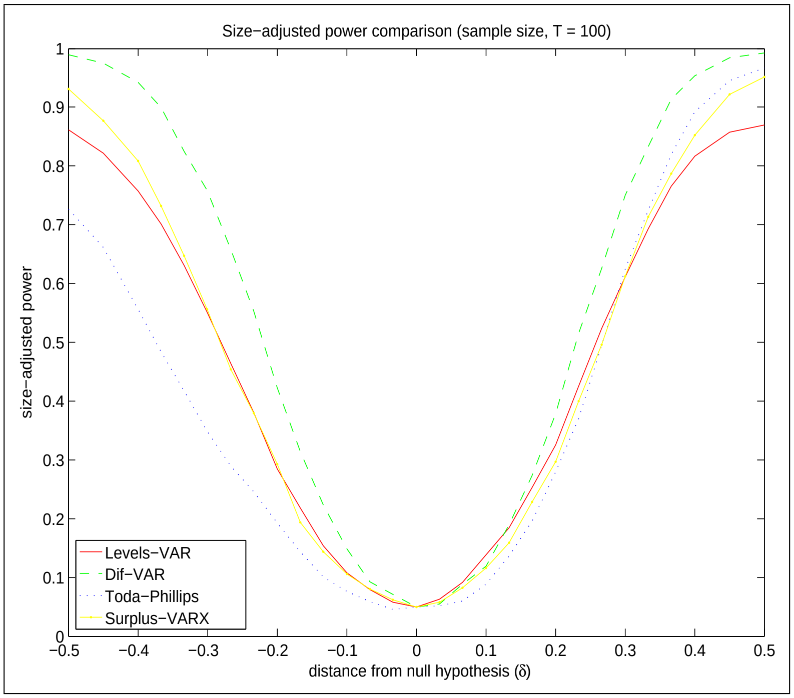

Figure 2: Size-Adjusted Power (power - actual size + nominal size): DGP is the VAR in differences in (16): Non-cointegrated I(1) series; $T = 100$.

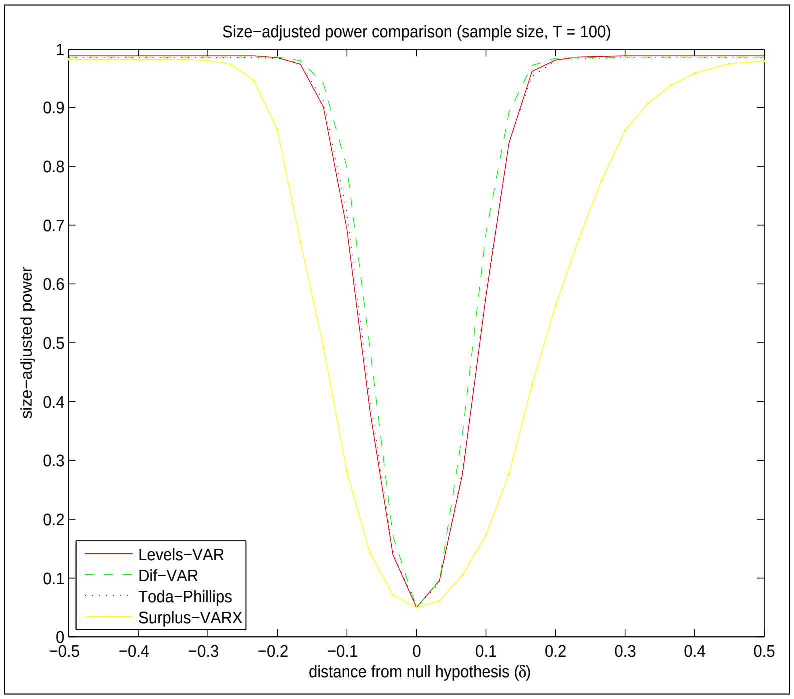

Figure 3: Size-Adjusted Power (power - actual size + nominal size): DGP is the VECM in (17) with $(\delta_1, \delta_2) = (0, 1)$ and δ_3 varying across the x-axis, $T = 100$.

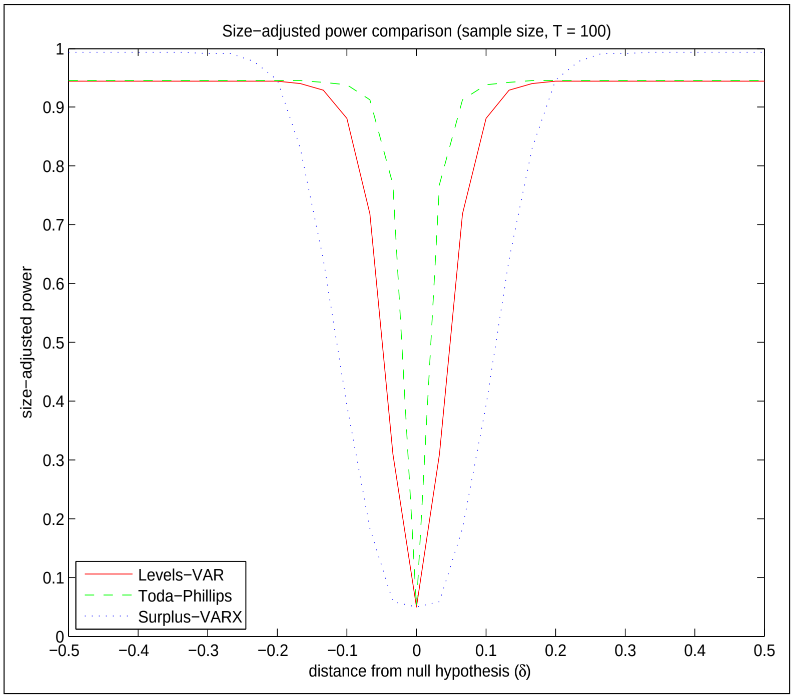

Figure 4: Size-Adjusted Power (power - actual size + nominal size): DGP is the VECM in (17) with $(\delta_1, \delta_3) = (1, 0)$ and $|\delta_2|$ varying across the x-axis; $T = 100$.

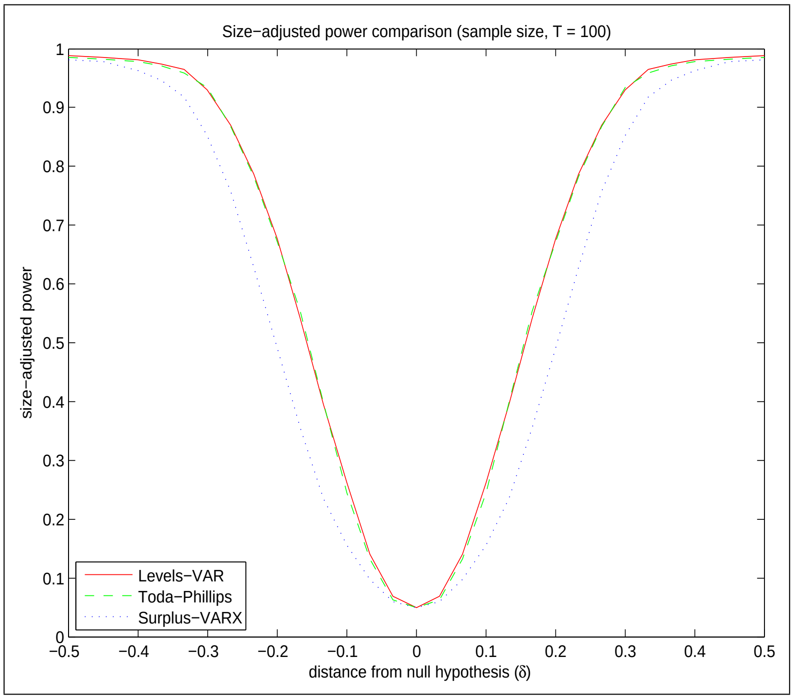

Figure 5: Size-Adjusted Power (power - actual size + nominal size): DGP is the VECM in (17) with $(\delta_2, \delta_3) = (1, 0)$ and $|\delta_1|$ varying across the x-axis; $T = 100$.

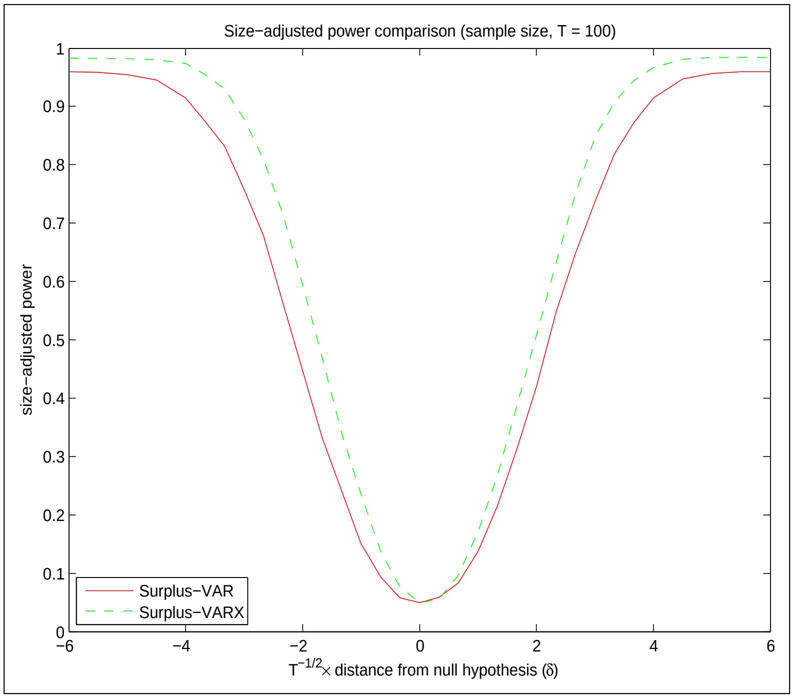

Figure 6: Size-Adjusted Power (power - actual size + nominal size): DGP is the quarterly seasonal levels VAR in (eq. 24, $s = 4$, $a_{22}(s) = 0.3$); $T = 100$.

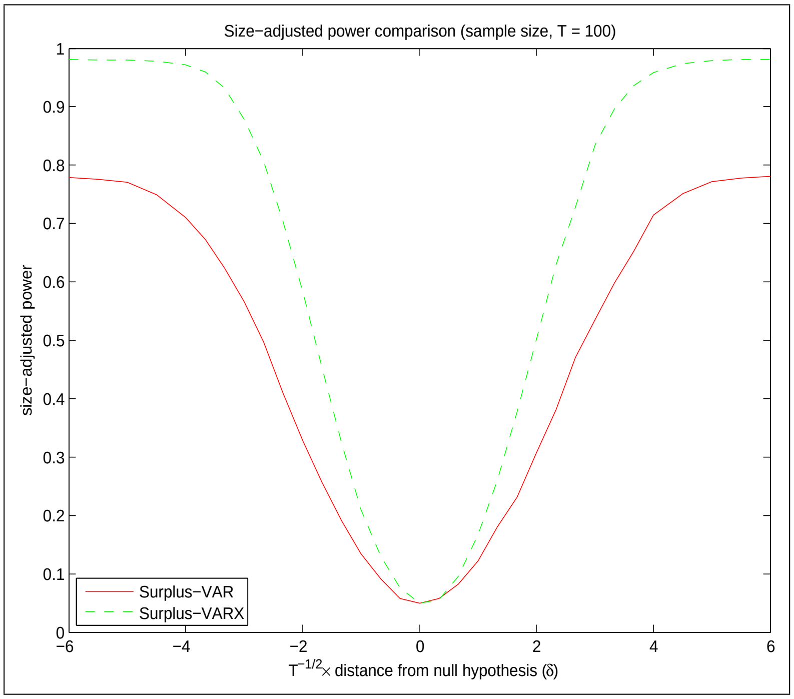

Figure 7: Size-Adjusted Power (power - actual size + nominal size): DGP is monthly seasonal levels VAR in (eq. 24, $s = 12$, $a_{22}(s) = 0.3$); $T = 100$.

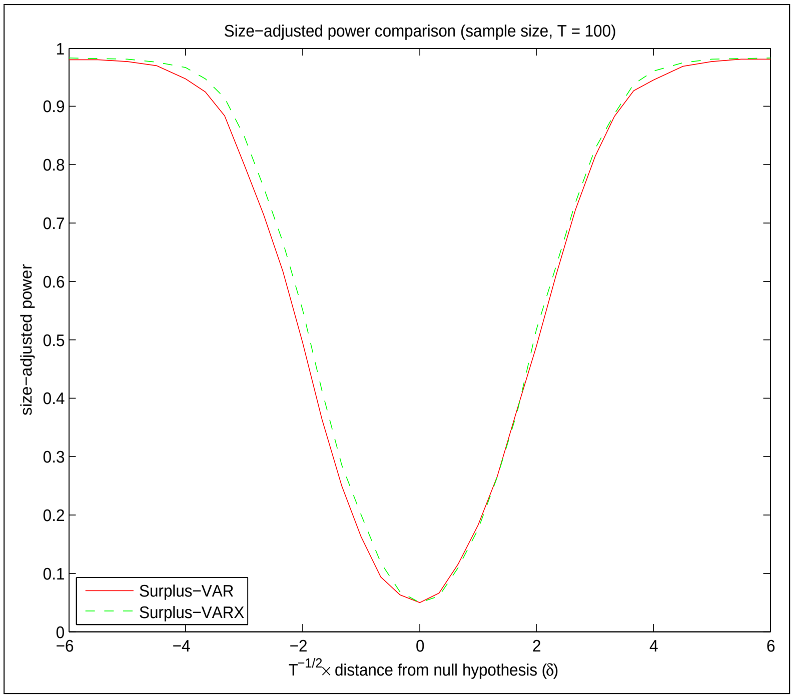

Figure 8: Size-Adjusted Power (power - actual size + nominal size): DGP is VAR(1) in levels without seasonal component (eq. 24 with $a_{22}(s) = 0$); $T = 100$.