Causal Change Detection in Possibly Integrated Systems: Revisiting the Money-Income Relationship[*]

Shuping Shi
Macquarie University

Stan Hurn
Queensland University of Technology

Peter C. B. Phillips
Yale University, University of Auckland,
University of Southampton & Singapore Management University

Abstract

This paper re-examines changes in the causal link between money and income in the United States over the past half century (1959 - 2014). Three methods for the data-driven discovery of change points in causal relationships are proposed, all of which can be implemented without prior detrending of the data. These methods are a forward recursive algorithm, a rolling window algorithm and a recursive evolving algorithm all of which utilize subsample tests of Granger causality within a lag-augmented vector autoregressive framework. The limit distributions for these subsample Wald tests are provided. Bootstrap methods are developed to control family-wise size in the implementation of the recursive testing algorithms. The results from a suite of simulation experiments suggest that the recursive evolving window algorithm provides the most reliable results, followed by the rolling window method. The forward expanding window procedure is shown to have the worst performance. Both the rolling window and recursive evolving approaches find evidence of Granger causality running from money to income during the Volcker period in the 1980s. The forward algorithm does not find any evidence of causality over the entire sample period.

Keywords: Time-varying Granger causality, subsample Wald tests, Money-income causality

JEL classification: C12, C15, C32, E47

1   Introduction

The question of whether the money supply Granger-causes (Granger, 1969) aggregate nominal income has been central to many discussions between major macroeconomic schools of thought over the last fifty years. This debate has also been a source of fascination for applied research starting with the paper of Sims (1972), which provided empirical evidence of a unidirectional causal relationship from money to income. Almost half a century after this seminal work and despite a proliferation of subsequent studies, a general consensus on the money-output question remains an elusive goal.[1]

The predominant methodology is the vector autoregressive approach with recent work augmenting the traditional VAR with a Markov switching mechanism (Psaradakis, *et al.* 2005). Yet even within the same VAR methodology there is the added problem, as noted by Stock and Watson (1989), that minor procedural differences are inescapably linked to arbitrary choices imposed by the researcher. Important points of difference include: data frequency (annual, quarterly, monthly); the treatment of the time-series properties of the data; the order of the VAR; the choice of monetary aggregate (M0, M1 and M2); the proxy used for economic activity (GDP, industrial production); and the control variables – the interest rate (Treasury bill rate, commercial paper rate) and inflation (CPI, WPI, GDP/GNP deflator).

In terms of the details of the construction of the VAR framework within which to conduct the tests of Granger causality, the literature offers little firm guidance. For the monetary aggregate, all studies invariably use M1 as one of the aggregates of choice, with the monetary base, M0, and broad money, M2, also being used. Swanson (1998) breaks with that tradition and includes Divisia M1 and M2 in the list of monetary aggregates used in the empirical analysis. The choice of proxies for real economic activity and prices are driven by the sampling frequency of the data: when the data are quarterly GDP or GNP and the associated deflator are used; when monthly industrial production data is used a price index is usually preferred. In terms of interest rates, replacing the Treasury Bill rate with a commercial paper rate or even an interest rate spread

---

[1] In this literature, causal relationships are defined as in Granger (1969), focusing on short term predictability. Subsequent developments have also considered long-horizon causality (Lütkepohl, 1993; Dufour and Renault, 1998; Dufour et al., 2006), which can arise from indirect links in a multivariate system. Long-horizon perspectives are of interest but have not received as much attention in the literature. Importantly for the present paper, long-run causal links are investigated under the maintained assumption that causal structure remains largely unchanged over the sample period, a condition that is at odds with the central thrust of the current research where causal links may change during the sample period. The long-run notion of causality is therefore not pursued here.

does not seem to yield any conclusive results.

Likewise, there is little consensus on the number of variables to include in the VAR. Christiano and Ljungqvist (1988) and Psaradakis, *et al.* (2005) report significant Granger-causality using bivariate money-income specifications. But the majority of empirical applications use a four-variable VAR that includes money, income, interest rates and prices. Friedman and Kuttner (1993) and Swanson (1998) also use a five variable VAR specification that includes two interest rates (the Treasury Bill rate and commercial paper rate), but there is no conclusive evidence that this model is superior to the more traditional four variable specification. In addition, the early literature (Thornton and Batten, 1985) found that tests for Granger causality were extremely sensitive to lag-length selection and advocated a thorough search of the lag space. This advice has been largely ignored in the subsequent literature with many studies arbitrarily fixing the number of lags at 6 or 12, Swanson (1998) being a notable exception where lag length choice is based on information criteria.

One recurring theme in the literature is subsample instability in the money-income relationship during the decade of the 1980s. Stock and Watson (1989) find that money has less predictive power for output for the sample period excluding data from the 1980s. By contrast, Friedman and Kuttner (1993) conclude that including data from the 1980s sharply weakens the significance of any relationship between money (however defined) and nominal output or between money and either real output or prices separately. Thoma (1994) shows that money Granger causes output only in the period of 1982-1987. Swanson (1998) finds that it is almost always present between 1960 and 1994. More recently Psaradakis, Ravn, and Sola (2005) conclude that money causes output in the first half of the 1980s but does not cause output in the second half of 1980s. They also find that money growth has more predictive power for output growth during recessions than during expansions.

Another important methodological issue to be resolved in developing a suitable testing framework is the role of trends, both deterministic and stochastic, in the analysis. Benanke (1986) observes that detrending data with deterministic trends enhances the explanatory power of money in output autoregressions. Writing around the same time, Eichenbaum and Singleton (1986) conclude that specifications using log-differenced data result in a small role for money in explaining output fluctuations. Sims (1987) argues that evidence of deterministic trends may be symptomatic of misspecification, which is ignored when the trends are removed. Similarly, Christiano and Ljungqvist (1988) suggest that differencing may also lead to specification error.

Stock and Watson (1989) advocate a careful methodological treatment of unit root behavior and time trends in the data. Their evidence suggests that output (industrial production), the money supply (M1), wholesale prices, and the 90-day Treasury-bill interest rate are characterised by independent unit roots, suggesting that previous inference using data expressed in levels may be inaccurate. They further conclude that the growth rate of the money stock contains a significant deterministic trend which should be removed (by quadratic detrending) in order to allow reliable inference on potential Granger-causality from money to income. But this approach has not been universally followed in the subsequent literature, more recent studies adopting a variety of specifications that include data in levels (Thoma, 1994), levels plus detrending (Hafer and Kutan, 1997), levels within a cointegrated VECM framework (Swanson, 1998) and first-differenced data (Psaradakis, *et al.*, 2005).

The often contradictory and sometimes confusing evidence does provide at least two general lessons. First, econometric methods that do not require choices of detrending or differencing to be made at the outset are potentially preferable to those that do. Second, causal relationships do change over time and links between money and output can be very sensitive to the sample period. Consequently, the approach taken in this paper is to utilize robust econometric methods that do not require choices of detrending or differencing to be made at the outset and which also explicitly allow for unknown change points in the causal relationships. The ultimate goal is to develop a testing procedure which allows for endogenously determined change points in any causal relation while at the same time treating trends, both deterministic and stochastic, in a way that does not require pretesting or prior removal of trend components, and to allow for potential heteroskedasticity in the testing process, an aspect which has largely been ignored in the existing literature.

One approach that might be expected to handle trends in causal testing would be to use reduced rank or VECM regression. Unfortunately, pre-testing for cointegrating rank inevitably produces size distortions and Granger causality tests suffers from nuisance parameter dependencies and nonstandard limit theory (Toda and Phillips, 1994). Alternative procedures that are applicable with such data are the fully modified VAR approach (Phillips, 1995) and the lag-augmented VAR (LA-VAR) approach (Dolado and Lütkepohl, 1996; Toda and Yamamoto, 1995). The Wald test statistics for both procedures follow standard chi-squared distribution. Yamada and Toda (1998) show that the LA-VAR test outperforms the fully modified VAR and the VECM approaches in terms of size stability, although fully modified VAR and VECM pro-

cedures generally have higher power than the LA-VAR test. In view of its superior size control properties, the LA-VAR approach is used in the causality tests proposed in the present paper. This method is specifically designed to be robust to the integration and cointegration properties of the time series used in the regressions and can therefore be applied without detailed or accurate prior knowledge of the presence (or absence) of unit roots.

Existing methods for dealing effectively with non-stability in causal relationships include the forward expanding window causality test (Thoma, 1994), the rolling window causality test (Swanson, 1998, Balcilar, Ozdemir, and Arslanturk, 2010, Arora and Shi, 2016, among others). While the forward and rolling window methods have been applied in empirical work, the asymptotic, finite sample, and relative performance properties of the these methods are relatively unexplored. In addition to exploring the asymptotic and finite-sample performance of these tests, this paper contributes a new time-varying Granger causality test based on a recursive evolving windowing procedure. The recursive evolving window approach was proposed in Phillips, Shi, and Yu (2015a, 2015b) for monitoring financial bubbles and is adapted here to detect Granger causality and possible changes in causal direction. The asymptotic properties of this test are developed and the performance of the LA-VAR based forward, rolling and recursive evolving approaches are examined and benchmarked in a simulation study. Since recursive testing methods all involve multiple testing, bootstrap methods are used to control the empirical size in their implementation.[2]

The empirical work in this paper focuses on the United States and ignores potential between-country differences.[3] In the present study, a traditional four variable VAR is chosen with M1 as the monetary aggregate and information criteria are used to aid the selection of lag length.

---

[2]Time-varying Granger causality tests are closely linked to the literature on instability tests for subsets of model parameters. These includes the Markov-switching Granger causality test of Psaradakis *et al.* (2005) and several tests proposed by Rossi (2005). See Rossi (2013) for a review. These methods are not implemented here because they are mainly ex-post testing procedures that are known to have difficulty in identifying breaks occurring at the end of the relevant sample. The focus instead is on tests that can be implemented in a real-time manner.

[3]The non-United States studies by, for example, Williams, Goodhart and Gowland (1976) and Mills and Wood (1978) (United Kingdom), Barth and Bennett 1974 (Canada), Komura (1982) (Japan) and Kamas and Joyce (1993) (India and Mexico) and the multi-country study by Krol and Ohanian (1990) (United Kingdom, West Germany, Canada and Japan) are all illustrative of the broad appeal of this problem. Of course, country differences are to be expected given the vastly differing institutional and policy settings.

2 Time-varying Granger causality tests

In order to conduct a Granger causality test for a possibly integrated variable, Toda and Yamamoto (1995) and Dolado and Lütkepohl (1996) suggest estimating a lag-augmented VAR model. As an illustration, consider the bivariate case (y_{1t} and y_{2t}) with a maximum order of integration d. The lag-augmented VAR model is

$$
\begin{aligned}
y_{1t} &= \alpha_{10} + \alpha_{11}t + \sum_{i=1}^{k+d} \beta_{1i} y_{1t-i} + \sum_{i=1}^{k+d} \delta_{1i} y_{2t-i} + \varepsilon_{1t}, \\
y_{2t} &= \alpha_{20} + \alpha_{21}t + \sum_{i=1}^{k+d} \beta_{2i} y_{1t-i} + \sum_{i=1}^{k+d} \delta_{2i} y_{2t-i} + \varepsilon_{2t},
\end{aligned}
$$

where t is a time trend, k is the lag order of the original VAR model, and ε_{it} are the error terms. The additional d lags in the VAR augment the system for the possible maximum order of integration of the variables. The non-causality from y_{2t} to y_{1t} (denoted by $y_{2t} \nrightarrow^{GC} y_{1t}$) refers to the situation that the predictions of y_{1t} conditional on its own history cannot be improved by incorporating the past k lags of y_{2t} in the model. The null hypothesis of $y_{2t} \nrightarrow^{GC} y_{1t}$ is

$$
H_0 : \delta_{11} = \cdots = \delta_{1k} = 0.
$$

More generally, for a n-dimensional vector y_t, the lag-augmented VAR model is

$$
y_t = \gamma_0 + \gamma_1 t + \sum_{i=1}^{k} J_i y_{t-i} + \sum_{j=k+1}^{k+d} J_j y_{t-j} + \varepsilon_t, \tag{1}
$$

where $J_{k+1} = \cdots = J_{k+d} = 0$ and d is the maximum order of integration in variable y_t. The regression equation may be rewritten as

$$
y_t = \Gamma \tau_t + \Phi x_t + \Psi z_t + \varepsilon_t, \tag{2}
$$

where $\Gamma = (\gamma_0, \gamma_1)_{n \times (q+1)}$, $\tau_t = (1, t)'_{2 \times 1}$, $x_t = (y'_{t-1}, ..., y'_{t-k})'_{nk \times 1}$, $z_t = (y'_{t-k-1}, ..., y'_{t-k-d})'_{nd \times 1}$, $\Phi = (J_1, ..., J_k)_{n \times nk}$, and $\Psi = (J_{k+1}, ..., J_{k+d})_{n \times nd}$. The null hypothesis of Granger non-causality is given by the restrictions

$$
H_0 : \mathbf{R}\phi = 0 \tag{3}
$$

on the parameter $\phi = \text{vec}(\Phi)$ using row vectorization, and $\mathbf{R}$ is a $m \times n^2 k$ matrix. The coefficient

matrix Ψ of the final d lagged vectors is ignored as its elements are taken to be zero.

The expression in (1) may be written in a more compact form as

$$Y = \tau\Gamma' + X\Phi' + Z\Psi' + \varepsilon$$

where $Y = (y_1, y_2 ..., y_T)'_{T \times n}$, $\tau = (\tau_1, ..., \tau_T)'_{T \times 2}$, $X = (x_1, ..., x_T)'_{T \times nk}$, $Z = (z_1, ..., z_T)'_{T \times nd}$, and $\varepsilon = (\varepsilon_1, \cdots, \varepsilon_T)'_{T \times n}$. Let $Q_\tau = I_T - \tau (\tau'\tau)^{-1} \tau'$ and $Q = Q_\tau - Q_\tau Z (Z'Q_\tau Z)^{-1} Z'Q_\tau$. The OLS estimator is

$$\hat{\Phi} = Y'QX (X'QX)^{-1}.$$

The standard Wald statistic $\mathcal{W}$ to test the hypothesis H_0 is

$$\mathcal{W} = \left(\mathbf{R}\hat{\phi} \right)' \left[\mathbf{R} \left\{ \hat{\Sigma}_\varepsilon \otimes (X'QX)^{-1} \right\} \mathbf{R}' \right]^{-1} \mathbf{R}\hat{\phi}, \tag{4}$$

where $\hat{\phi} = \text{vec}(\hat{\Phi})$, $\hat{\Sigma}_\varepsilon = \frac{1}{T}\hat{\varepsilon}'\hat{\varepsilon}$, and $\otimes$ is the Kronecker product. Toda and Yamamoto (1995) and Dolado and Lütkepohl (1996) show that this Wald statistic has the usual χ^2_m asymptotic null distribution with m being the number of restrictions.

The recursive Granger causality tests calculate Wald statistics from subsamples of the data. Suppose f_1 and f_2 are the (fractional) starting and ending points of the regression sample and $f_w = f_2 - f_1$. The Wald statistic (based on the lag-augmented VAR) calculated from this subsample is denoted by $\mathcal{W}_{f_1}^{f_2}$. Let $\tau_1 = \lfloor f_1 T \rfloor$, $\tau_2 = \lfloor f_2 T \rfloor$, $\tau_w = \lfloor f_w T \rfloor$, where T is the total number of observations, and $\tau_0 = \lfloor f_0 T \rfloor$ be the minimum number of observations required to estimate the VAR system. For the forward expanding window procedure[4], the starting point τ_1 is fixed at the first observation (i.e. $\tau_1 = 1$) and the regression window expands from τ_0 to T. This process is equivalent to having τ_2 move from τ_0 to T.

The regression window size for the rolling procedure is fixed. Here, it is assumed the window size equals τ_0. The start point τ_1 moves from the first observation to $T - \tau_0 + 1$ and the end point $\tau_2 = \tau_1 + \tau_0 - 1$. Alternatively, one can write τ_1 and τ_2 of the procedure as $\tau_2 = \{\tau_0, ..., T\}$ and $\tau_1 = \tau_2 - \tau_0 + 1$. The end point of the regression runs from τ_0 to the last observation of the sample T and the start point follows, keeping a fixed window size τ_0.

For the recursive evolving window procedure, like the rolling window procedure, the end point of the regression $\tau_2 = \{\tau_0, ..., T\}$. However, the starting point of the regression τ_1, instead

---

[4]The forward expanding window Granger causality test has been considered in Thoma (1994), but in the (unaugmented) original VAR model for systems containing integrated variables.

of keeping a fixed distance with τ_2 as in the rolling window procedure, varies from 1 to $\tau_2 - \tau_0 + 1$ (covering all possible values). For each observation of interest f, one obtain a sequence of Wald statistics $\{\mathcal{W}_{f_1,f_2}\}_{f_2=f}^{f_1 \in [0, f_2 - f_0]}$. The test statistic is defined to be the supremum of the Wald statistic sequence

$$\mathcal{SW}_f(f_0) = \sup_{f_2=f, f_1 \in [0, f_2 - f_0]} \{\mathcal{W}_{f_1,f_2}\}.$$

Inference on Granger non-causality for observation $\lfloor fT \rfloor$ is based on the sup Wald statistic $\mathcal{SW}_f(f_0)$.

3 Limiting distributions

In this subsection, the limiting distributions for the Wald and sup Wald statistics are derived under the assumptions that the process $\{y_t\}$ is I(1) or I(2) around a linear trend. Those assumptions are sufficient but not necessary.

Let an n-vector time series $\{y_t\}_{t=-k+1}^{\infty}$ be generated by the following model

$$y_t = \beta_0 + \beta_1 t + \eta_t \tag{5}$$

with η_t following a VAR(k) process

$$\eta_t = J_1 \eta_{t-1} + ... + J_k \eta_{t-k} + \varepsilon_t, \tag{6}$$

where ε_t is the error term. Substituting $\eta_t = y_t - (\beta_0 + \beta_1 t)$ into (5), we have

$$y_t = \gamma_0 + \gamma_1 t + J_1 y_{t-1} + ... + J_k y_{t-k} + \varepsilon_t, \tag{7}$$

where γ_i are function of β_i and J_h with $i = 0, 1$ and $h = 1, .., k$. The equation is initialized at $t = -k + 1, ...0$ and the initial values $\{\eta_{-k+1}, ...\eta_0\}$ are any random vectors including constant.

Equation (6) can be rewritten as

$$\eta_t = J_1 \eta_{t-1} + ... + J_k \eta_{t-k} + J_{k+1} \eta_{t-k-1} + J_{k+2} \eta_{t-k-2} + \varepsilon_t \tag{8}$$

where $J_{k+1} = J_{k+2} = 0$. Alternatively, equation (8) can be written in error correction format as

$$\Delta \eta_t = J_1^{\dagger} \Delta \eta_{t-1} + ... + J_{k+1}^{\dagger} \Delta \eta_{t-k-1} + \Pi_2 \eta_{t-k-2} + \varepsilon_t,$$

where $J_i^\dagger = \sum_{h=1}^i J_h - I_n$ $(i = 1, \ldots, k+1)$ and $\Pi_2 = -J(1)$. The conditions for η_t to be I(1) or I(2) are presented below, which are the same as those in Toda and Yamanoto (1995).

Assumption 1 ε_t *is a iid sequence of n-dimensional random vectors with mean zero and variance matrix $\Sigma_\varepsilon > 0$ such that $E\,|\varepsilon_{it}|^{2+\delta} < \infty$ for some $\delta > 0$.*

Assumption 2 $|J(z)| = 0$ *implies $|z| > 1$ or $z = 1$, where $J(z) = I_n - J_1 z - \ldots - J_{k+2} z^{k+2}$.*

Assumption 3 $\Pi_2 = AB'$ *for some A and B, where A and B are $n \times r$ matrices of rank r $(0 < r < n)$. If $\Pi_2 = 0$, we say $r = 0$.*

Assumption 4 $A'_\perp \Pi_1 B_\perp$ *is nonsingular, where $\Pi_1 = -J^\dagger(1)$ with $J^\dagger(z) = I_n - J_1^\dagger z - \cdots - J_{k+1}^\dagger z^{k+1}$, and $A_\perp$ an $B_\perp$ are $n \times (n-r)$ matrices of rank $n - r$ such that $AA'_\perp = B'B_\perp = 0$. (If $r = 0$, we take $A_\perp = B_\perp = I_n$).*

Assumption 2 excludes explosive processes, but allows for the model to have some unit roots. Assumption 3 defines the cointegration space to be of rank r and B is a matrix whose columns span this space. Assumption 4 ensures that $\Delta \eta_t$ is stationary with a Wold representation, $B' \eta_t$ is stationary. Under assumption 1-4, the process η_t is I(1) and is cointegrated if $r > 0$ (Theorem 2 of Johansen (1992)).

Furthermore, we can rewrite equation (8) as

$$\Delta^2 \eta_t = J_1^* \Delta^2 \eta_{t-1} + \ldots + J_k^* \Delta^2 \eta_{t-k} + \Pi_1 \Delta \eta_{t-k-1} + \Pi_2 \eta_{t-k-2} + \varepsilon_t, \tag{9}$$

where $J_i^* = \sum_{h=1}^i J_h^\dagger - I_n$ $(i = 1, \ldots, k)$.

Assumption 5 $\bar{A}'_\perp \Pi_1 \bar{B}_\perp = FG'$ *for some F and G, where $\bar{A}_\perp = A_\perp (A'_\perp A_\perp)^{-1}$ and $\bar{B}_\perp = B_\perp (B'_\perp B_\perp)^{-1}$, and F and G are $(n-r) \times s$ matrices of rank s $(0 < s < n - r)$. If $\Pi_1 = 0$, we say $s = 0$.*

According to Theorem 3 of Johansen (1992), the process η_t is I(2) and is cointegrated unless $r = s = 0$, under Assumptions 1-3, 5, and Assumption (2.8) of Johansen (1992), which prevents η_t from being I(3).

Theorem 1 *Suppose the process $\{y_t\}$ is stationary, I(1) or I(2), possibly around a linear trend in each case.*[5] *The subsample Wald statistic $\mathcal{W}_{f_1,f_2}$ has limit distribution given by*

$$\mathcal{W}_{f_1,f_2} \Rightarrow \left[\frac{W_m(f_2) - W_m(f_1)}{f_w^{1/2}}\right]' \left[\frac{W_m(f_2) - W_m(f_1)}{f_w^{1/2}}\right],$$

which is a quadratic function of the limit process, $W_m(.)$, a vector standard Brownian motion with covariance matrix I_m where m is the number of restrictions. The sup Wald statistic converges to

$$\mathcal{SW}_f(f_0) \Rightarrow \sup_{f_1 \in [0, f_2 - f_0], f_2 = f} \left[\frac{W_m(f_2) - W_m(f_1)}{f_w^{1/2}}\right]' \left[\frac{W_m(f_2) - W_m(f_1)}{f_w^{1/2}}\right]$$

as $T \to \infty$.

The proof of the Theorem is given in the Online Supplement(Shi, Hurn, and Phillips, 2019).[6]

4 Dealing with multiplicity in recursive testing

It is well known that the probability of making a Type I error rises with the number of hypotheses in a test sequence, a phenomenon known as multiplicity. In the current application and for a data series of sample size T, the recursive test statistics run from $\lfloor T f_0 \rfloor$ to T, so the number of hypotheses tested equals $T - \lfloor T f_0 \rfloor + 1$. Furthermore, it is obvious from Theorem 1 that the test statistics are not independent and hence a simple Bonferroni correction will be too conservative and hence lead to lower power (Romano and Wolf, 2005). Instead, a bootstrap method to resolve the multiplicity issue is explored and implemented here in our simulations and empirical work.

This solution is very much in the spirit of the bootstrap reality check of White (2000). The essence of the idea is as follows. Let τ_b be the number of observations in the window over which size is to be controlled. The bootstrap procedure allows the probability of having at least one false positive detection over the τ_b period to be 5%. The derivation of the limiting properties of the bootstrap procedure is not straightforward because it involves integrated variables, subsample analysis, and family wise size control. These complications are considerable and analysis is left

---

[5]We derive the asymptotic distribution of the subsample Wald and the sup Wald statistics under the null hypothesis with maximum integration order two, as in Toda and Yamamoto (1995).

[6]The Online Supplement, together with Matlab code and data to implement and to reproduce the results, is downloadable from `https://sites.google.com/site/shupingshi`.

for future work.[7]

The bootstrap procedure may be described most simply in the context of a bivariate VAR(1) model for the null hypothesis of no Granger causality from y_2 to y_1.

Step 1: Estimate a bivariate VAR(1) model which imposes the null hypothesis of no causality from y_2 to y_1 such that

$$
\begin{bmatrix} y_{1t} \\ y_{2t} \end{bmatrix} = \begin{bmatrix} \phi_{11} & 0 \\ \phi_{12} & \phi_{22} \end{bmatrix} \begin{bmatrix} y_{1t-1} \\ y_{2t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{bmatrix}.
$$

Denote the estimated coefficients by $\widehat{\phi}_{11}$, $\widehat{\phi}_{12}$, and $\widehat{\phi}_{22}$ and the estimated residuals by e_{1t} and e_{2t}.

Step 2: Let $T_b = \tau_0 + \tau_b - 1$ be the sample size of the bootstrapped data series. Generate a bootstrap sample given by

$$
\begin{bmatrix} y_{1t}^b \\ y_{2t}^b \end{bmatrix} = \begin{bmatrix} \hat{\phi}_{11} & 0 \\ \hat{\phi}_{12} & \hat{\phi}_{22} \end{bmatrix} \begin{bmatrix} y_{1t-1}^b \\ y_{2t-1}^b \end{bmatrix} + \begin{bmatrix} e_{1t}^b \\ e_{2t}^b \end{bmatrix} \tag{10}
$$

where the residuals e_{1t}^b (resp. e_{2t}^b) are randomly drawn with replacement from the estimated residuals e_{1t} (resp. e_{2t}). The initial values $y_{11}^b = y_{11}$ and $y_{21}^b = y_{21}$.

Step 3: Using the bootstrapped series, compute the test statistic sequences for the forward, $\{\mathcal{W}_{1,t}^b\}_{t=\tau_0}^{\tau_0+\tau_b-1}$, rolling, $\{\mathcal{W}_{t-\tau_0+1,t}^b\}_{t=\tau_0}^{\tau_0+\tau_b-1}$, and recursive evolving, $\{\mathcal{SW}_t^b(\tau_0)\}_{t=\tau_0}^{\tau_0+\tau_b-1}$, algorithms, respectively. The maximum values of each bootstrapped test statistic sequence is calculated such that

$$
\text{Forward: } \mathcal{M}_{1,t}^b = \max_{t\in[\tau_0,\tau_0+\tau_b-1]} \left(\mathcal{W}_{1,t}^b \right),
$$
$$
\text{Rolling: } \mathcal{M}_{t-\tau_0+1,t}^b = \max_{t\in[\tau_0,\tau_0+\tau_b-1]} \left(\mathcal{W}_{t-\tau_0+1,t}^b \right), \tag{11}
$$
$$
\text{Recursive: } \mathcal{SM}_t^b(\tau_0) = \max_{t\in[\tau_0,\tau_0+\tau_b-1]} \left(\mathcal{SW}_t^b(\tau_0) \right).
$$

Step 4: Repeat Steps 2-3 for $B = 499$ times.

---

[7]We note that the methods of the paper rely on lag-augmented VAR regression for which standard limit theory for the individual coefficient estimates applies. This means that problems of bootstrap inconsistency from the presence of unit roots in the model do not apply. We therefore believe that it will be possible to establish bootstrap validity under certain regularity conditions. But the difficulty of analysis is still considerable and is left for further study.

Step 5: The critical values of the forward, rolling and recursive evolving approaches are, respectively, the 95% percentiles of the $\left\{\mathcal{M}_{1,t}^b\right\}_{b=1}^B$, $\left\{\mathcal{M}_{t-\tau_0+1,t}^b(\tau_0)\right\}_{b=1}^B$, and $\left\{\mathcal{SM}_t^b(\tau_0)\right\}_{b=1}^B$ sequences.

For practical implementation, one would first select the optimal lag order of the VAR system using the information criteria and then estimate the restrictive model in Step 1. Similarly, in Step 3, one would estimate the lag order before calculating the test statistics. The next subsection investigates the performance of the forward expanding, rolling and recursive evolving causality tests with the bootstrapped critical values under the DGP (12) with different parameter settings for Cases 2-4. The calculation is repeated $1,000$ times for every parameter constellation. The lag length p is fixed at one.

5 Simulation experiments

Consider an n-vector time series $\{Z_t\}_{t=-1}^T$ generated by a first order VAR model

$$Z_t = \Phi_{s_t} Z_{t-1} + u_t, u_t \overset{i.i.d}{\sim} N\left(0, \Sigma_u\right), \tag{12}$$

in which Φ_{s_t} is the first-order autoregressive coefficient, and Σ_u is the covariance matrix of u_t. Now introduce a transformation of the model, which is equivalent to the original, specified as

$$Z_t^* = \Phi_{s_t}^* Z_{t-1}^* + u_t^*, \qquad u_t^* \overset{i.i.d}{\sim} N\left(0, \Sigma_u^*\right), \tag{13}$$

where $Z_t^* = P^{-1}Z_t$, $u_t^* = P^{-1}u_t$, $\Phi_{s_t}^* = P^{-1}\Phi_{s_t}P$, $\Sigma_u^* = P^{-1}\Sigma_u P^{-1\prime}$ and P is an $n \times n$ block lower-triangular matrix. Notice that for any combination of Φ_{s_t} and Σ_u that give the same value of $\Phi_{s_t}^*$, the Wald statistics are the same (Toda and Yamamoto, 1995). Let $\Sigma_u = PP'$. Then $\Sigma_u^* = I_n$. The transformed model (13) gives a standardized version of the VAR model (12).

The simulation is based on a bivariate version of (13) with $Z_t^* = (y_{1t}^*, y_{2t}^*)'$. For simplicity, the causal relationship from y_{1t}^* to y_{2t}^* is shut down so that

$$\Phi^* = \begin{pmatrix} \phi_{11}^* & \phi_{s_t}^* \\ 0 & \phi_{22}^* \end{pmatrix}.$$

Under the null hypothesis of no causality, $\phi_{s_t}^*$ equals zero. Under the alternative hypothesis, the causation relationship runs from y_{2t}^* to y_{1t}^* for certain periods of the sample. Let s_t be the causality indicator which takes the value 1 for the causality periods and is zero otherwise. The

autoregressive coefficient $\phi_{s_t}^*$ is defined as $\phi_{s_t}^* = \phi_{12}^* s_t$. The non-explosive conditions for the system are $|\phi_{11}^*| \leq 1$ and $|\phi_{22}^*| \leq 1$. There are four different cases to consider.

Case 1. If both $|\phi_{11}^*|$ and $|\phi_{22}^*|$ are smaller than unity, the system is stationary. The maximum lag order is zero and hence the lag augmented VAR is exactly the same as the original VAR model. This case has been examined in detail in Shi, Phillips, and Hurn (2018), with $(\phi_{11}^*, \phi_{22}^*)$ equaling $(0.5, 0.5), (0.5, 0.8), (-0.5, 0.8)$, and $(0.5, -0.8)$.

Case 2. If $\phi_{11}^* = 1$ and $|\phi_{22}^*| < 1$, we have

$$\Delta y_{1t}^* = \phi_{s_t}^* y_{2t-1}^* + u_{1t}^*$$
$$y_{2t}^* = \phi_{22}^* y_{2t-1}^* + u_{2t}^*$$

where $y_{1t}^* \sim I(1)$ and $y_{2t}^* \sim I(0)$ and hence $d_{\max} = 1$. In the simulation study, we consider the following settings of $(\phi_{11}^*, \phi_{22}^*)$: $(1, 0.8), (1, -0.8)$. The test statistics are calculated based on a VAR(2) model.

Case 3. If $\phi_{22}^* = 1$ and $|\phi_{11}^*| < 1$, we have

$$\Delta y_{1t}^* = (\phi_{11}^* - 1) y_{1t-1}^* + \phi_{s_t}^* y_{2t-1}^* + u_{1t}^*$$
$$\Delta y_{2t}^* = u_{2t}^*.$$

If $\phi_{s_t}^* = 0$, $y_{1t}^* \sim I(0)$ and $y_{2t}^* \sim I(1)$. If $\phi_{s_t}^* \neq 0$, y_{1t}^* and y_{2t}^* are cointegrated with a cointegration vector of $[\phi_{11}^* - 1, \phi_{s_t}^*]$. In addition, y_{2t}^* is weakly exogenous for the cointegration parameters. The value of $(\phi_{11}^*, \phi_{22}^*)$ are set as $(0.8, 1), (-0.8, 1)$ in the simulation studies. The test statistics are obtained from a VAR(2) model.

Case 4. If both ϕ_{11}^* and ϕ_{22}^* are unity, we have a full unit root model, namely

$$\Delta y_{1t}^* = \phi_{s_t}^* y_{2t-1}^* + u_{1t}^*$$
$$\Delta y_{2t}^* = u_{2t}^*.$$

If $\phi_{s_t}^* = 0$, both y_{1t}^* and y_{2t}^* are $I(1)$. If $\phi_{s_t}^* \neq 0$, y_{1t}^* is $I(2)$ and y_{2t}^* is $I(1)$. The maximum integration order $d_{\max}$ equals two. The regression model is VAR(3).

We calculate the sizes (probability of rejecting at least one true null hypothesis) and powers (probability of rejecting at least one false null hypothesis over the sample period) of the three

13

procedures. The family-wise error rate over the entire sample period is controlled by setting $T_b = T$. Table 1 reports the impact of the persistence parameters $\{\phi_{11}, \phi_{22}\}$ (top panel), the minimum window size f_0 (middle panel), and the sample size T (bottom panel), respectively, on the sizes and powers of the three algorithms. Table 2 focuses on the impact of causality characteristics (causal duration $\mathcal{D}$ and the location of the causal episode f_e) on the empirical powers of the tests.

The data generating procedure under the null is (12) with $\phi_{s_t}^* = 0$. Under the alternative, the DGP contains a single causality episode from $y_{2t} \rightarrow y_{1t}$. Let f_e and f_f be the origination and termination dates of causality. The causality indicator s_t is one for $\lfloor f_e T \rfloor \leq t \leq \lfloor f_f T \rfloor$ and zero otherwise. In Table 1, we let causality switch on in the middle of the sample (i.e. $f_e = 0.5$) and last for 20% of the sample (i.e. $f_f = 0.7$). The causality strength parameter ϕ_{12} is either 0.8 or 1.5. In Table 2, we fix the strength of causality such that $\phi_{12} = 1.5$.

As evident in Table 1, the empirical sizes of all three procedures are very close to the nominal size of 5%.[8] This reveals the effectiveness of the bootstrap procedure in controlling family-wise size and addressing multiplicity in recursive testing. As for power, it is apparent from the results (Tables 1 and 2) that the recursive evolving window procedure has the highest power overall, followed closely by the rolling algorithm. Although the performance of the rolling and recursive evolving procedures are almost identical in most cases, the power improvement of the recursive evolving method is quite obvious when the causal strength is moderate $\phi_{12} = 0.8$ and the sample size is large ($T = 200$ and $T = 300$). The power of the forward approach is far below that of the rolling and recursive evolving algorithms.

The detective power of the rolling and recursive evolving procedures increases when the minimum window length rises from 18 to 24 observations but decreases upon further extension to 36 observations. Unlike these two procedures, the power of the forward algorithm increases as the initialization f_0 becomes larger. The powers of all three methods increase (but at a decreasing rate) with the sample size T. The additional information in the data as T rises from 100 to 300 delivers more accurate estimates of the model parameters and this accuracy naturally translates into more accurate tests. As expected, powers of all the tests rise with the strength of causality. Additionally, from Table 2, powers of the forward and recursive rolling procedures

---

[8]Unreported simulations show that with the standard bootstrap critical values (i.e. obtained from $\{\mathcal{W}_{1,t}^b\}_{b=1}^{499}$, $\{\mathcal{W}_t(\tau_0)^b\}_{b=1}^{499}$, and $\{\mathcal{SW}_t(\tau_0)^b\}_{b=1}^{499}$), the three algorithms are significantly over-sized. The empirical sizes of the forward, rolling and recursive evolving algorithms are, respectively, around 15%, 50% and 40% when $f_0 = 0.24$ and $T = 100$.

Table 1: The empirical sizes and powers of the testing procedures. The parameter setting under the alternative is: $f_e = 0.5$ and $\mathcal{D} = 0.2$.

	Size			Power ($\phi_{12} = 0.8$)			Power ($\phi_{12} = 1.5$)		
	Forward	Rolling	Recursive	Forward	Rolling	Recursive	Forward	Rolling	Recursive
$(\phi_{11}^*, \phi_{22}^*)$: $f_0 = 0.24$ and $T = 100$									
Case 2: d = 1									
(1,0.8)	0.06	0.06	0.06	0.22	0.51	0.53	0.51	0.94	0.93
(1,-0.8)	0.05	0.05	0.05	0.23	0.54	0.57	0.53	0.96	0.96
Case 3: d = 1									
(0.8,1)	0.06	0.09	0.08	0.17	0.40	0.42	0.32	0.67	0.67
(-0.8,1)	0.05	0.05	0.05	0.17	0.38	0.41	0.33	0.67	0.67
Case 4: d = 2									
(1,1)	0.06	0.08	0.08	0.15	0.40	0.41	0.29	0.87	0.87
$\lfloor T f_0 \rfloor$: $T = 100$									
Case 2: d = 1 and $(\phi_{11}^*, \phi_{22}^*) = (1, 0.8)$									
18	0.06	0.07	0.07	0.18	0.39	0.43	0.44	0.87	0.90
24	0.06	0.06	0.06	0.22	0.51	0.53	0.51	0.94	0.93
36	0.05	0.05	0.06	0.28	0.51	0.51	0.57	0.86	0.86
Case 3: d = 1 and $(\phi_{11}^*, \phi_{22}^*) = (0.8, 1)$									
18	0.05	0.09	0.8	0.14	0.39	0.43	0.27	0.85	0.89
24	0.06	0.09	0.08	0.17	0.40	0.42	0.32	0.67	0.67
36	0.06	0.06	0.06	0.22	0.36	0.35	0.39	0.56	0.56
Case 4: d = 2 and $(\phi_{11}^*, \phi_{22}^*) = (1, 1)$									
18	0.07	0.08	0.08	0.11	0.26	0.29	0.18	0.69	0.74
24	0.06	0.08	0.08	0.15	0.40	0.41	0.29	0.87	0.87
36	0.06	0.07	0.07	0.21	0.32	0.33	0.36	0.53	0.53
T: $f_0 = 0.24$									
Case 2: d = 1 and $(\phi_{11}^*, \phi_{22}^*) = (1, 0.8)$									
100	0.06	0.06	0.06	0.22	0.51	0.53	0.51	0.94	0.93
200	0.06	0.07	0.08	0.33	0.67	0.81	0.79	1.00	1.00
300	0.05	0.08	0.08	0.55	0.78	0.93	0.95	1.00	1.00
Case 3: d = 1 and $(\phi_{11}^*, \phi_{22}^*) = (0.8, 1)$									
100	0.06	0.09	0.08	0.17	0.40	0.42	0.32	0.67	0.67
200	0.06	0.08	0.07	0.22	0.70	0.80	0.50	1.00	1.00
300	0.04	0.08	0.08	0.35	0.79	0.94	0.70	1.00	1.00
Case 4: d = 2 and $(\phi_{11}^*, \phi_{22}^*) = (1, 1)$									
100	0.06	0.08	0.08	0.15	0.40	0.41	0.29	0.87	0.87
200	0.06	0.08	0.09	0.19	0.56	0.67	0.45	0.99	1.00
300	0.05	0.09	0.09	0.30	0.69	0.85	0.65	1.00	1.00

Note: Calculations are based on 1,000 replications, with the 5% bootstrapped critical values of the statistics defined in (11).

increases with the duration of the causality episode. Moreover, the empirical powers of all tests are higher when the causal episode occurs earlier in the sample.

Table 2: The impact of causal characteristics on empirical powers of the testing procedures. The causality strength $\phi_{12}^* = 1.5$, the minimum window $f_0 = 0.24$, and $T = 100$.

|	Case 2: $(\phi_{11}^*, \phi_{22}^*) = (1, 0.8)$			Case 3: $(\phi_{11}^*, \phi_{22}^*) = (0.8, 1)$			Case 4: $(\phi_{11}^*, \phi_{22}^*) = (1, 1)$		
	Forward	Rolling	Recursive	Forward	Rolling	Recursive	Forward	Rolling	Recursive
Causality Duration $\mathcal{D}$: $f_e = 0.5$									
$\mathcal{D} = 0.1$	0.20	0.34	0.34	0.13	0.21	0.21	0.12	0.18	0.19
$\mathcal{D} = 0.2$	0.51	0.94	0.93	0.32	0.67	0.67	0.29	0.87	0.87
$\mathcal{D} = 0.3$	0.80	1.00	1.00	0.57	1.00	1.00	0.51	0.98	0.99
Causality Location f_e: $\mathcal{D} = 0.2$									
$f_e = 0.3$	0.67	0.95	0.95	0.44	0.73	0.74	0.42	0.88	0.88
$f_e = 0.5$	0.51	0.94	0.93	0.32	0.67	0.67	0.29	0.87	0.87
$f_e = 0.7$	0.38	0.95	0.95	0.23	0.59	0.60	0.19	0.85	0.85

Note: Calculations are based on 1,000 replications, with the 5% bootstrapped critical values of the statistics defined in (11).

6   The money-income relationship

The money-income relationship in the United States is examined using a four-variable VAR model comprising the logarithm of industrial production (ip_t), the logarithm of the money base (denoted by m_t), the logarithm of the price index (p_t), and the interest rate (i_t). Money base is measured as M1 (seasonally adjusted), the interest rate is the secondary market rate on three-month Treasury bills, and prices are measured by the consumer price index for all urban consumers (all items, seasonally adjusted). All data are monthly observations for the period January 1959 to April 2014 (664 observations) and are obtained from the Federal Reserve Economic Database. The data are plotted in Figure 1.

Figure 1 provides visual evidence that at least three of the series, namely, ip_t, m_t and p_t, are non-stationary. While the testing procedure based on the LA-VAR does not require pre-filtering the data series by de-trending or taking differences, it does need information about the maximum possible order of integration. To find the maximum order of integration of the system, we conduct augmented Dickey-Fuller (ADF) tests for all data series (Dickey and Fuller, 1979) with a constant and a linear time trend in the regression equation. In addition, to account for potential structural breaks in the data series we use the unit root tests of Perron and Vogelsang (1992) and Clemente, Montanes, and Reyes (1998): each of these tests searches for unknown structural breaks with either additive outliers (AO) or innovational outliers (IO). The Perron-Vogelsang test allows for one break, while the Clemente *et al.* (1998) test allows for two breaks under both the null hypothesis of a unit root null and alternative hypotheses of stationarity.

Figure 1: Time series plots of the logarithms of industrial production, money (M1) and price index, and the interest rate in the United States from January 1959 to April 2014.
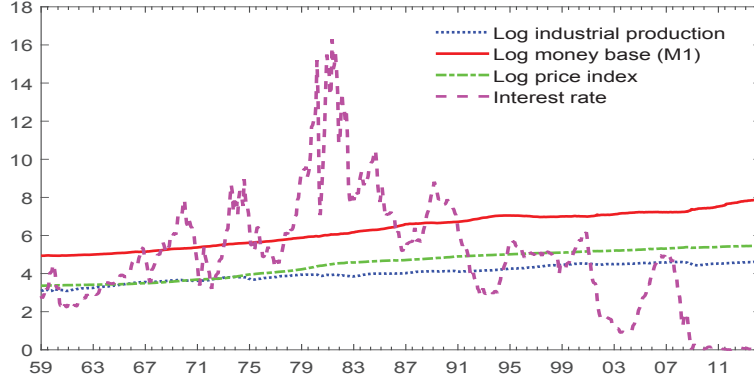
The test statistics and their respective finite sample critical values are displayed in Table 3. Lag orders of all tests are selected using BIC with maximum lag order of 5. The finite sample critical values are obtained from Monte Carlo simulation with 5,000 replications.

All data series are found to be I(1) when assuming no structural break (the ADF test) or one unknown structural break with additive outliers – the AO test of Perron and Vogelsang (1992) – or two unknown structural breaks with additive outliers – the AO test of Clemente et $al.$ (1998). When assuming structural break(s) with innovational outliers, p_t is found to stationary while i_t is stationary only in the case of the IO test with two unknown breaks. The other data series are found to be I(1). This result implies that the maximum order of integration is I(1). We therefore include both a constant and a time trend in the regression as in (1) with $y_t = (ip_t, m_t, p_t, i_t)'$ and set the lag addition parameter d to unity.

We now investigate the existence of causal relationship from money to income using the forward, rolling and recursive evolving procedures. The minimum window size is set to be 72 (six-years). The lag length is selected using BIC applied to the whole sample period with a maximum lag order of 12. The selected lag order is four and applied to all subsample regressions. We control the overall size over a one-year period to be 5%. In bootstrapping the critical values, model parameters under the null are estimated using the whole sample period. The sample size of the bootstrapped data series is $T_b = \tau_0 + 11$ and hence the maximum value in (11) is taken over a sequence of 12 test statistics.

Figure 2 displays the test statistic sequences and their corresponding bootstrapped 5% criti-

Figure 2: Does money Granger cause income? Tests are obtained from a VAR model (1) with $y_t = (ip_t, m_t, p_t, i_t)'$ and lag augmentation $d = 1$. The 5% bootstrapped critical values are obtained from (11) with 499 repetitions and controlled over a one-year period. The sequence of tests for the forward recursive, rolling window, and recursive evolving procedures run from November 1964 to April 2014 with 72 observations for the minimum window size. Lag orders are assumed to be constant and selected using BIC with a maximum length of 12 for the whole sample period.
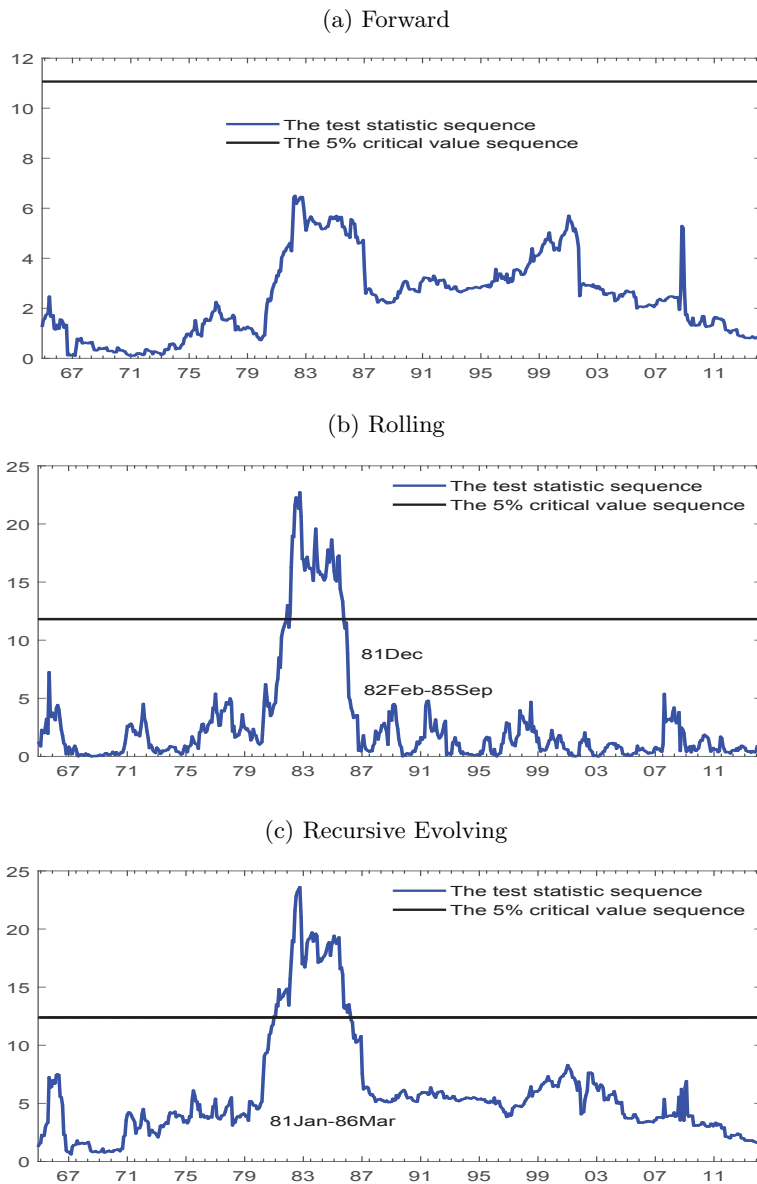
(a) Forward

(b) Rolling

(c) Recursive Evolving

Table 3: Unit Root Tests

| | ADF | Perron and Vogelsang (1992) | | Clemente *et al.* (1998) | |
	cst. & trend	AO	IO	AO	IO
ip_t	-1.81	-2.43	-3.21	-3.20	-4.26
m_t	-0.68	-2.26	-1.56	-3.00	-4.24
p_t	2.67	-2.68	-7.76	-3.49	-8.98
i_t	-2.91	-3.60	-3.66	-4.87	-4.96
Δip_t	-17.80	-13.84	-13.57	-14.56	-14.17
Δm_t	-20.25	-15.82	-15.55	-17.01	-16.59
Δp_t	-12.62	-10.85	-10.64	-13.67	-13.74
Δi_t	-18.21	-19.58	-19.50	-21.03	-20.75
10%	-3.15	-3.95	-4.14	-4.98	-4.27
5%	-3.43	-4.20	-4.41	-5.23	-5.49
1%	-3.96	-4.79	-5.01	-5.76	-6.09

Note: The finite sample critical values are obtained from Monte Carlo simulation with 5,000 replications. The lag orders are selected using BIC with maximum lag order of 5.

cal values (controlled over a one-year period) for the forward, rolling and recursive evolving procedures. The forward algorithm finds no evidence of Granger causality from money to income over the entire sample period. By contrast, both the rolling and recursive evolving procedures identify causality running from money to income during the Volcker period in the early 1980s. The recursive evolving procedure identifies the starting date for this causal episode as January 1981, which 11 months earlier than that suggested by the rolling window algorithm. This episode lasts until March 1986 (September 1985) according to the recursive evolving (rolling) algorithm.

This causal episode coincides with a period of contractionary monetary policy with higher interest rates which slowed the money supply growth rate and started to bring inflation under control. On the basis of the results reported here, we concur with Stock and Watson (1989) that money will have less predictive power for output if data from the 1980s is excluded. Our results are consistent with those of Thoma (1994) and Psaradakis *et al.* (2005) in the sense that there is, we believe, reliable evidence that money Granger causes output in the first part of the 1980s but not the second part. Our findings differ from Swanson (1998) as we do not find persistent causality between 1960 and 1994. In general, it is fairly clear why the 1980s is problematic from the perspective of causality tests. Our subsample findings suggest that studies in which the end

of the sample is around the mid-1980s are likely to find evidence of a causal relationship running from money to income, while studies in which the sample ends towards the end of the decade will not.

Two comments of a general nature emerge from the results. The first relates to the claim of Psaradakis, *et al.* (2005) that money growth has more predictive power for output growth during recessions than during expansions. Our analysis gives no real support for this claim on two counts. First, the recession from November 1973 to March 1975 does not appear to coincide with any indication of a causal relationship between money and income. Second, there is no evidence of a causal relationship during the recession caused by the global financial crisis of 2008-2009.

The second comment relates to the period from late 2008 to late 2013 during which the Federal Reserve implemented three rounds of quantitative easing. This strategy aimed to prevent further deflation by buying financial assets from the banking sector to increase prices and lower their yields while at the same time increasing the money supply. Surprisingly, there is no empirical evidence of any causality from money to income during this period when deliberate increases in the money supply were employed to stimulate GDP. Part of the explanation for this finding is that despite the increased reserves delivered by quantitative easing, a credit crunch continued and banks continued to hold reserves instead of adopting the riskier strategy of increasing lending in a stagnant economy.

Finally, for the sake of completeness, we conduct an analysis of Granger causality running from income to money. The results are provided in the Online Supplement (Shi, Hurn, and Phillips, 2018) and are summarized briefly here. Interestingly, all three procedures identify an episode in the late 1970s and early 1980s, with the indicator from the recursive evolving procedure being switched on the earliest (May 1975). This observation supports the existence of a bi-directional causal relationship between money and income during the Volcker period. Additionally, the recursive evolving algorithm detects an episode during the early 1990s recession period and an episode which lasted for eight years starting from January 2001. The causal relationship terminates in December 2008. This period is often referred to as the Great Moderation and the relative price stability of that period may account for the strong relationship between income and money.

7 Robustness checks

In order to investigate the robustness of the conclusions on Granger causality from money to income, a sensitivity analysis in undertaken. The following variants of the basic test statistics from the LA-VAR model are presented: heteroskedastic consistent test statistics of Shi, Phillips, and Hurn (2018) are estimated; 5% bootstrapped critical values controlled over a five-year period are used; the regression model is augmented by two lags ($d = 2$) to protect against potentially higher levels of integration in the data; a minimum window size of 60 observations is chosen to search for finer local variability in the test statistics; and an alternative implementation of the bootstrapping procedure is conducted.

The heteroskedastic-consistent sequences of the Granger causal relationship from money to income obtained from (1) with $y_t = (ip_t, m_t, p_t, i_t)'$ and $d = 1$ are presented in panels (a), (c) and (e) of Figure 3 showing the results, respectively, for the forward, rolling window and recursive evolving procedures. Also shown are the 5% bootstrapped critical values (described in Section 3.2) and controlled over a one-year period. The minimum window size is 72. Lag orders are selected using BIC with a maximum length of 12 for the whole sample period (four) and assumed to be constant for all subsamples. The heteroskedastic-consistent subsample Wald test statistic is defined as

$$\mathcal{W}^*_{f_1,f_2} = T_w \left(\mathbf{R}\hat{\phi}_{f_1,f_2} \right)' \left[\mathbf{R} \left\{ \hat{V}^{-1}_{f_1,f_2} \hat{\Sigma}_{f_1,f_2} \hat{V}^{-1}_{f_1,f_2} \right\} \mathbf{R}' \right]^{-1} \mathbf{R}\hat{\phi}_{f_1,f_2}, \tag{14}$$

where $\hat{\phi}_{f_1,f_2} = \text{vec}(\hat{\Phi}_{f_1,f_2})$ with $\hat{\Phi}_{f_1,f_2}$ being the OLS estimate of Φ from the sample running from f_1 to f_2,

$$\hat{V}_{f_1,f_2} = I_n \otimes \hat{Q}_{f_1,f_2} \text{ with } \hat{Q}_{f_1,f_2} = \frac{1}{T_w} \sum_{t=\lfloor Tf_1 \rfloor}^{\lfloor Tf_2 \rfloor} x_t x_t',$$

$$\hat{\Sigma}_{f_1,f_2} = \frac{1}{T_w} \sum_{t=\lfloor Tf_1 \rfloor}^{\lfloor Tf_2 \rfloor} \hat{\xi}_t \hat{\xi}_t' \text{ with } \hat{\xi}_t = \hat{\varepsilon}_t \otimes x_t.$$

The heteroskedastic-consistent sup Wald statistic is

$$\mathcal{SW}^*_f (f_0) = \sup_{f_2=f, f_1 \in [0, f_2 - f_0]} \left\{ \mathcal{W}^*_{f_1,f_2} \right\}.$$

It is apparent that the assumption about the behaviour of the variance has no significant influence on the outcomes in all three testing procedures. As in Figure 2, the forward procedure

Figure 3: Results from the heteroskedastic consistent test statistics are displayed in panel (a), (c) and (e) and from the size controlled over a five-year period are in panel (b), (d) and (f). The regression model is (1) with $y_t = (ip_t, m_t, p_t, i_t)'$ and lag augmentation $d = 1$. The sequence of tests for the forward recursive, rolling window, and recursive evolving procedures run from November 1964 to April 2014 with 72 observations for the minimum window size. Lag orders are assumed to be constant and selected using BIC with a maximum length of 12 for the whole sample period.

(a) Forward - Heteroskedasticity

(b) Forward - 5-year period

(c) Rolling - Heteroskedasticity

(d) Rolling - 5-year period

(e) Recursive Evolving - Heteroskedasticity

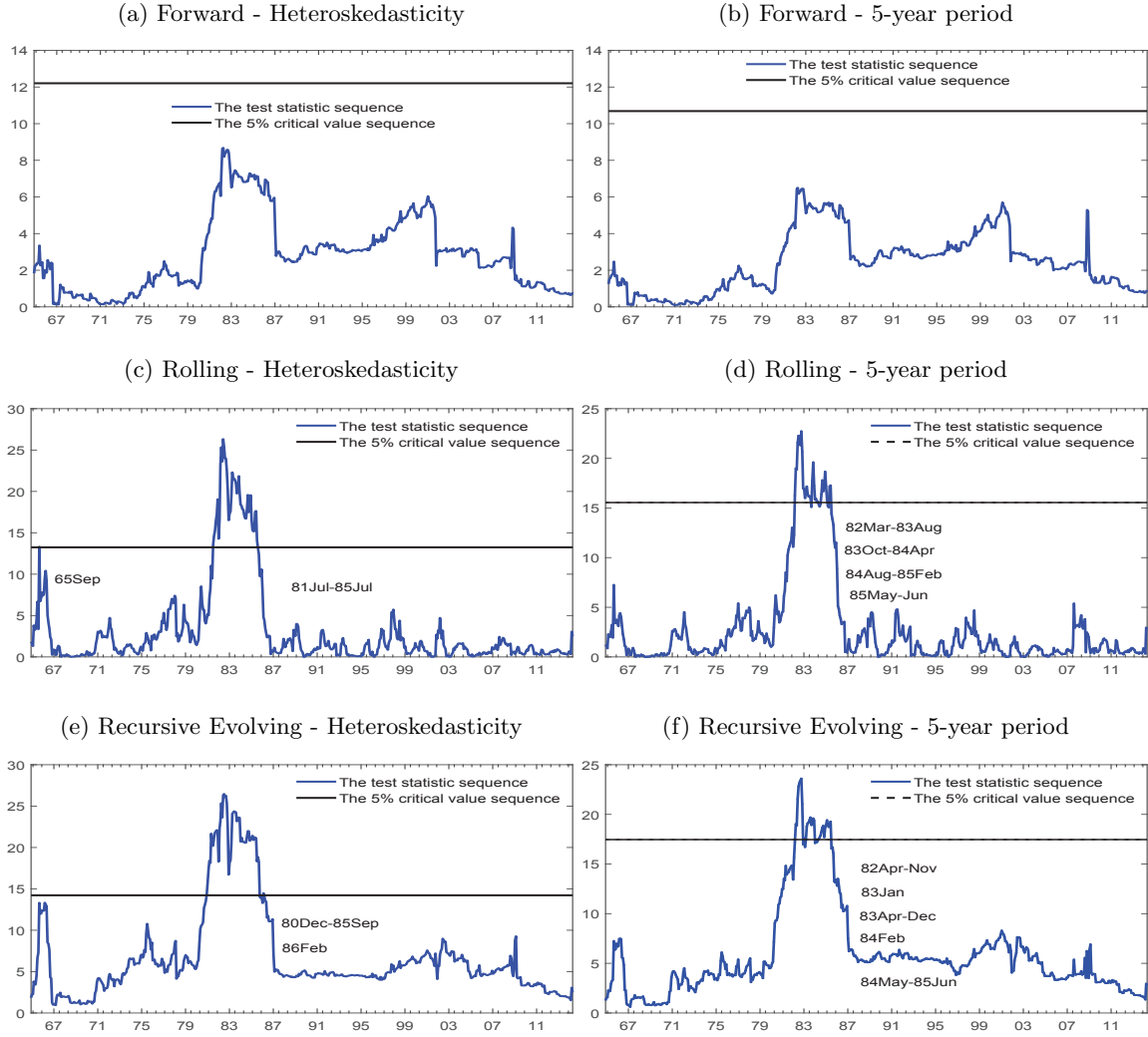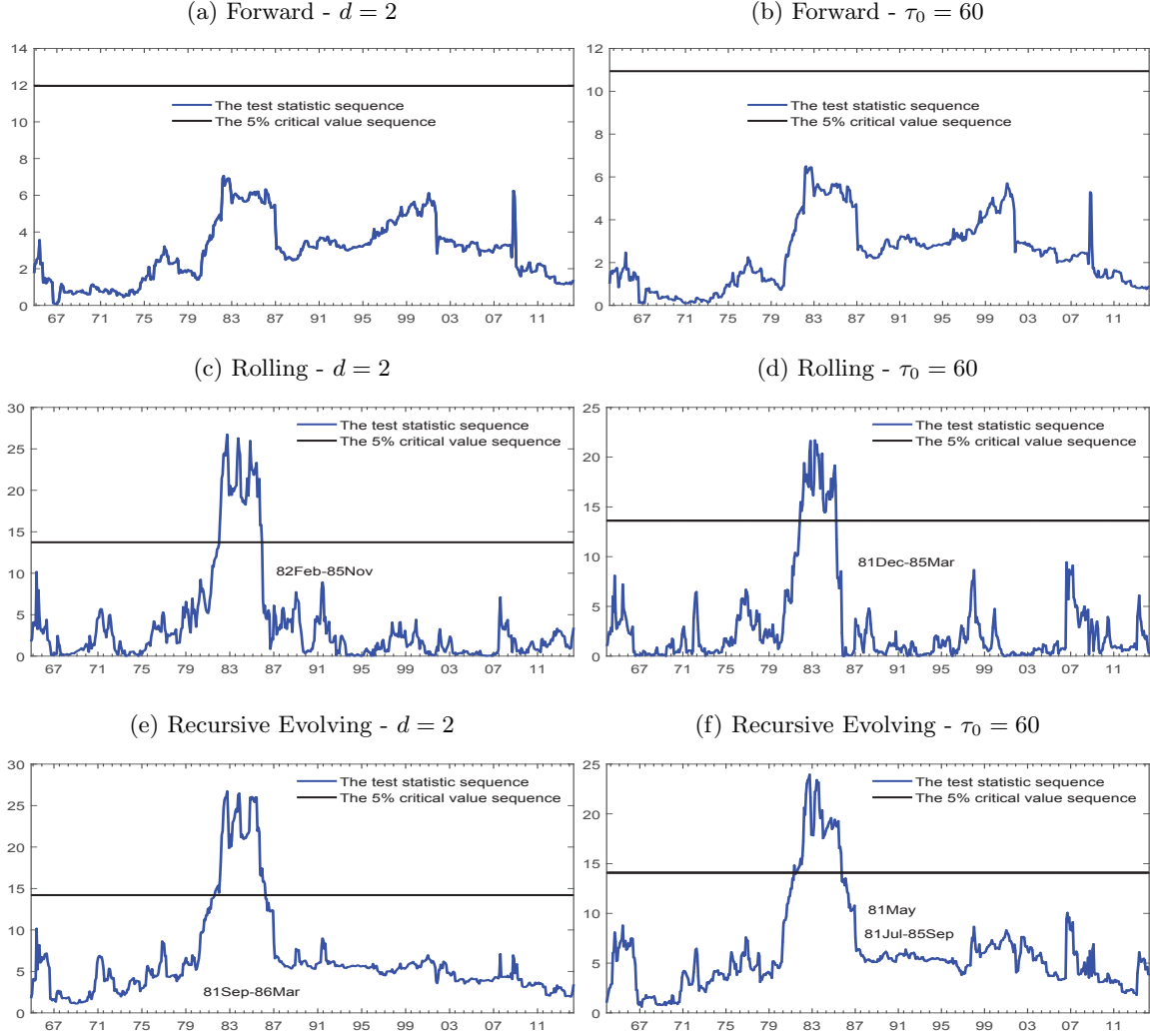(f) Recursive Evolving - 5-year period

22

Figure 4: The forward, rolling, and recursive evolving test results with lag augmentation $d = 2$ are displayed in panel (a), (c) and (e) and those with a minimum window size $\tau_0 = 60$ are in (b), (d) and (f). The regression model is (1) with $y_t = (ip_t, m_t, p_t, i_t)'$, allowing for homoskedastic errors. The empirical size is 5%, controlled over a one-year period and obtained from bootstrapping. Lag orders are assumed to be constant and selected using BIC with a maximum length of 12 for the whole sample period.

(a) Forward - $d = 2$

(b) Forward - $\tau_0 = 60$

(c) Rolling - $d = 2$

(d) Rolling - $\tau_0 = 60$

(e) Recursive Evolving - $d = 2$

(f) Recursive Evolving - $\tau_0 = 60$

23

does not find any episodes of Granger causality from money to income and both the rolling and recursive evolving algorithms detect the early 1980s episode. It is, however, to be noted that the starting dates identified by heteroskedastic-consistent test statistic sequences are slightly earlier than those based on the assumption of homoskedastic errors.

Next, instead of controlling the size of the test sequence over a one-year period, the probability of drawing at least one false positive conclusion is taken to be 5% over a five-year period. The 5% bootstrapped critical values are obtained from (11) with 499 repetitions and bootstrap sample size $T_b = \tau_0 + 59$. It is expected that with a stricter rejection criteria there will be a low chance of making false positive conclusion but, the other hand, a lower power for detection. The basic estimation setup remains otherwise unaltered and the results are presented in panels (b), (d) and (f) of Figure 3 panel. As expected the forward algorithm does not find any evidence of Granger causality, but the pattern of causality identified by the other two algorithms is solid, despite the expected result that the starting date is found to be later than that in Figure 2.

Figure 4 presents result for a larger potential maximum order of integration, $d = 2$, in panels (a), (c), and (e), while those for the minimum window size $\tau_0 = 60$ are reported in panels (b), (d) and (f). Once more all other settings of the LA-VAR model remain the same. Once more the identification of the causal episode appears to be robust to these changes with only small variations in the starting date indicated by the changed parameter settings. The overall conclusion reached here therefore is that the pattern of Granger causality identified using sequences of Wald statistics is remarkably robust to changes in the testing setup.

Finally, the null hypothesis of no Granger causality does not preclude the existence of structural breaks in other model coefficients. To allow for this possibility, one could modify the bootstrap procedure. Specifically, instead of estimating the null model using the full sample and applying the same bootstrap critical value for all individual tests, one could conduct the bootstrap procedure repeatedly for each observation of interest and generate the bootstrap samples from estimates and residuals obtained from the nearest sample. Suppose the observation of interest is t. We can estimate the restricted model (in the first step of the bootstrap procedure) using the nearest (past) window, i.e., observations from $t - \tau_0$ to t. Note that this bootstrap procedure can be implemented in real-time as it only relies on past information.[9] The results for the money to income Granger causality are displayed in Figure 5. We can see that all three

---

[9]The two bootstrap approaches are expected to provide similar simulations results in Section 5, given the data generating processes considered.

Figure 5: The forward, rolling, and recursive evolving test results with the alternative bootstrapping procedure, allowing for homoskedastic errors. The lag augmentation $d = 1$. The empirical size is 5%, controlled over a one-year period and obtained from bootstrapping. Lag orders are assumed to be constant and selected using BIC with a maximum length of 12 for the whole sample period.

(a) Forward - real-time bootstrap

(b) Rolling - real-time bootstrap

(c) Recursive Evolving - real-time bootstrap
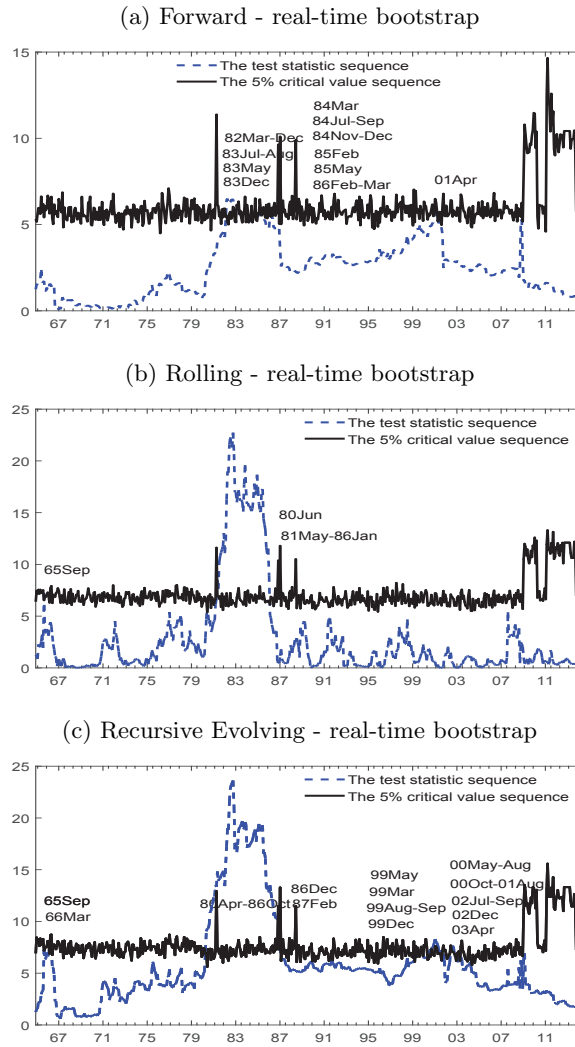
algorithms detect the early 1980s episode, with the signal from the recursive evolving (forward) approach being the strongest (weakest). Note that the forward procedure did not find any evidence of Granger causality when using the bootstrap procedure outlined in Section 4. Both

the rolling and the recursive approaches locate one short episode in mid 1960s. Importantly, the recursive evolving method finds one additional episode around the early 2000s recession, although the signal turns on and off over the identified period.

8 Conclusion

This paper has proposed a recursive evolving Granger causality test based on lag-augmented VAR models that allow for possibly integrated and deterministically trending time series data. The performance of this test is compared with forward recursive regression and rolling window regression tests in a comprehensive simulation study. The modeling framework accommodates multiple configurations of stationary, nonstationary, and cointegrated variables. None of the methods require prefiltering the data to remove potential nonstationarity and bootstrap methods are introduced to ensure family-wise size control in the recursive regression tests.

Simulation results show the effectiveness of the bootstrap procedure in achieving error rates that are close to nominal size, thereby resolving the multiplicity issue that arises in recursive test procedures. The powers of the recursive test algorithms increase with sample size, minimum window length, and the strength and duration of the causal relationship. All three methods perform better when the change in causality occurs earlier in the sample, although, as expected, the impact of location is much more dramatic for the forward procedure than the other two algorithms.

A number of general conclusions emerge. The power of the forward recursive test procedure is well below that of the recursive evolving and rolling approaches. The recursive evolving window approach provides higher power than the simple rolling window algorithm and both have good size control. The results therefore favor recursive evolving tests performed in conjunction with a bootstrap engine for maintaining family-wise size control.

The much-studied causal relationship between money and income in the United States is the focus of the empirical application of these algorithms. The results obtained here confirm some of the conclusions reported in the literature, provide limited support for others and also shed some new light on recent monetary policy experience. One major period of money-income causality is detected, namely, from 1981-1986. This period has been the subject of intense study and speculation in the literature. A very interesting feature of the results is that the policy of quantitative easing following the financial crisis of 2008 does not reveal any statistically

significant causality running from money to income.

References

Arora, V. & Shi, S. (2016). Energy Consumption and Economic Growth in the United States. *Applied Economics*, **48**: 3763-3773.

Balcilar, M., & Ozdemir, Z.A., & Arslanturk, Y. (2010). Economic growth and energy consumption causal nexus viewed through a bootstrap rolling window. *Energy Economics*, **32**: 1398-1410.

Barth, J. R., & Bennett, J. T. (1974). The role of money in the Canadian economy: An empirical test. *The Canadian Journal of Economics*, **7**: 306-311.

Bernanke, B. S. (1986). Alternative Explanations of the Money-Income Correlation. *Carnegie-Rochester Conference Series on Public Policy*, **25**: 49-99.

Christiano, L. J., & Ljungqvist, L. (1988). Money does Granger-cause output in the bivariate money-output relation. *Journal of Monetary Economics*, **22**: 217-235.

Clemente, J., Montanes, A., & Reyes, M. (1998). Testing for a unit root in variables with a double change in the mean. *Economics Letters*, **59**: 175-182.

Dickey, D. A., & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, **74**: 427-431.

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

Dufour, J. M., & Renault, E. (1998). Short run and long run causality in time series: theory. *Econometrica*, **66**: 1099-1125.

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

Eichenbaum, M., & Singleton, K.J. (1986) Do equilibrium real business cycle theories explain postwar U.S. business cycles. *NBER Macroeconomics Annual*, **1986**: 91-146.

Friedman, B. M., & Kuttner, K.N. (1993) Another look at the evidence on money-income causality. *Journal of Econometrics*, **57**: 189-203.

Friedman, M. & Schwartz, A. (1963) Money and business cycles. *Review of Economics and Statistics, Supplement*, **45**: 32-78.

Granger, C.W.J. (1969). Investigating causal relations by econometric models and cross-spectral methods, *Econometrica*, **37**: 424-438.

Granger, C.W.J. (1988). Some recent development in a concept of causality, *Journal of Econometrics*, **39**: 199-211.

Hafer, R. W. & Kutan, A. M. (1997). More evidence on the money-output relationship. *Economic Inquiry*, **35**: 48-58.

Johansen, S. (1992). A representation of vector autoregressive processes integrated of order 2. *Econometric Theory*, **8**: 188-202.

Kamas, L., & Joyce, J. P. (1993). Money, income and prices under fixed exchange rates: evidence from causality tests and VARs. *Journal of Macroeconomics*, **15**: 747-768.

Komura, C. (1982). Money, income, and causality: The Japanese case. *Southern Economic Journal*, **49**: 19-34.

Krol, R., & Ohanian, L. E. (1990). The impact of stochastic and deterministic trends on money–output causality: a multi-country investigation. *Journal of Econometrics*, **45**: 291-308.

Lütkepohl, H. (1993). *Testing for causation between two variables in higher-dimensional VAR models.* In Studies in Applied Econometrics (pp. 75-91). Physica-Verlag HD.

Mills, T. C., & Wood, G. E. (1978). Money-income relationships and the exchange rate regime. *Federal Reserve Bank of St. Louis Review*, **60**: 22-27.

Perron, P. and Vogelsang, T. J. (1992). Nonstationarity and level shifts with an application to purchasing power parity. *Journal of Business & Economic Statistics*, **10**: 301-320.

Phillips, P.C.B. (1995). Fully modified least squares and vector autoregression. *Econometrica*, **63**: 1023-1078.

Phillips, P.C.B., Shi, S., Yu, J. (2015a). Testing for multiple bubbles: Historical episodes of exuberance and collapse in the S&P 500. *International Economic Review*, **56**: 1043-1078.

Phillips, P.C.B., Shi, S., Yu, J. (2015b). Testing for Multiple bubbles: Limit theory of real time detectors. *International Economic Review*, **56**: 1079-1134.

Psaradakis, Z., Ravn, M. O. and Sola, M. (2005). Markov switching causality and the money-output relationship. *Journal of Applied Econometrics*, **20**: 665-683.

Romano, J. P., & Wolf, M. (2005). Stepwise multiple testing as formalized data snooping. *Econometrica*, **73**: 1237-1282.

Rossi, B. (2005). Optimal tests for nested model selection with underlying parameter instability. *Econometric theory*, **21**: 962-990.

Rossi, B. (2013). Advances in forecasting under instability. In *Handbook of economic forecasting*, **2**: 1203-1324, Elsevier.

Shi, S., Phillips, P. C. B., & Hurn, S. (2018). Change detection and the causal impact of the yield curve, *Journal of Time Series Analysis,* **39**: 966-987.

Shi, S., Hurn, S., & Phillips, P. C. B. (2019). Online Supplement to: Causal Change Detection in Possibly Integrated Systems: Revisiting the Money-Income Relationship.

Sims, C. (1972). Money, Income, and Causality. *American Economic Review*, **62**: 540-552.

Sims, C. A. (1987). Comment. *Journal of Business & Economic Statistics*, **5**: 443-449.

Stock, J. H. and Watson, M. W. (1989). Interpreting the evidence on money-income causality. *Journal of Econometrics*, **40**: 161-181.

Swanson, N. R. (1998). Money and output viewed through a rolling window. *Journal of Monetary Economics*, **41**: 455-474

Thoma, M. A. (1994). Subsample instability and asymmetries in money-income causality. *Journal of Econometrics*, **64**: 279-306.

Thornton, D. L. and Batten, D. S. (1985). Lag-length selection and tests of Granger causality between money and income. *Journal of Money, Credit and Banking*, **17**: 164-178.

Toda, H.Y., & Phillips, P.C.B. (1994). Vector autoregression and causality: a theoretical overview and simulation study. *Econometric Reviews*, **13**: 259-285.

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

White, H. (2000). A reality check for data snooping. *Econometrica*, 68(5), 1097-1126.

Williams, D., Goodhart, C. A., & Gowland, D. H. (1976). Money, income, and causality: the UK experience. *The American Economic Review*, **66**: 417-423.

Yamada, H., & Toda, H. Y. (1998). Inference in possibly integrated vector autoregressive models: Some finite sample evidence. *Journal of Econometrics*, **86**: 55-95..