A better understanding of Granger causality analysis: A big data environment*

Xiaojun Song† Abderrahim Taamouti‡

October 9, 2018

ABSTRACT

This paper aims to provide a better understanding of the causal structure in a multivariate time series by introducing several statistical procedures for testing indirect and spurious causal effects. In practice, detecting these effects is a complicated task, since the auxiliary variables that transmit/induce indirect/spurious causality are very often unknown. The availability of hundreds of economic variables makes this task even more difficult since it is generally infeasible to find the appropriate auxiliary variables among all the available ones. In addition, including hundreds of variables and their lags in a regression equation is technically difficult. The paper proposes several statistical procedures to test for the presence of indirect/spurious causality based on big data analysis. Furthermore, it suggests an identification procedure to find the variables that transmit/induce the indirect/spurious causality. Finally, it provides an empirical application where 135 economic variables were used to study a possible indirect causality from money/credit to income.

Keywords: Indirect causality, spurious causality, big data analysis, auxiliary variable(s), asymptotic theory, Monte Carlo simulations, money, credit, income.

Journal of Economic Literature classification: C12; C32; C38; C53; E60.

---

*We thank Anindya Banerjee (Editor) and two anonymous referees for several useful comments. We also thank Han Hong and the participants at the Second Conference on High-Dimensional Statistics in the Age of Big Data in Beijing, 2017 Asian Meeting of the Econometric Society in Hong Kong and IAAE 2017 Annual Conference in Sapporo (Japan) for their very useful comments. Special thanks to Raffaella Giacomini for her discussion which helped us write this paper. All errors are our own.

†Department of Business Statistics and Econometrics, Guanghua School of Management and Center for Statistical Science, Peking University, Beijing, 100871, China. E-mail: sxj@gsm.pku.edu.cn. Financial support from the National Natural Science Foundation of China (Grant No. 71532001) is acknowledged.

‡Department of Economics and Finance, Durham University Business School. Address: Mill Hill Lane, Durham, DH1 3LB, UK. TEL: +44-1913345423. E-mail: abderrahim.taamouti@durham.ac.uk.

1 Introduction

The concept of causality introduced by Wiener (1956) and Granger (1969) constitutes a basic notion for analyzing dynamic relationships between time series. In studying Wiener-Granger causality, predictability is the central issue, hence its importance to economists and policymakers. In practice, Granger-causality is often investigated for bivariate processes. However, different conclusions may be reached when more than two variables are considered. If more than two variables are present, non-causality conditions become more complicated; see e.g. Lütkepohl (1993) and Dufour and Renault (1998). In other words, even if a variable is Granger-causal in a bivariate model, it may not be Granger-causal in a larger model involving more variables. In this case, we talk about an *indirect causality* transmitted through a third variable(s); hereafter referred as auxiliary variable(s). For instance, there may be a variable that drives both variables in the bivariate process, such that when this variable is included into the model, a bivariate causal structure may disappear. In turn, it is also possible that a variable is non-causal for another one in a bivariate model and becomes causal if the information set is extended to include other variables as well. The latter situation corresponds to what is known as a *spurious causality.* Ignoring these causal effects can lead to wrong economic analysis, and consequently to inaccurate policy decisions. In this paper, we borrow from Hsiao (1982) and the literature on factor analysis to introduce statistical procedures that help us detect indirect and spurious causal effects.

The literature on Granger causality analysis is extensive and many tests and measures have been introduced to detect and quantify both linear and non-linear Granger causality; for review see Dufour and Taamouti (2010), Bouezmarni et al. (2012), and Song and Taamouti (2018). The original definition of Granger (1969) that have been adopted in this literature implicitly assumes that all the relevant information is available and used for the causality analysis. However, in practice only a very limited information is considered and the omission of key variables (auxiliary variables) could lead to a spurious causality or might not help detect a possible indirect causality between the variables of interest. The relevance of the information set for Granger causality analysis was first pointed out by Hsiao (1982) [see also Eichler (2007, 2012)], who formally introduced the concept of indirect/spurious causality in a trivariate model. Hsiao (1982) provides a basic framework to explain the causal relationships in a multivariate time series model based on Wiener-Granger notion of causality. He focuses on establishing a Granger causal ordering of the events and on the reconciliation of the disparity between the results obtained from the bivariate and multivariate analysis. He generalizes the Granger's concept of causality to make some provision for spurious/indirect causality which may arise in multivariate analysis. In particular, he shows that a certain type of spurious causality vanishes when the information set is reduced. This observation leads to a strengthened definition of (direct) causality by requiring an improvement in prediction irrespective of the used information set. Finally, Hsiao (1982) characterizes the indirect/spurious causality in the context of VAR models and discusses how to test these causal effects in the presence of *known* auxiliary variables.

It is worth mentioning that indirect/spurious causality might be linked to the omitted variables bias problem. In the context of vector moving average model, Sims (1980) points out that the Granger causal relations may appear in the model because of the omitted variables problem. Furthermore, Lütkepohl (1982) shows that on the one hand Granger-causality in a bivariate system may be due to omission of relevant variables, and on the other hand non-causality in a bivariate system may theoretically result from neglected variables. For Lütkepohl (1982) the structure of the causal relation between the variables of interest can *only* be obtained by including all relevant variables in the model. He adds that "*since many economic variables are important in the sense that they interact, highdimensional time series model-building seems to be required*", but he also recognizes that the latter " *does not seem to be an easy task.*" This paper aims to use big data analysis techniques to proposes statistical procedures that help to test for the presence of indirect/spurious causality.

The main issue of Hsiao (1982)'s framework is that the auxiliary variables that transmit/induce the indirect/spurious causality are implicitly assumed to be *known.* However, in practice these variables are unknown, except in the presence of an economic theory that explicitly specifies the auxiliary variables, which complicates very much the task of testing for the presence of indirect/spurious causality. The availability of hundreds of economic variables makes this task even harder as it is generally infeasible to find the appropriate auxiliary variable(s) among all the available ones. In addition, including hundreds of variables and their lags in a regression equation is technically difficult.

In this paper, we introduce several statistical procedures to test for the presence of indirect/spurious causality using big data analysis. To overcome the problem of unknown relevant auxiliary variables, a diffusion index, extracted using principal component analysis, is included in the regression equation to represent all the variables that are available to practitioners. We derive the asymptotic distributions of the tests in the presence of the estimated index. Furthermore, we conduct a Monte Carlo simulation to evaluate the performance of the proposed statistical procedures. The results show that our procedures have good power for detecting indirect/spurious causality.

Unfortunately, the above statistical procedures only test for the presence/absence of indirect/spurious causality and cannot inform us about the variables of the big data that are responsible for the transmission/induction of this indirect/spurious causality. Another contribution of this paper is we provide an identification procedure which helps us identify the variables in the big data that transmit/induce the indirect/spurious causality.

Finally, to show the practical relevance of the proposed tests, we use 135 economic variables to examine the causality from money/credit to income. In particular, we test whether or not there is an *indirect* causality from monetary policy/credit to income. Thereafter, if this indirect causality exists, then we use the identification procedure discussed above to identify the auxiliary variable(s) that are responsible for the transmission of this indirect causality. Our results show that there is an indirect causality from credit to income, but not from money to income. In addition, the identification procedure indicates that this indirect causality is mainly transmitted through short and long-term interest rates. Hence, interest rates are responsible for the indirect causality from credit to income.

The plan of the paper is as follows. Section 2 presents the general theoretical framework which underlies the definition of indirect/spurious causality. Section 3 provides some motivations for deriving statistical procedures that help detect indirect/spurious causality. In Section 4, we define the regression models and hypotheses that we consider to test for indirect/spurious causality. In Section 5, we provide the asymptotic distributions of the tests. These distributions are derived based on the asymptotic theory from the factor analysis. In Section 6, we propose a statistical procedure that allows us to identify the auxiliary variables that transmit/induce the indirect/spurious causality. In Section 7, we run a Monte Carlo simulation to investigate the finite sample properties of the tests of indirect/spurious causality. Section 8 is devoted to an empirical application. The conclusion is given in Section 9. Finally, the proofs, the parameter values of the data generating processes (DGPs) used in the simulation, the simulation results for the empirical size and power, and the data and the empirical results can be found in a separate companion Appendix, which is available online.

2 Framework

We consider three stochastic processes $\{X_t : t \in \mathbb{Z}\}$, $\{Y_t : t \in \mathbb{Z}\}$, and $\{Z_t : t \in \mathbb{Z}\}$. For simplicity of exposition, we assume that these processes are univariate. We denote $I_X(t) = \{X(s) : s \leq t\}$, $I_Y(t) = \{Y(s) : s \leq t\}$ and $I_Z(t) = \{Z(s) : s \leq t\}$ the information sets which contain all the past and present values of X, Y, and Z until time t, respectively. We denote $I(t)$ the information set that contains $I_X(t)$, $I_Y(t)$ and $I_Z(t)$. $I(t)-A_t$, with $A_t = I_X(t), I_Y(t), I_Z(t)$, contains all the elements of $I(t)$ except those of A_t. Following Florens and Mouchart (1985a,b), the notion of non-causality considered here is defined in terms of orthogonality conditions between subspaces of a Hilbert space of random variables with finite second moments. We denote $L^2 \equiv L^2(\Omega, \mathcal{A}, Q)$ the Hilbert space of random variables defined on a common probability space $(\Omega, \mathcal{A}, Q)$, with covariance as inner product.

For any information set B_t [some Hilbert subspace of L^2], we denote $P[X_{t+1} \mid B_t]$ the best linear forecast of X_{t+1} based on the information set B_t. The corresponding prediction error is $u\left(X_{t+1} \mid B_t\right) = X_{t+1} - P[X_{t+1} \mid B_t]$, and $\sigma^2\big(X_{t+1} \mid B_t\big)$ is the variance of the prediction error. $P[X_{t+1} \mid B_t]$ is the orthogonal projection of X_{t+1} on the subspace B_t. We now remind the reader of the following definitions of *indirect* causality and *spurious* causality from Hsiao (1982) [see also the discussions in Eichler, 2007, 2012]. In the following, Z is used as a *known* auxiliary random variable. However, in the next sections, when we describe the statistical procedures for testing indirect/spurious causality, Z will be treated as an *unknown* auxiliary variable.

Definition 1 (Indirect Causality) *: Y is an indirect cause of X, denoted $Y \overset{ind}{\mapsto} X \mid I(t) - I_Y(t)$, iff*

(i): *Y Granger causes X with respect to the information set $I_X(t)$:*

$$P[X_{t+1} \mid I_X(t)] \neq P[X_{t+1} \mid I(t) - I_Z(t)], \textit{ for some } t > w,$$

(ii): *Y does not Granger cause X with respect to the information set $I(t) - I_Y(t)$:*

$$P[X_{t+1} \mid I(t) - I_Y(t)] = P[X_{t+1} \mid I(t)], \ \forall t > w,$$

(iii): (a) *Y Granger causes Z and* ***(b)*** *Z Granger causes X with respect to the information sets $I(t) - I_Y(t)$ and $I(t) - I_Z(t)$, respectively:*

$$P[Z_{t+1} \,|I(t) - I_Y(t)] \neq P[Z_{t+1} \mid I(t)], \quad P[X_{t+1} \,|I(t) - I_Z(t)] \neq P[X_{t+1} \mid I(t)], \textit{ for some } t > w,$$

where w is a "starting point" which is typically equal to a finite initial date [such as $w = 0$ or 1] or to $-\infty$; in the latter case $I(t)$ is defined for all $t \in \mathbb{Z}$.

Thus, the conditions [(i), (ii), (iii)] of Definition 1 must be satisfied in order to have an indirect causality from Y to X in the presence of an auxiliary variable Z. Similar conditions can be obtained for an indirect causality from X to Y. We now provide the necessary conditions for a spurious causality from Y to X. We distinguish between two types of spurious causality.

Definition 2 (Spurious Causality) *:*

1. Y ***is a spurious cause of type I for*** X ***if***

*1.**(i):*** Y *Granger causes* X *with respect to the information set* $I(t) - I_Y(t)$:

$$P[X_{t+1} \,|\, I(t) - I_Y(t)] \neq P[X_{t+1} \,|\, I(t)], \textit{ for some } t > w,$$

*1.**(ii):*** Y *does not Granger cause* X *with respect to the information set* $I_X(t)$:

$$P[X_{t+1} \,|\, I_X(t)] = P[X_{t+1} \,|\, I(t) - I_Z(t)], \; \forall t > w,$$

*1.**(iii):** **(a)*** Y *Granger causes* Z *and **(b)*** Z *Granger causes* X, *both with respect to the information sets* $I(t) - I_Y(t)$ *and* $I(t) - I_Z(t)$, *respectively,*

$$P[Z_{t+1} \,|I(t) - I_Y(t)] \neq P[Z_{t+1} \,|\, I(t)], \quad P[X_{t+1} \,|I(t) - I_Z(t)] \neq P[X_{t+1} \,|\, I(t)], \textit{ for some } t > w.$$

2. Y ***is a spurious cause of type II for*** X ***if***

*2.**(i):*** Y *Granger causes* X *with respect to the information set* $I_X(t)$:

$$P[X_{t+1} \,|\, I_X(t)] \neq P[X_{t+1} \,|\, I(t) - I_Z(t)], \textit{ for some } t > w,$$

*2.**(ii):*** Y *does not Granger cause* X *with respect to the information set* $I(t) - I_Y(t)$:

$$P[X_{t+1} \,|\, I(t) - I_Y(t)] = P[X_{t+1} \,|\, I(t)], \; \forall t > w,$$

*2.**(iii):** **(a)*** Z *Granger causes* Y *and **(b)*** Z *Granger causes* X, *both with respect to the information set* $I(t) - I_Z(t)$:

$$P[Y_{t+1} \,|I(t) - I_Z(t)] \neq P[Y_{t+1} \,|\, I(t)], \quad P[X_{t+1} \,|I(t) - I_Z(t)] \neq P[X_{t+1} \,|\, I(t)], \textit{ for some } t > w.$$

Definition 2 shows that there are three conditions to satisfy for each type of spurious causality from Y to X. Similar conditions can be obtained for the spurious causality from X to Y.

The above definitions will be used to construct statistical procedures that test for the presence of indirect/spurious causality when the auxiliary variable Z is unknown.

3 Motivation

Unfortunately, most empirical studies on Granger causality analysis ignore indirect and spurious causal effects. This might be explained by the lack of statistical procedures that detect these effects. Up to now, the detection of indirect/spurious causality depends on the knowledge of relevant auxiliary variables, which can happen only in rare cases such as the existence of an economic theory that identifies these variables. The following examples illustrate situations in which indirect/spurious causality happens. To better understand these examples, we need the following lemma [Lütkepohl (1993, pages 231-232)].

Lemma 1 *[**Linear transformation of a VARMA**(p, q) **process**] Let W_t be a K-dimensional, stable, invertible VARMA(p, q) process and let F be an $M \times K$ matrix of rank M. Then the process $W_{t,0} = FW_t$ has a VARMA$(\bar{p}, \bar{q})$ representation with $\bar{p} \leq Kp$ and $\bar{q} \leq (K-1)p + q$.*

If we assume that W_t follows a VAR(p) [i.e. VARMA$(p, 0)$] model, then its linear transformation $W_{t,0} = FW_t$ has a VARMA$(\bar{p}, \bar{q})$ representation with $\bar{p} \leq Kp$ and $\bar{q} \leq (K-1)p$. We now start with the following example on indirect causality.

Example 1 *[**Indirect Causality**] This example illustrates an indirect causality between X and Y transmitted through an auxiliary variable Z. We consider the following VAR(2) model:*

$$\begin{bmatrix} X_t \\ Y_t \\ Z_t \end{bmatrix} = \begin{bmatrix} \mu_X \\ \mu_Y \\ \mu_Z \end{bmatrix} + \begin{bmatrix} \phi^1_{xx} & 0 & \phi^1_{xz} \\ \phi^1_{yx} & \phi^1_{yy} & \phi^1_{yz} \\ 0 & \phi^1_{zy} & \phi^1_{zz} \end{bmatrix} \begin{bmatrix} X_{t-1} \\ Y_{t-1} \\ Z_{t-1} \end{bmatrix} + \begin{bmatrix} 0 & \phi^2_{xy} & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} X_{t-2} \\ Y_{t-2} \\ Z_{t-2} \end{bmatrix} + \begin{bmatrix} \varepsilon_{X_t} \\ \varepsilon_{Y_t} \\ \varepsilon_{Z_t} \end{bmatrix}, \tag{1}$$

with $\phi^1_{xz} \neq 0$, $\phi^1_{zy} \neq 0$, and $\phi^1_{zy} = -\frac{\phi^2_{xy}}{\phi^1_{xz}}$. The error terms ε_{X_t}, ε_{Y_t}, and ε_{Z_t} are assumed to be independent of each other, but they can be relaxed to be serially correlated.

From the first equation of VAR system in (1), we have

$$X_t = \mu_X + \phi^1_{xx} X_{t-1} + \phi^1_{xz} Z_{t-1} + \phi^2_{xy} Y_{t-2} + \varepsilon_{X_t}. \tag{2}$$

From the third equation of VAR system, we get

$$Z_t = \mu_Z + \phi^1_{zy} Y_{t-1} + \phi^1_{zz} Z_{t-1} + \varepsilon_{Z_t}. \tag{3}$$

Using equations (2) and (3), we can easily check that condition (ii) of Definition 1 is satisfied. In other words, if we replace Z_{t-1} in Equation (2) by its expression from Equation (3), we obtain

$$X_t = \mu_X + \phi^1_{xz}\mu_Z + \phi^1_{xx} X_{t-1} + \left(\phi^1_{xz}\phi^1_{zy} + \phi^2_{xy}\right) Y_{t-2} + \phi^1_{xz}\phi^1_{zz} Z_{t-2} + \phi^1_{xz}\varepsilon_{Z_{t-2}} + \varepsilon_{X_t},$$

and since $\phi^1_{zy} = -\phi^2_{xy}/\phi^1_{xz}$, we have $\phi^1_{xz}\phi^1_{zy} + \phi^2_{xy} = -\phi^2_{xy} + \phi^2_{xy} = 0$, hence

$$X_t = \mu_X + \phi^1_{xz}\mu_Z + \phi^1_{xx} X_{t-1} + \phi^1_{xz}\phi^1_{zz} Z_{t-2} + \phi^1_{xz}\varepsilon_{Z_{t-2}} + \varepsilon_{X_t},$$

which indicates that Y does not cause X in the presence of Z. Furthermore, since $\phi^1_{xz} \neq 0$ and $\phi^1_{zy} \neq 0$, from equations (2) and (3) we have Z Granger causes X and Y Granger causes Z, respectively, hence condition (iii) of Definition 1 is satisfied.

Now, to check condition (i) of Definition 1, we use Lemma 1. The marginal process of $(X, Y)'$ can be obtained by taking the matrix F in Lemma 1 as:

$$F = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}.$$

By Lemma 1, the process $(X, Y)'$ has a VARMA($\bar{p}$, $\bar{q}$) representation with $\bar{p} \leq 6$ and $\bar{q} \leq 4$. Thus, in the absence of Z and under some restrictions on the coefficients of VAR(2), X_t can be expressed as a function of the past of Y_t: $X_t = \bar{\mu}_X + \alpha_X X_{t-1} + \alpha_Y Y_{t-1} + \bar{\varepsilon}_{X_t}$, where $\bar{\mu}_X$, α_X, and α_Y are some functions of the coefficients of VAR(2), and $\bar{\varepsilon}_{X_t}$ is a new error term which depends on the error terms of VAR(2) process in (1).

For a real example on indirect causality, the reader can consult the paper by Fackler (1985) who found that neither money nor credit directly cause real output, but these variables play an indirect role in income determination. Fackler also found that interest rates provide the link between the financial and real sectors, thus they can be viewed as the auxiliary variables that transmit the indirect causality from money/credit to income. In Section 8, we use the statistical procedures that we propose in this paper to re-examine these findings and check if interest rates are effectively the appropriate auxiliary variables. Our approach is practical because it does not require a priori knowledge of auxiliary variables. The next example is about spurious causality.

Example 2 *[**Spurious Causality**] This example illustrates a spurious causality of type I from Y to X induced by an auxiliary variable Z. We consider the following VAR(1) model:*

$$\begin{bmatrix} X_t \\ Y_t \\ Z_t \end{bmatrix} = \begin{bmatrix} \mu_X \\ \mu_Y \\ \mu_Z \end{bmatrix} + \begin{bmatrix} \phi^1_{xx} & 0 & \phi^1_{xz} \\ \phi^1_{yx} & \phi^1_{yy} & \phi^1_{yz} \\ \phi^1_{zx} & \phi^1_{zy} & \phi^1_{zz} \end{bmatrix} \begin{bmatrix} X_{t-1} \\ Y_{t-1} \\ Z_{t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{X_t} \\ \varepsilon_{Y_t} \\ \varepsilon_{Z_t} \end{bmatrix}, \tag{4}$$

with $\phi^1_{xz} \neq 0$ and $\phi^1_{zy} \neq 0$. The error terms ε_{X_t}, ε_{Y_t}, and ε_{Z_t} are assumed to be independent of each other, but they can be relaxed to be serially correlated.

From the first equation of VAR system in (4), we have

$$X_t = \mu_X + \phi^1_{xx} X_{t-1} + \phi^1_{xz} Z_{t-1} + \varepsilon_{X_t}. \tag{5}$$

From the third equation of VAR system, we get

$$Z_t = \mu_Z + \phi^1_{zx} X_{t-1} + \phi^1_{zy} Y_{t-1} + \phi^1_{zz} Z_{t-1} + \varepsilon_{Z_t}. \tag{6}$$

Using equations (5) and (6), we can easily check that condition 1.(i) of Definition 2 is satisfied. In other words, if we replace Z_{t-1} in Equation (5) by its expression from Equation (6), we obtain

$$X_t = \mu_X + \phi^1_{xz}\mu_Z + \phi^1_{xx} X_{t-1} + \phi^1_{xz}\phi^1_{zx} X_{t-2} + \phi^1_{xz}\phi^1_{zy} Y_{t-2} + \phi^1_{xz}\phi^1_{zz} Z_{t-2} + \phi^1_{xz}\varepsilon_{Z_{t-1}} + \varepsilon_{X_t},$$

which, for $\phi^1_{xz} \neq 0$ and $\phi^1_{zy} \neq 0$, indicates that Y does cause X in the presence of Z. Furthermore, since $\phi^1_{xz} \neq 0$ and $\phi^1_{zy} \neq 0$, from equations (5) and (6) we have Z Granger causes X and Y Granger causes Z, respectively, hence condition 1.(iii) of Definition 2 is satisfied.

To check condition 1.(ii) of Definition 2, we use Lemma 1. The marginal process of X can be obtained by taking the matrix F in Lemma 1 as $F = [\ 1 \quad 0 \quad 0\ ]$. According to Lemma 1, the process X has an ARMA($\bar{p}$, $\bar{q}$) representation with $\bar{p} \leq 6$ and $\bar{q} \leq 4$. Thus, in the absence of Z and under some restrictions on the coefficients of VAR(1), X_t can be expressed as a function of its own past:

$$X_t = \bar{\mu}_X + \phi_X X_{t-1} + \bar{\varepsilon}_{X_t},$$

where $\bar{\mu}_X$ and ϕ_X are some functions of the coefficients of VAR(1) and $\bar{\varepsilon}_{X_t}$ is a moving average process which depends on the error terms of VAR(1) process in (4).

4 Testing for indirect and spurious causalities

In this section, we describe the testing procedures that we use to test for indirect and spurious causality. Our statistical procedures are based on testing each condition in Definition 1 (or Definition 2) separately. As discussed in Hsiao (1980, page 22), the advantage of testing separately each condition against the maintained hypothesis (unconstrained model) is that the size of the test for each tested condition is the same compared to a sequential test where the size has to be different in order to guarantee the control of the Type I error. However, as have been argued by Hsiao (1980, page 22), the disadvantage of running separate tests is that it is difficult to reject the null hypothesis- of indirect or spurious no causality- because "*an otherwise significant coefficients might be contaminated by other insignificant coefficients*" of the estimated unconstrained model. Hsiao (1980) discusses an alternative sequential test for testing indirect and spurious no-causality. Regarding the sequential test, Hsiao (1980, page 22) writes: "*The advantage of testing each hypothesis [condition] sequentially* *by treating the previously accepted null hypothesis as the maintained hypothesis is that the test is more sharply focused. The disadvantage is that a test of ψ_{ij} [$\psi_{ij} = 0$, which corresponds to one of the conditions of indirect or spurious no-causality] may either be accepted or rejected depending on the order it is tested.*" He adds that using a fixed significance level α in a sequential test will "*increase the Type I error from $P_1 = \alpha$ to $P_2 = \alpha + (1-\alpha)\alpha$, to $P_3 = \alpha + 2(1-\alpha)\alpha + (1-\alpha)^2\alpha$,... when we change the order of testing $\psi_{ij} = 0$ from the first to the second and so on.*" For these reasons we decide to use the first alternative namely separate tests.

4.1 Testing for indirect causality

Definition 1 shows that there are three conditions that must be satisfied in order to have an indirect causality from Y to X. The first one [condition (i)] is simple to test as it only involves the observed variables X and Y. However, the other two conditions [conditions (ii) and (iii)] are difficult because the auxiliary variable Z is unknown, thus not observed. In the following, we propose to use factor analysis to identify Z. In particular, we use as a proxy of Z the factor(s) that we extract from a big data using principal component analysis. Formally, condition (i) can be checked using the regression

$$X_{t+1} = \mu + \sum_{i=1}^{p} \beta_i X_{t+1-i} + \sum_{j=1}^{q} \alpha_j Y_{t+1-j} + \varepsilon_{t+1} \tag{7}$$

and a Wald-test for testing the null hypothesis

$$H_0 : \alpha_1 = ... = \alpha_q = 0 \quad \text{vs} \quad H_1 : \text{No } H_0.$$

If H_0 is rejected and Z is observed, we proceed to verify the condition (ii) using the regression

$$X_{t+1} = \eta + \sum_{i=1}^{\bar{p}} \gamma_i X_{t+1-i} + \sum_{j=1}^{\bar{q}} \lambda_j Y_{t+1-j} + \sum_{l=1}^{\bar{h}} \theta_l Z_{t+1-l} + e_{t+1} \tag{8}$$

and a Wald-test for testing the null hypothesis

$$\bar{H}_0 : \lambda_1 = ... = \lambda_{\bar{q}} = 0 \quad \text{vs} \quad \bar{H}_1 : \text{No } \bar{H}_0.$$

If $\bar{H}_0$ is not rejected, we proceed to check the condition (iii) using the regressions

$$Z_{t+1} = \nu + \sum_{i=1}^{\dot{p}} \kappa_i X_{t+1-i} + \sum_{j=1}^{\dot{q}} \psi_j Y_{t+1-j} + \sum_{l=1}^{\dot{h}} \rho_l Z_{t+1-l} + u_{t+1}, \tag{9}$$

$$X_{t+1} = \varpi + \sum_{i=1}^{\ddot{p}} \xi_i X_{t+1-i} + \sum_{j=1}^{\ddot{q}} \delta_j Y_{t+1-j} + \sum_{l=1}^{\ddot{h}} \varsigma_l Z_{t+1-l} + \epsilon_{t+1}, \tag{10}$$

and the Wald-tests for testing the null hypotheses

$$\dot{H}_0 : \psi_1 = ... = \psi_{\dot{q}} = 0 \quad \text{vs} \quad \dot{H}_1 : \text{No } \dot{H}_0,$$
$$\ddot{H}_0 : \varsigma_1 = ... = \varsigma_{\ddot{h}} = 0 \quad \text{vs} \quad \ddot{H}_1 : \text{No } \ddot{H}_0.$$

If both $\dot{H}_0$ and $\ddot{H}_0$ are rejected, we conclude that Y indirectly causes X.

In practice, however, Z is not observed but it can be proxied by the factors that we extract from a big data that contains all economic variables that are available to econometricians. Formally, we consider an N-dimensional vector of large number of economic time series $w_t = (w_{t,1}, ..., w_{t,N})'$ observed at each time t. We denote by $W = (w_1, ..., w_T)'$ the $(T \times N)$-dimensional matrix in which the t-th row is given by w_t. We assume that k common factors f_t are associated with the N-dimensional vector w_t according to the following equation:

$$w_t = \Lambda f_t + \varepsilon_t, \tag{11}$$

where Λ is an $(N \times k)$-dimensional matrix of factor loadings and ε_t's are vectors of idiosyncratic shocks that could be cross-sectionally and temporally dependent.

To extract f from the $(T \times N)$-dimensional matrix W, we consider the factor model in Equation (11), which associates the N-dimensional vector w_t with the k common factors f_t. However, we remind the reader that the factors f_t and loadings Λ are not identified simultaneously since

$$w_t = \Lambda f_t + \varepsilon_t = \Lambda\Delta^{-1}\Delta f_t + \varepsilon_t = \Lambda^* f_t^* + \varepsilon_t, \tag{12}$$

for $\Lambda^* = \Lambda\Delta^{-1}$, $f_t^* = \Delta f_t$, and Δ a $(k \times k)$-dimensional positive definite matrix. Thus, we will only estimate the space spanned by the true factors and not the factors themselves. Simultaneous identification of the factors is, however, not essential for the statistical procedures that we propose to test for indirect/spurious causality. In other words, we only need to control for all available variables and it doesn't matter how they are weighted in the factors. Using Equation (11) and each element $w_{t,j}$ of the vector w_t, the factor model is given by

$$w_{t,j} = \vartheta_j' f_t + \varepsilon_{t,j}, \text{ for } j = 1, \ldots, N,$$

where ϑ_j is a k-dimensional vector of factor loadings given by the j-th row of the matrix Λ. The factors f_t will be extracted using the principal component analysis (PCA) based on the following nonlinear least squares objective function

$$V(\tilde{f},\ \tilde{\Lambda}) = \frac{1}{NT}\sum_{j=1}^{N}\sum_{t=1}^{T}\left(w_{t,j} - \tilde{\vartheta}_j'\tilde{f}_t\right)^2. \tag{13}$$

The function $V(\tilde{f}, \tilde{\Lambda})$ depends on the hypothetical values of the factors $\tilde{f} = (\tilde{f}_1, ..., \tilde{f}_T)'$ and factor loadings $\tilde{\Lambda} = (\tilde{\vartheta}_1, ..\tilde{\vartheta}_N)'$. Let $\hat{f}$ and $\hat{\Lambda}$ be the minimizers of $V(\tilde{f}, \tilde{\Lambda})$. After concentrating out $\hat{f}$, minimizing $V(\tilde{f}, \tilde{\Lambda})$ is equivalent to maximizing $tr(\tilde{\Lambda}'Y'Y\tilde{\Lambda})$ subject to $\frac{\tilde{\Lambda}'\tilde{\Lambda}}{N} = I_k$, where $tr(\cdot)$ denotes the trace of a matrix and I_k is a $(k \times k)$-dimensional identity matrix. This represents the classical principal components problem that can be solved by setting the columns of $\hat{\Lambda}$ to be equal to the eigenvectors of $W'W$ corresponding to the k largest eigenvalues. The resulting principal components estimator of the matrix of the factors $f = (f_1, ..., f_T)'$ is:

$$\hat{f} = (\hat{f}_1, ..., \hat{f}_T)' = \frac{W\hat{\Lambda}}{N}. \tag{14}$$

The computation of $\hat{f}$ requires the calculation of the eigenvectors of the $N \times N$ matrix $W'W$ for $N > T$. Under some regularity conditions, Bai and Ng (2002) show that $\hat{f}$ is a consistent estimator of f; see Theorem 1 of Bai and Ng (2002).

We next replace Z in equations (8)-(10) by the extracted factors $\hat{f}$. To simplify our analysis, we let $k = 1$. In practice the selection of number of factors can be performed using the information criteria suggested by Bai and Ng (2002). However, using only the first factor (that explains most of the variation in the data) should be enough as it represents a linear combination of all variables in the big data W. Now, to check condition (ii) of Definition 1, we use the feasible regression

$$X_{t+1} = \eta + \sum_{i=1}^{\bar{p}} \gamma_i X_{t+1-i} + \sum_{j=1}^{\bar{q}} \lambda_j Y_{t+1-j} + \sum_{l=1}^{\bar{h}} \theta_l \hat{f}_{t+1-l} + e_{t+1} \tag{15}$$

and the Wald-test for testing

$$\bar{H}_0 : \lambda_1 = ... = \lambda_{\bar{q}} = 0 \quad \text{vs} \quad \bar{H}_1 : \text{No } \bar{H}_0. \tag{16}$$

Similarly, to verify condition (iii), we use the feasible regressions

$$\hat{f}_{t+1} = \nu + \sum_{i=1}^{\dot{p}} \kappa_i X_{t+1-i} + \sum_{j=1}^{\dot{q}} \psi_j Y_{t+1-j} + \sum_{l=1}^{\dot{h}} \rho_l \hat{f}_{t+1-l} + u_{t+1}, \tag{17}$$

$$X_{t+1} = \varpi + \sum_{i=1}^{\ddot{p}} \xi_i X_{t+1-i} + \sum_{j=1}^{\ddot{q}} \delta_j Y_{t+1-j} + \sum_{l=1}^{\ddot{h}} \varsigma_l \hat{f}_{t+1-l} + \epsilon_{t+1} \tag{18}$$

and the Wald-tests for testing the null hypotheses:

$$\dot{H}_0 : \psi_1 = ... = \psi_{\dot{q}} = 0 \text{ vs } \dot{H}_1 : \text{No } \dot{H}_0 \tag{19}$$

$$\ddot{H}_0 : \varsigma_1 = ... = \varsigma_{\ddot{h}} = 0 \quad \text{vs} \quad \ddot{H}_1 : \text{No} \quad \ddot{H}_0. \tag{20}$$

If both $\dot{H}_0$ and $\ddot{H}_0$ are rejected, we conclude that Y indirectly causes X.

4.2 Testing for spurious causality

To test for the spurious causality from Y to X, we need to check the conditions of Definition 2. For the spurious causality of type I, we have to check if conditions 1.(i) -1.(iii) are satisfied. To test condition 1.(i), we use the feasible regression

$$X_{t+1} = \mu + \sum_{i=1}^{p} \beta_i X_{t+1-i} + \sum_{j=1}^{q} \alpha_j Y_{t+1-j} + \sum_{l=1}^{k} \pi_j \hat{f}_{t+1-l} + \varepsilon_{t+1}, \tag{21}$$

and the Wald-test for testing the null hypothesis

$$H_0^{(1)} : \alpha_1 = ... = \alpha_q = 0 \quad \text{vs} \quad H_1^{(1)} : \text{No } H_0^{(1)}. \tag{22}$$

If $H_0^{(1)}$ is rejected, we proceed to test condition 1.(ii) using the regression

$$X_{t+1} = \eta + \sum_{i=1}^{\bar{p}} \beta_i X_{t+1-i} + \sum_{j=1}^{\bar{q}} \alpha_j Y_{t+1-j} + e_{t+1},$$

and a Wald-test for testing the null hypothesis

$$\bar{H}_0^{(1)} : \alpha_1 = ... = \alpha_{\bar{q}} = 0 \quad \text{vs} \quad \bar{H}_1^{(1)} : \text{No } \bar{H}_0^{(1)}. \tag{23}$$

If $\bar{H}_0^{(1)}$ is not rejected, we proceed to check the condition 1.(iii) before deciding about the presence of spurious causality of type I. Condition 1.(iii) can be verified using the following feasible regressions:

$$\hat{f}_{t+1} = \nu + \sum_{i=1}^{\dot{p}} \kappa_i X_{t+1-i} + \sum_{j=1}^{\dot{q}} \psi_j Y_{t+1-j} + \sum_{l=1}^{\dot{h}} \rho_j \hat{f}_{t+1-l} + u_{t+1}, \tag{24}$$

$$X_{t+1} = \varpi + \sum_{i=1}^{\ddot{p}} \xi_i X_{t+1-i} + \sum_{j=1}^{\ddot{q}} \delta_j Y_{t+1-j} + \sum_{l=1}^{\ddot{h}} \varsigma_j \hat{f}_{t+1-l} + \epsilon_{t+1}, \tag{25}$$

and the Wald-tests for testing the null hypotheses:

$$\dot{H}_0^{(1)} : \psi_1 = ... = \psi_{\dot{q}} = 0 \quad \text{vs} \quad \dot{H}_1^{(1)} : \text{No } \dot{H}_0^{(1)}, \tag{26}$$

$$\ddot{H}_0^{(1)} : \varsigma_1 = ... = \varsigma_{\ddot{h}} = 0 \quad \text{vs} \quad \ddot{H}_1^{(1)} : \text{No } \ddot{H}_0^{(1)}. \tag{27}$$

If both $\dot{H}_0^{(1)}$ and $\ddot{H}_0^{(1)}$ are rejected, we conclude that Y spuriously (type I) causes X.

For the spurious causality of type II, we need to check conditions 2.(i)-2.(iii). To test condition 2.(i), we use the regression

$$X_{t+1} = \mu + \sum_{i=1}^{p} \beta_i X_{t+1-i} + \sum_{j=1}^{q} \alpha_j Y_{t+1-j} + \varepsilon_{t+1},$$

and a Wald-test for testing the null hypothesis

$$H_0^{(2)} : \alpha_1 = ... = \alpha_q = 0 \quad \text{vs} \quad H_1^{(2)} : \text{No } H_0^{(2)}.$$

If $H_0^{(2)}$ is rejected, we proceed to test condition 2.(ii) using the feasible regression

$$X_{t+1} = \eta + \sum_{i=1}^{\bar{p}} \gamma_i X_{t+1-i} + \sum_{j=1}^{\bar{q}} \lambda_j Y_{t+1-j} + \sum_{l=1}^{\bar{h}} \theta_l \hat{f}_{t+1-l} + e_{t+1}$$

and a Wald-test for testing the null hypothesis

$$\bar{H}_0^{(2)} : \lambda_1 = ... = \lambda_{\bar{q}} = 0 \quad \text{vs} \quad \bar{H}_1^{(2)} : \text{No } \bar{H}_0^{(2)}. \tag{28}$$

Thereafter, if $\bar{H}_0^{(2)}$ is not rejected, we next check condition 2.(iii) before deciding about the presence of spurious causality of type II. Condition 2.(iii) can be verified using the following feasible regressions:

$$\begin{aligned} Y_{t+1} &= \nu + \textstyle\sum_{i=1}^{\dot{p}} \kappa_i X_{t+1-i} + \sum_{j=1}^{\dot{q}} \psi_j Y_{t+1-j} + \sum_{l=1}^{\dot{h}} \rho_j \hat{f}_{t+1-l} + u_{t+1}, \\ Y_{t+1} &= \nu + \textstyle\sum_{i=1}^{\dot{p}} \kappa_i X_{t+1-i} + \sum_{j=1}^{\dot{q}} \psi_j Y_{t+1-j} + \sum_{l=1}^{\dot{h}} \rho_j \hat{f}_{t+1-l} + u_{t+1}, \end{aligned}$$

and the Wald-tests for testing the null hypotheses:

$$\begin{aligned} \dot{H}_0^{(2)} &: \rho_1 = ... = \rho_{\dot{h}} = 0 \quad \text{vs} \quad \dot{H}_1^{(2)} : \text{No } \dot{H}_0^{(2)}, \\ \ddot{H}_0^{(2)} &: \varsigma_1 = ... = \varsigma_{\ddot{h}} = 0 \quad \text{vs} \quad \ddot{H}_1^{(2)} : \text{No } \ddot{H}_0^{(2)}. \end{aligned} \tag{29}$$

If both $\dot{H}_0^{(2)}$ and $\ddot{H}_0^{(2)}$ are rejected, we conclude that Y spuriously (type II) causes X.

5 Asymptotic distributions

In this section, we study the properties of indirect/spurious causality tests described in Section 4. In particular, we use the results of Bai and Ng (2006) [see also their working paper Bai and Ng (2005)] to provide the asymptotic distributions of these tests in the presence of a consistent estimator of factors f. Bai and Ng's (2006) methodology consists of estimating the common factors f from a panel of data- which includes a large number of series (large N)- by the method of principal components and then augmenting a standard regression with the estimated factors. They show that the ordinary least squares estimates obtained from these factor-augmented regressions are square root T consistent and asymptotically normal if $\sqrt{T}/N \to 0$. Their approach, however, was not motivated by the importance of detecting indirect and spurious causalities. In this paper, we show that their methodology can be applied to test for the presence of indirect/spurious causality by helping overcome the problem of unknown relevant auxiliary variables that we discussed in the previous sections. The assumptions required for the consistency of f are given by the following conditions.

Assumption A: *For a positive generic constant* δ, *we assume that:* $(\mathbf{A1})$ $E\|f_t\|^4 \leqslant \delta < \infty$ *and* $T^{-1}\sum_{t=1}^{T} f_t f_t' \xrightarrow{p} \Sigma_f > 0$, *where* Σ_f *is a* $(k \times k)$ *non-random positive definite matrix and* $\|.\|$ *denotes the Euclidean norm;* $(\mathbf{A2})$ *If* Λ *is deterministic, then* $\|\lambda_j\| \leqslant \delta < \infty$, *where* λ_j *is the* j*-th row of the factor loadings matrix* Λ. *If it is stochastic, then* $E\|\lambda_j\|^4 \leqslant \delta$. *Furthermore,* $N^{-1}\Lambda'\Lambda \xrightarrow{p} \Sigma_\Lambda > 0$, *as* $N \longrightarrow \infty$, *where* Σ_Λ *is a* $(k \times k)$ *non-random matrix;* $(\mathbf{A3})$ *For all* N *and* T, *we have:* ***(i)*** $E(\varepsilon_{t,i}) = 0$ *and* $E|\varepsilon_{t,i}|^8 \leq \delta$; ***(ii)*** *For* $N^{-1}E(\varepsilon_s'\varepsilon_t) = \gamma_N(s,t)$, *we have* $|\gamma_N(s,s)| \leq \delta$, *where* $s = 1, ...T$, *and* $T^{-1}\sum_{1\leq s,t\leq N} |\gamma_N(s,t)| \leq \delta$; ***(iii)*** *For* $E(\varepsilon_{t,i}\varepsilon_{t,j}) = \tau_{ij,t}$, *we have* $|\tau_{t,ij}| \leq \tau_{ij}$, $\forall t$, *and* $N^{-1}\sum_{1\leq i,j\leq N} |\tau_{ij}| \leq \delta$; ***(iv)*** *For* $E(\varepsilon_{t,i}\varepsilon_{s,j}) = \tau_{ts,ij}$, *we have* $(TN)^{-1}\sum_{1\leq i,j,s,t\leq N} |\tau_{ts,ij}| \leq \delta$; *and* ***(v)*** $E\left|N^{-\frac{1}{2}}\sum_{i=1}^{N}\left[\varepsilon_{t,i}\varepsilon_{s,j} - E(\varepsilon_{t,i}\varepsilon_{s,j})\right]\right|^4 \leq \delta, \forall s,t$; and $(\mathbf{A4})$ $E\left(N^{-1}\sum_{i=1}^{N}\left\|T^{-1}\sum_{t=1}^{T} f_t\varepsilon_{t,i}\right\|^2\right) \leq \delta$.

Assumptions $(\mathbf{A1})$ and $(\mathbf{A2})$ are standard in the literature on factor analysis; see Stock and Watson (2002), Bai and Ng (2002) and Bai (2003). They represent moment conditions on factors f_t and factor loadings Λ, and they ensure that the factors are non-degenerate and their contribution to the variance of the data is nontrivial. Assumptions $(\mathbf{A3})$**-(i)** to $(\mathbf{A3})$**-(v)** allow for heteroskedasticity and weak correlation between the components of the vector of idiosyncratic shocks ε_t in (11). Under these assumptions both cross-sectional and serial correlations are allowed.

5.1 Indirect causality

We derive the asymptotic distributions of tests of conditions of indirect causality in Definition 1. We focus on testing conditions (ii) and (iii), because the test of condition (i) depends only on the observed variables X and Y. We introduce the following notations. Define the vector of parameters $\tau_1^{Ind} = (\eta, \gamma', \lambda', \theta')'$, where $\gamma = (\gamma_1, \ldots, \gamma_{\bar{p}})'$, and λ and θ can be defined in similar way. Let $\hat{z}_{t1} = (1, X_t, \ldots, X_{t+1-\bar{p}}, Y_t, \ldots, Y_{t+1-\bar{q}}, \hat{f}_t, \ldots, \hat{f}_{t+1-\bar{h}})'$ and $\hat{\tau}_1^{Ind} = (\hat{\eta}, \hat{\gamma}', \hat{\lambda}', \hat{\theta}')'$, where $\hat{\tau}_1^{Ind}$ is the least squares estimates obtained from the regression of X_{t+1} on the constant, $(X_t, \ldots, X_{t+1-\bar{p}})'$, $(Y_t, \ldots, Y_{t+1-\bar{q}})'$, and the estimated factors $(\hat{f}_t, \ldots, \hat{f}_{t+1-\bar{h}})'$. Henceforth, $0_{m,n}$ is the $m \times n$ matrix of zeros and I_n is the $n \times n$ identity matrix.

First, we test condition (ii) using the following Wald-test statistic, which tests that all the coefficients of $Y_t, \ldots, Y_{t+1-\bar{q}}$ in the regression (15) are jointly equal to zero:

$$W_T^{Ind,\lambda} = \left(\sqrt{T}R^{Ind,\lambda}\hat{\tau}_1^{Ind}\right)\left(R^{Ind,\lambda}\hat{\Sigma}_{\tau_1^{Ind}}R^{Ind,\lambda\prime}\right)^{-1}\left(\sqrt{T}R^{Ind,\lambda}\hat{\tau}_1^{Ind}\right)', \tag{30}$$

where the selection matrix $R^{Ind,\lambda} = (0_{\bar{q},1+\bar{p}}, I_{\bar{q}}, 0_{\bar{q},\bar{h}})$ and

$$\hat{\Sigma}_{\tau_1^{Ind}} = \left(\frac{1}{T}\sum_{t=1}^{T-1} \hat{z}_{t1}\hat{z}_{t1}'\right)^{-1} \left(\frac{1}{T}\sum_{t=1}^{T-1} \hat{e}_{t+1}^2 \hat{z}_{t1}\hat{z}_{t1}'\right) \left(\frac{1}{T}\sum_{t=1}^{T-1} \hat{z}_{t1}\hat{z}_{t1}'\right)^{-1} \tag{31}$$

is the estimated variance-covariance matrix, with $\hat{e}_{t+1} := X_{t+1} - \hat{z}_{t1}'\hat{\tau}_1^{Ind}$ the least squares residuals. The following theorem demonstrates that the $W_T^{Ind,\lambda}$ test statistic is asymptotically distributed as a chi-squared distribution with $\bar{q}$ degrees of freedom [The proof of Theorem 1 and those of the theoretical results below can be found in a separate companion Appendix, which is available online].

Theorem 1 *: Let Assumption **A** hold. Under the null hypothesis (16), if $\sqrt{T}/N \to 0$,*

$$W_T^{Ind,\lambda} \to_d \chi^2_{\bar{q}},$$

where the $W_T^{Ind,\lambda}$ test statistic is defined in (30).

Theorem 1 is stated under the general case of heteroskedasticity. However, the above result is still valid under homoskedasticity, with a consistent estimator of variance-covariance matrix given by

$$\hat{\Sigma}^*_{\tau_1^{Ind}} = \hat{\sigma}_e^2 \left(\frac{1}{T}\sum_{t=1}^{T-1} \hat{z}_{t1}\hat{z}_{t1}'\right)^{-1}, \tag{32}$$

where $\hat{\sigma}_e^2 = T^{-1}\sum_{t=1}^{T-1} \hat{e}_{t+1}^2$. It can be shown that the difference between the estimators in (31) and (32) is asymptotically negligible under homoskedasticity. Furthermore, the proof of Theorem 1 indicates that for $\sqrt{T}/N \to 0$, having estimated factors as regressors does not affect the root-T consistency of the least squares estimates of τ_1^{Ind}, except that the variance-covariance matrix $\Sigma_{\tau_1^{Ind}}$ will be different, which can be consistently estimated by $\hat{\Sigma}_{\tau_1^{Ind}}$. However, if $\sqrt{T}/N \to c > 0$, then there are two additional terms that do not vanish asymptotically, thus $\hat{\tau}_1^{Ind}$ will have an asymptotic bias term, reflecting the contribution of factors estimation uncertainty. For details on the source of the bias term the reader is referred to the proof of Lemma 3 in the separate companion Appendix. Unfortunately, this bias term complicates our analysis and requires additional work that is beyond the scope of this paper [e.g., to allow for $\sqrt{T}/N \to c > 0$, Ludvigson and Ng (2011) propose an analytical bias correction and Gonçalves and Perron (2014) propose a residual-based bootstrap method]. Hence, we focus on the case of $\sqrt{T}/N \to 0$ and leave the general case for future study.

We now provide the test statistics that one can use to test condition (iii) or equivalently the null hypotheses (19) and (20). The latter hypotheses can be tested using the following $W_T^{Ind,\psi}$ and $W_T^{Ind,\varsigma}$ test statistics that test if all the coefficients of the vectors $(Y_t, \ldots, Y_{t+1-\dot{q}})'$ and $(\hat{f}_t, \ldots, \hat{f}_{t+1-\ddot{h}})'$ in the regressions (17) and (18) are jointly equal to zero, respectively. The $W_T^{Ind,\psi}$ test statistic for testing the null hypothesis (19) is given by:

$$W_T^{Ind,\psi} = \left(\sqrt{T} R^{Ind,\psi} \hat{\tau}_2^{Ind}\right) \left(R^{Ind,\psi} \hat{\Sigma}_{\tau_2^{Ind}} R^{Ind,\psi\prime}\right)^{-1} \left(\sqrt{T} R^{Ind,\psi} \hat{\tau}_2^{Ind}\right)', \tag{33}$$

where $\hat{\tau}_2^{Ind} = (\hat{\nu}, \hat{\kappa}', \hat{\psi}', \hat{\rho}')'$, the selection matrix $R^{Ind,\psi} = (0_{\dot{q},1+\dot{p}}, I_{\dot{q}}, 0_{\dot{q},\dot{h}})$, and

$$\hat{\Sigma}_{\tau_2^{Ind}} = \left(\frac{1}{T}\sum_{t=1}^{T-1} \hat{z}_{t2}\hat{z}_{t2}'\right)^{-1} \left(\frac{1}{T}\sum_{t=1}^{T-1} \hat{u}_{t+1}^2 \hat{z}_{t2}\hat{z}_{t2}'\right) \left(\frac{1}{T}\sum_{t=1}^{T-1} \hat{z}_{t2}\hat{z}_{t2}'\right)^{-1},$$

with $\hat{u}_{t+1} = \hat{f}_{t+1} - \hat{z}_{t2}'\hat{\tau}_2^{Ind}$ the least squares residuals from the regression in (17) and

$$\hat{z}_{t2} = (1, X_t, \ldots, X_{t+1-\dot{p}}, Y_t, \ldots, Y_{t+1-\dot{q}}, \hat{f}_t, \ldots, \hat{f}_{t+1-\dot{h}})'.$$

Similarly, the $W_T^{Ind,\varsigma}$ test statistic for testing the null hypothesis (20) is given by:

$$W_T^{Ind,\varsigma} = \left(\sqrt{T} R^{Ind,\varsigma} \hat{\tau}_3^{Ind}\right) \left(R^{Ind,\varsigma} \hat{\Sigma}_{\tau_3^{Ind}} R^{Ind,\varsigma\prime}\right)^{-1} \left(\sqrt{T} R^{Ind,\varsigma} \hat{\tau}_3^{Ind}\right)', \tag{34}$$

where $\hat{\tau}_3^{Ind} = (\hat{\varpi}, \hat{\xi}', \hat{\delta}', \hat{\varsigma}')'$, the selection matrix $R^{Ind,\varsigma} = (0_{\ddot{h},1+\ddot{p}+\ddot{q}}, I_{\ddot{h}})$, and

$$\hat{\Sigma}_{\tau_3^{Ind}} = \left(\frac{1}{T}\sum_{t=1}^{T-1} \hat{z}_{t3}\hat{z}_{t3}'\right)^{-1} \left(\frac{1}{T}\sum_{t=1}^{T-1} \hat{\epsilon}_{t+1}^2 \hat{z}_{t3}\hat{z}_{t3}'\right) \left(\frac{1}{T}\sum_{t=1}^{T-1} \hat{z}_{t3}\hat{z}_{t3}'\right)^{-1},$$

with $\hat{\epsilon}_{t+1} = X_{t+1} - \hat{z}_{t3}'\hat{\tau}_3^{Ind}$ the least squares residuals from the regression in (18) and

$$\hat{z}_{t3} = (1, X_t, \ldots, X_{t+1-\ddot{p}}, Y_t, \ldots, Y_{t+1-\ddot{q}}, \hat{f}_t, \ldots, \hat{f}_{t+1-\ddot{h}})'.$$

The following theorem shows that the Wald-type test statistics $W_T^{Ind,\psi}$ and $W_T^{Ind,\varsigma}$ are asymptotically distributed as chi-squared distributions with $\dot{q}$ and $\ddot{h}$ degrees of freedom, respectively.

Theorem 2 : *Let Assumption **A** hold. Under the null hypotheses (19) and (20), if $\sqrt{T}/N \to 0$,*

$$W_T^{Ind,\psi} \to_d \chi^2_{\dot{q}} \quad \text{and} \quad W_T^{Ind,\varsigma} \to_d \chi^2_{\ddot{h}},$$

where the $W_T^{Ind,\psi}$ and $W_T^{Ind,\varsigma}$ test statistics are defined in (33) and (34), respectively.

Similar to Theorem 1, the result in Theorem 2 is still valid under homoskedasticity, with consistent estimators of the variance-covariance matrices $\Sigma_{\tau_2^{Ind}}$ and $\Sigma_{\tau_3^{Ind}}$ given by $\hat{\Sigma}^*_{\tau_2^{Ind}} = \hat{\sigma}_u^2(\sum_{t=1}^{T-1} \hat{z}_{t2}\hat{z}_{t2}'/T)^{-1}$ and $\hat{\Sigma}^*_{\tau_3^{Ind}} = \hat{\sigma}_\epsilon^2(\sum_{t=1}^{T-1} \hat{z}_{t3}\hat{z}_{t3}'/T)^{-1}$, respectively, where $\hat{\sigma}_u^2 = \sum_{t=1}^{T-1} \hat{u}_{t+1}^2/T$ and $\hat{\sigma}_\epsilon^2 = \sum_{t=1}^{T-1} \hat{\epsilon}_{t+1}^2/T$. We can show that the difference between the estimators $\hat{\Sigma}_{\tau_2^{Ind}}$ and $\hat{\Sigma}^*_{\tau_2^{Ind}}$ (resp. $\hat{\Sigma}_{\tau_3^{Ind}}$ and $\hat{\Sigma}^*_{\tau_3^{Ind}}$) is negligible under homoskedasticity assumption. Moreover, our results show that for $\sqrt{T}/N \to 0$, having estimated factors as regressand or regressors does not affect the root-T consistency of the least squares estimates of τ_2^{Ind} and τ_3^{Ind}, except that the variance-covariance matrices $\Sigma_{\tau_2^{Ind}}$ and $\Sigma_{\tau_3^{Ind}}$ will have different expressions, which can be consistently estimated by $\hat{\Sigma}_{\tau_2^{Ind}}$ and $\hat{\Sigma}_{\tau_3^{Ind}}$, respectively. However, if $\sqrt{T}/N \to c > 0$, then there are additional terms that do not vanish even asymptotically, thus $\hat{\tau}_2^{Ind}$ and $\hat{\tau}_3^{Ind}$ will have asymptotic bias terms; see the remarks after Theorem 1.

5.2 Spurious causality

We now study the asymptotic properties of tests of conditions of Definition 2. For type I spurious causality, we focus on providing the asymptotic distributions of tests of conditions (i) and (iii), since the test of condition (ii) only depends on observed variables X and Y and can be tested using the standard test. For type II spurious causality, we only derive the asymptotic distributions of tests of conditions (ii) and (iii), again because the test of condition (i) involves observed variables only.

Regarding the type I spurious causality, conditions (i), (iii)-(a) and (iii)-(b) can be tested using the following Wald-type test statistics

$$\begin{aligned} W_T^{SI,\alpha} &= \left(\sqrt{T}R^{si,\alpha}\hat{\tau}_1^{si}\right)\left(R^{si,\alpha}\hat{\Sigma}_{\hat{\tau}_1^{si}}R^{si,\alpha\prime}\right)^{-1}\left(\sqrt{T}R^{si,\alpha}\hat{\tau}_1^{si}\right)', \\ W_T^{SI,\psi} &= \left(\sqrt{T}R^{si,\psi}\hat{\tau}_2^{si}\right)\left(R^{si,\psi}\hat{\Sigma}_{\hat{\tau}_2^{si}}R^{si,\psi\prime}\right)^{-1}\left(\sqrt{T}R^{si,\psi}\hat{\tau}_2^{si}\right)', \\ W_T^{SI,\varsigma} &= \left(\sqrt{T}R^{si,\varsigma}\hat{\tau}_3^{si}\right)\left(R^{si,\varsigma}\hat{\Sigma}_{\hat{\tau}_3^{si}}R^{si,\varsigma\prime}\right)^{-1}\left(\sqrt{T}R^{si,\varsigma}\hat{\tau}_3^{si}\right)', \end{aligned} \tag{35}$$

respectively, where $\hat{\tau}_1^{si} = (\hat{\mu}, \hat{\beta}', \hat{\alpha}', \hat{\pi}')'$, $\hat{\tau}_2^{si} = (\hat{\nu}, \hat{\kappa}', \hat{\psi}', \hat{\rho}')'$, $\hat{\tau}_3^{si} = (\hat{\varpi}, \hat{\xi}', \hat{\delta}', \hat{\varsigma}')'$, the selection matrices $R^{si,\alpha} = (0_{q,1+p}, I_q, 0_{q,k})$, $R^{si,\psi} = (0_{\dot{q},1+\dot{p}}, I_{\dot{q}}, 0_{\dot{q},\dot{h}})$, $R^{si,\varsigma} = (0_{\ddot{h},1+\ddot{p}+\ddot{q}}, I_{\ddot{h}})$, and

$$\begin{aligned} \hat{\Sigma}_{\tau_1^{si}} &= \left(\frac{1}{T}\sum_{t=1}^{T-1}\hat{z}_{1t}^{si}\hat{z}_{t1}^{si\prime}\right)^{-1}\left(\frac{1}{T}\sum_{t=1}^{T-1}\hat{\varepsilon}_{t+1}^2\hat{z}_{t1}^{si}\hat{z}_{t1}^{si\prime}\right)\left(\frac{1}{T}\sum_{t=1}^{T-1}\hat{z}_{t1}^{si}\hat{z}_{t1}^{si\prime}\right)^{-1}, \\ \hat{\Sigma}_{\tau_2^{si}} &= \left(\frac{1}{T}\sum_{t=1}^{T-1}\hat{z}_{t2}^{si}\hat{z}_{t2}^{si\prime}\right)^{-1}\left(\frac{1}{T}\sum_{t=1}^{T-1}\hat{u}_{t+1}^2\hat{z}_{t2}^{si}\hat{z}_{t2}^{si\prime}\right)\left(\frac{1}{T}\sum_{t=1}^{T-1}\hat{z}_{t3}^{si}\hat{z}_{t3}^{si\prime}\right)^{-1}, \\ \hat{\Sigma}_{\tau_3^{si}} &= \left(\frac{1}{T}\sum_{t=1}^{T-1}\hat{z}_{t3}^{si}\hat{z}_{t3}^{si\prime}\right)^{-1}\left(\frac{1}{T}\sum_{t=1}^{T-1}\hat{\epsilon}_{t+1}^2\hat{z}_{t3}^{si}\hat{z}_{t3}^{si\prime}\right)\left(\frac{1}{T}\sum_{t=1}^{T-1}\hat{z}_{t3}^{si}\hat{z}_{t3}^{si\prime}\right)^{-1}, \end{aligned}$$

with $\hat{\varepsilon}_{t+1} = X_{t+1} - \hat{z}_{1t}^{si\prime}\hat{\tau}_1^{si}$, $\hat{u}_{t+1} = \hat{f}_{t+1} - \hat{z}_{t2}^{si}\hat{\tau}_2^{si}$, $\hat{\epsilon}_{t+1} = X_{t+1} - \hat{z}_{t3}^{si\prime}\hat{\tau}_3^{si}$ the least squares residuals, where

$$\begin{aligned} \hat{z}_{t1}^{sii} &= (1, X_t, \ldots, X_{t+1-p}, Y_t, \ldots, Y_{t+1-q}, \hat{f}_t, \ldots, \hat{f}_{t+1-k})', \\ \hat{z}_{t2}^{si} &= (1, X_t, \ldots, X_{t+1-\dot{p}}, Y_t, \ldots, Y_{t+1-\dot{q}}, \hat{f}_t, \ldots, \hat{f}_{t+1-\dot{h}})', \\ \hat{z}_{t3}^{si} &= (1, X_t, \ldots, X_{t+1-\ddot{p}}, Y_t, \ldots, Y_{t+1-\ddot{q}}, \hat{f}_t, \ldots, \hat{f}_{t+1-\ddot{h}})'. \end{aligned}$$

The following theorem shows that the test statistics $W_T^{SI,\lambda}$, $W_T^{SI,\rho}$, and $W_T^{SI,\varsigma}$ are asymptotically distributed as chi-squared distributions with q, $\dot{q}$, and $\ddot{h}$ degrees of freedom, respectively. The proof of Theorem 3 is omitted since it is similar to those of Theorems 1 and 2.

Theorem 3 *: Let Assumption **A** holds. Under the null hypotheses (22), (26), and (27), respectively, if $\sqrt{T}/N \to 0$, then we have*

$$W_T^{SI,\alpha} \to_d \chi^2_q, \;\; W_T^{SI,\psi} \to_d \chi^2_{\dot{q}}, \;\; \text{and} \;\; W_T^{SI,\varsigma} \to_d \chi^2_{\ddot{h}},$$

where the test statistics $W_T^{SI,\alpha}$, $W_T^{SI,\psi}$ and $W_T^{SI,\varsigma}$ are defined in (35), respectively.

Concerning the type II spurious causality, conditions (ii), (iii)-(a) and (iii)-(b) can be tested using the following three Wald test statistics:

$$\begin{aligned}
W_T^{SII,\lambda} &= \left(\sqrt{T} R^{sii,\lambda} \hat{\tau}_1^{sii}\right) \left(R^{sii,\lambda} \hat{\Sigma}_{\hat{\tau}_1^{sii}} R^{sii,\lambda\prime}\right)^{-1} \left(\sqrt{T} R^{sii,\lambda} \hat{\tau}_1^{sii}\right)', \\
W_T^{SII,\rho} &= \left(\sqrt{T} R^{sii,\rho} \hat{\tau}_2^{sii}\right) \left(R^{sii,\rho} \hat{\Sigma}_{\hat{\tau}_2^{sii}} R^{sii,\rho\prime}\right)^{-1} \left(\sqrt{T} R^{sii,\rho} \hat{\tau}_2^{sii}\right)', \\
W_T^{SII,\varsigma} &= \left(\sqrt{T} R^{sii,\varsigma} \hat{\tau}_3^{sii}\right) \left(R^{sii,\varsigma} \hat{\Sigma}_{\hat{\tau}_3^{sii}} R^{sii,\varsigma\prime}\right)^{-1} \left(\sqrt{T} R^{sii,\varsigma} \hat{\tau}_3^{sii}\right)',
\end{aligned} \tag{36}$$

respectively, where $\hat{\tau}_1^{sii} = (\hat{\eta}, \hat{\gamma}', \hat{\lambda}', \hat{\theta}')'$, $\hat{\tau}_2^{sii} = (\hat{\nu}, \hat{\kappa}', \hat{\psi}', \hat{\rho}')'$, $\hat{\tau}_3^{sii} = (\hat{\varpi}, \hat{\xi}', \hat{\delta}', \hat{\varsigma}')'$, the selection matrices $R^{sii,\lambda} = (0_{\bar{q},1+\bar{p}}, I_{\bar{q}}, 0_{\bar{q},\bar{h}})$, $R^{sii,\rho} = (0_{\dot{h},1+\dot{p}+\dot{q}}, I_{\dot{h}})$, $R^{sii,\varsigma} = (0_{\ddot{h},1+\ddot{p}+\ddot{q}}, I_{\ddot{h}})$, and

$$\begin{aligned}
\hat{\Sigma}_{\tau_1^{sii}} &= \left(\tfrac{1}{T}\textstyle\sum_{t=1}^{T-1} \hat{z}_{t1}^{sii} \hat{z}_{t1}^{sii\prime}\right)^{-1} \left(\tfrac{1}{T}\textstyle\sum_{t=1}^{T-1} \hat{e}_{t+1}^2 \hat{z}_{t1}^{sii} \hat{z}_{t1}^{sii\prime}\right) \left(\tfrac{1}{T}\textstyle\sum_{t=1}^{T-1} \hat{z}_{t1}^{sii} \hat{z}_{t1}^{sii\prime}\right)^{-1}, \\
\hat{\Sigma}_{\tau_2^{sii}} &= \left(\tfrac{1}{T}\textstyle\sum_{t=1}^{T-1} \hat{z}_{t2}^{sii} \hat{z}_{t2}^{sii\prime}\right)^{-1} \left(\tfrac{1}{T}\textstyle\sum_{t=1}^{T-1} \hat{u}_{t+1}^2 \hat{z}_{t2}^{sii} \hat{z}_{t2}^{sii\prime}\right) \left(\tfrac{1}{T}\textstyle\sum_{t=1}^{T-1} \hat{z}_{t2}^{sii} \hat{z}_{t2}^{sii\prime}\right)^{-1}, \\
\hat{\Sigma}_{\tau_3^{sii}} &= \left(\tfrac{1}{T}\textstyle\sum_{t=1}^{T-1} \hat{z}_{t3}^{sii} \hat{z}_{t3}^{sii\prime}\right)^{-1} \left(\tfrac{1}{T}\textstyle\sum_{t=1}^{T-1} \hat{\epsilon}_{t+1}^2 \hat{z}_{t3}^{sii} \hat{z}_{t3}^{sii\prime}\right) \left(\tfrac{1}{T}\textstyle\sum_{t=1}^{T-1} \hat{z}_{t3}^{sii} \hat{z}_{t3}^{sii\prime}\right)^{-1},
\end{aligned}$$

with $\hat{e}_{t+1} = X_{t+1} - \hat{z}_{t1}^{sii\prime} \hat{\tau}_1^{sii}$, $\hat{u}_{t+1} = Y_{t+1} - \hat{z}_{t2}^{sii} \hat{\tau}_2^{sii}$, $\hat{\epsilon}_{t+1} = X_{t+1} - \hat{z}_{t3}^{sii\prime} \hat{\tau}_3^{sii}$, and

$$\begin{aligned}
\hat{z}_{t1}^{sii} &= (1, X_t, \ldots, X_{t+1-\bar{p}}, Y_t, \ldots, Y_{t+1-\bar{q}}, \hat{f}_t, \ldots, \hat{f}_{t+1-\bar{h}})', \\
\hat{z}_{t2}^{sii} &= (1, X_t, \ldots, X_{t+1-\dot{p}}, Y_t, \ldots, Y_{t+1-\dot{q}}, \hat{f}_t, \ldots, \hat{f}_{t+1-\dot{h}})', \\
\hat{z}_{t3}^{sii} &= (1, X_t, \ldots, X_{t+1-\ddot{p}}, Y_t, \ldots, Y_{t+1-\ddot{q}}, \hat{f}_t, \ldots, \hat{f}_{t+1-\ddot{h}})'.
\end{aligned}$$

The following theorem demonstrates that the tests $W_T^{SII,\lambda}$, $W_T^{SII,\rho}$, and $W_T^{SII,\varsigma}$ are asymptotically distributed as chi-squared distributions with $\bar{q}$, $\dot{h}$, and $\ddot{h}$ degrees of freedom, respectively. The proof of Theorem 4 follows naturally in a similar way as the one of Theorem 1, and therefore is omitted.

Theorem 4 *: Let Assumption **A** hold. Under the null hypotheses (28) and (29), respectively, if $\sqrt{T}/N \to 0$, then we have*

$$W_T^{SII,\lambda} \to_d \chi^2_{\bar{q}}, \;\; W_T^{SII,\rho} \to_d \chi^2_{\dot{h}}, \;\; \text{and} \;\; W_T^{SII,\varsigma} \to_d \chi^2_{\ddot{h}},$$

where the test statistics $W_T^{SII,\lambda}$, $W_T^{SII,\rho}$ and $W_T^{SII,\varsigma}$ are defined in (36), respectively.

As in Section 5.1, the results in Theorems 3 and 4 show that for $\sqrt{T}/N \to 0$, having estimated factors as regressand or regressors does not affect the root-T consistency of the least squares estimates of the coefficients used to test type I and type II spurious causalities. These results work under the general case of heteroskedasticity, and they are still valid under homoskedasticity.

6 Identification of the auxiliary variables

Unfortunately, the tests developed in the previous sections only detect the presence of indirect/spurious causality and cannot provide information about the nature of the auxiliary variables that transmit/induce these effects. In this section, we suggest a simple statistical procedure to identify the auxiliary variables that are responsible for the transmission/induction of indirect/spurious causality.

The literature on the interpretation of factors extracted using factor analysis suggests to use marginal regressions where each factor is regressed on each of the variables of the big data, see Ludvigson and Ng (2009) and the references therein. Thereafter, it uses the coefficient of determination R^2 to order the variables according to their importance in terms of explaining each factor. Thus, the variable(s) that produce(s) high R^2 are used to interpret the factor. Following this literature, we propose to identify the auxiliary variable(s) in the following way:

1. Test the conditions of indirect/spurious causality using as an auxiliary variable $f_t = w_{t,j}$, for $j = 1, ..., N$, where $w_{t,j}$ is one of the variables of the big data W defined in Section 4.1;

2. Eliminate all $w_{t,j}$, for $j = 1, ..., N$, which do not satisfy the conditions of indirect/spurious causality. We denote the subset of W with all variables satisfying the conditions of indirect/spurious causality by $W^{sub} = \left\{ \left(w_t^{(1)}, ..., w_t^{(\bar{N})} \right), \text{ for } t = 1, ..., T \text{ and where } \bar{N} \leq N \right\}$. W^{sub} can be viewed as a subset of auxiliary variables that transmit/induce indirect/spurious causality. If this subset is sufficiently large or because of possible interaction between the auxiliary variables, we can consider the next additional step;

3. Use the factor analysis where $\bar{k}$ common factors f_t^{sub} are associated with the $\bar{N}$-dimensional vector w_t^{sub} according to $w_t^{sub} = \Lambda^{sub} f_t^{sub} + \varepsilon_t$, with $w_t^{sub} = (w_t^{(1)}, ..., w_t^{(\bar{N})})'$ for $t = 1, ..., T$ and Λ^{sub} an $\left(\bar{N} \times \bar{k}\right)$-dimensional matrix of factor loadings and ε_t's vectors of idiosyncratic shocks that could be cross-sectionally and temporally dependent. We can then use the following marginal regressions, as in Ludvigson and Ng (2009), to interpret f_t^{sub}. Hereafter, we focus on one factor (f_t^{sub}), say the one that explains most of the variation in the data, as it represents all the variables in the big data W, since it is a linear combination of all these variables. Formally, we run the marginal regressions $f_t^{sub} = \alpha_j^{sub} + \beta_j^{sub} w_t^{(j)} + u_t^{(j)}$ for each $j = 1, ..., \bar{N}$ and obtain the corresponding R^2s. Thus, the auxiliary factor f_t^{sub} can be interpreted in terms of the variables in W^{sub} that generate high R^2s.

7 Monte Carlo simulations

We assess the sizes and powers of the tests stated in Theorems 1-3 under a variety of data generating processes (DGPs), different sample sizes and numbers of auxiliary variables.

7.1 Indirect causality

We first describe DGPs that we use in the simulations to assess the performance of the tests in Theorems 1 and 2. These DGPs are constructed based on the Example 1 of Section 3 and such that our assumptions are satisfied. Initially, we consider a set of DGPs in which indirect causality is transmitted through one auxiliary variable Z_1, among a total of N variables $\{Z_1, ..., Z_N\}$ with N large that represent the whole economy. In particular, we consider the following processes:

$$\begin{aligned} X_t &= \mu_X^{(1)} + \phi_X^{(1)} X_{t-1} + \phi_Y^{(1)} Y_{t-1} + \varepsilon_{X_t}^{(1)} \text{ with } \varepsilon_{X_t}^{(1)} \sim N(0,1), &(37)\\ X_t &= \mu_X^{(2)} + \phi_X^{(2)} X_{t-1} + \phi_Y^{(2)} Y_{t-2} + \phi_Z^{(2)} Z_{t-1,1} + \varepsilon_{X_t}^{(2)} \text{ with } \varepsilon_{X_t}^{(2)} \sim N(0,1), &(38)\\ Z_{t,1} &= \mu_Z - \phi_Y^{(2)}/\phi_Z^{(2)} Y_{t-1} + \phi_Z Z_{t-1,1} + \varepsilon_{Z_1,t} \text{ with } \varepsilon_{Z_1,t} \sim N(0,1) \text{ and } \phi_Z^{(2)} \neq 0, &(39)\\ Z_{t,i} &= \varepsilon_{Z_i,t} \sim N(0,1) \text{ for } i = 2, ..., N, \text{ with } \varepsilon_{Z_i,t} \text{ mutually independent,} \end{aligned}$$

where the error terms $\varepsilon_{X_t}^{(1)}$, $\varepsilon_{X_t}^{(2)}$, and $\varepsilon_{Z_i,t}$ for $i = 1, ..., N$, are assumed to be mutually independent and X has to satisfy both equations (37) and (38). Numerical values for the coefficients $\mu_X^{(1)}$, $\mu_X^{(2)}$, μ_Z, $\phi_X^{(1)}$, $\phi_X^{(2)}$, $\phi_Y^{(1)}$, $\phi_Y^{(2)}$, $\phi_Z^{(2)}$, ϕ_Z will be specified later. The functional forms of DGPs are selected such that there is an indirect causality from Y to X. According to Definition (1), indirect causality occurs if: (i) Y Granger causes X with respect to the information set $I_X(t)$; (ii) Y does not Granger cause X with respect to the information set $I(t) - I_Y(t)$; and (iii) Y Granger causes Z and Z Granger causes X with respect to the information sets $I(t) - I_Y(t)$ and $I(t) - I_Z(t)$, respectively. Thus, condition (i) will be satisfied if we choose the coefficient $\phi_Y^{(1)}$ to be different from zero. Furthermore, if we assume that $\phi_Z^{(2)} \neq 0$ and $\phi_Y^{(2)} \neq 0$, then Z Granger causes X and Y Granger causes Z, respectively, hence condition (iii) is satisfied. What about condition (ii)? Combining equations (38) and (39) leads to

$$X_t = \left(\mu_X^{(2)} + \phi_Z^{(2)} \mu_Z\right) + \phi_X^{(2)} X_{t-1} + \phi_Z^{(2)} \phi_Z Z_{t-2,1} + \phi_Z^{(2)} \varepsilon_{Z_1,t-1} + \varepsilon_{X_t}^{(1)},$$

which indicates that Y does not cause X in the presence of Z, hence condition (ii) is also satisfied.

Depending on whether or not $\phi_Y^{(1)}$, $\phi_Z^{(2)}$ and $\phi_Y^{(2)}$ are taken to be equal to zero, the following steps can be performed to simulate a sample of T observations on X, Y and Z under the absence/presence of indirect causality from Y to X transmitted through the auxiliary variable Z_1:

(1) Choose the initial values X_2, Z_2 and Y_1, and generate X_3 using (38): $X_3 = \mu_X^{(2)} + \phi_X^{(2)} X_2 + \phi_Y^{(2)} Y_1 + \phi_Z^{(2)} Z_2 + \varepsilon_{X_3}^{(1)}$, for $\varepsilon_{X_3}^{(2)} \sim N(0,1)$;

(2) Generate Y_2 using (37): $Y_2 = (X_3 - \mu_X^{(1)} - \phi_X^{(1)} X_2 - \varepsilon_{X_3}^{(1)})/\phi_Y^{(1)}$ for $\varepsilon_{X_3}^{(1)} \sim N(0,1)$ and $\varepsilon_{X_3}^{(1)} \perp \varepsilon_{X_3}^{(2)}$;

(3) Generate $Z_{3,1}$ using (39): $Z_{3,1} = \mu_Z - \phi_Y^{(2)}/\phi_Z^{(2)} Y_2 + \phi_Z Z_2 + \varepsilon_{Z_1,3}$ for $\varepsilon_{Z_1,3} \sim N(0,1)$ and $\varepsilon_{Z_1,3} \perp (\varepsilon_{X_3}^{(1)}, \varepsilon_{X_3}^{(2)})$;

(4) Generate $\{Z_{t,i}\}_{i=2}^{N}$ with $Z_{t,i}$ mutually independent, using $Z_{t,i} = \varepsilon_{Z_i,t} \sim N(0,1)$;

(5) Repeat steps (1)-(4) $T + 500$ times and discard the first 500 observations to eliminate the effect of initial values.

To examine the performance of tests in Theorem 1, Table 1 of the companion Appendix summarizes the DGPs, with a direct causality from Y to X, that we use in our simulations and provides four different sets of parameters that we choose to indicate various degrees of causality from Y to X and different degrees of serial dependence in Z. It is straightforward to notice that DGP1 and DGP3 are exhibiting a relatively weaker extent of direct causality from Y to X compared to DGP2 and DGP4, in terms of the coefficients in front of Y_{t-1}; i.e. 0.3 versus 0.7. For the direct causality set-up in Table 1, the auxiliary variables Zs give no extra useful information for predicting X and we have used two different types of Zs, which can be either i.i.d. or AR(1) processes.

Furthermore, we consider the additional DGPs in Table 2 of the companion Appendix, which correspond to equations (37)-(39) with four different sets of parameters that represent different scenarios of indirect causality. Since Z_1 is present in DGP5 to DGP8, different parameter values indicate different degrees of indirect causality from Y to X transmitted through Z_1. For instance, it seems at first sight that there is a high causality from Y to X in DGP6 given by $\phi_Y^{(1)} = 0.7$. However, this causality is not direct and it disappears once controlling the effect of Z, hence following into the context of indirect causality.

In the simulations, three different sample sizes are studied: $T = 100$, 200, and 400. In addition, N is chosen to be varying according to the number of time periods T and satisfies $\sqrt{T}/N \to 0$. In particular, for each DGP in Tables 1 and 2, three different values of N are considered to examine the effect of data richness on the performance of the tests. The nominal level 5% is used and results for other levels are available upon request. Finally, all the results are based on 2000 replications.

Simulation results using DGP1 to DGP8 are reported in Tables 3-5 of the companion Appendix. Observe that for testing condition (ii) of Definition (1), DGP1 to DGP4 are used to investigate the power performance of the tests, and DGP5 to DGP8 are used to examine their size properties [see Table 3 for details]. However, it is important to notice that the roles of null and alternative hypotheses are reversed when testing condition (iii) of Definition (1) based on the regression equations (17) and (18) [see Tables 4 and 5 for details]. The results show that the empirical sizes are accurate for different DGPs with various sample sizes T and numbers of auxiliary variables N. Furthermore, for our simulation settings, it seems that the test of non-causality from Y to Z is substantially more powerful than the test of non-causality from Z to X. Finally, the number N of auxiliary variables Zs does not seem to affect greatly the performance of the tests given various sample sizes T.

To further illustrate the performance of the proposed tests, we consider the interesting case where many auxiliary variables [10 variables in our simulation] are transmitting the indirect causality from Y to X. In particular, we shall consider the following processes:

$$
\begin{aligned}
X_t &= \mu_X^{(1)} + \phi_X^{(1)} X_{t-1} + \phi_Y^{(1)} Y_{t-1} + \varepsilon_{X_t}^{(1)} \text{ with } \varepsilon_{X_t}^{(1)} \sim N(0,1), \qquad (40) \\
X_t &= \mu_X^{(2)} + \phi_X^{(2)} X_{t-1} + \phi_Y^{(2)} Y_{t-2} + \sum_{j=1}^{10} \phi_{Z_j}^{(2)} Z_{t-1,j} + \varepsilon_{X_t}^{(2)} \text{ with } \varepsilon_{X_t}^{(2)} \sim N(0,1), \qquad (41) \\
Z_{t,j} &= \mu_{Z_j} - (\phi_Y^{(2)}/\phi_Z^{(2)}) Y_{t-1} + \phi_{Z_j} Z_{t-1,j} + \varepsilon_{Z_j,t} \text{ with } \{\varepsilon_{Z_j,t}\}_{j=1}^{10} \sim N(0,1) \text{ and } \phi_Z^{(2)} \neq 0, \qquad (42) \\
Z_{t,i} &= \varepsilon_{Z_i,t} \sim N(0,1) \text{ for } i = 11, ..., N, \text{ with } \varepsilon_{Z_i,t} \text{ mutually independent,}
\end{aligned}
$$

where $\phi_Z^{(2)} := \sum_{j=1}^{10} \phi_{Z_j}^{(2)}$. The latter condition guarantees that Y does not cause X in the presence of Z, which satisfies condition (ii) of Definition (1). We now consider the five DGPs in Table 6 of the companion Appendix and follow the steps described above to simulate the data on X, Y, and Z.

The DGPs in Table 6 are different from those in Tables 1 and 2. In particular, the values of parameters in the former are much smaller, which indicates that the causal links are significantly weaker when we consider ten auxiliary variables. The use of smaller values is also to ensure that the processes under consideration are stationary. Due to the weak degree of causality, the power of tests of indirect causality will be low when we use ten auxiliary variables instead of one, though the power still increases as sample size increases.

Tables 7-9 of the companion Appendix report the empirical sizes and powers of the tests of conditions (ii) and (iii) of an indirect causality from Y to X transmitted by 10 auxiliary variables. Before discussing the results, recall that DGPs 11 to 13 are used to assess the empirical size of test of condition (ii) [Theorem 1], whereas DGPs 9 and 10 are used to assess the empirical size of test of condition (iii) [Theorem 2]. The results show that the proposed tests control very well the size whatever the sample size T, the number of Zs, and the DGP under consideration. Regarding the power, on one hand we find that the test of condition (ii) in Theorem 1 has low power, but as expected it improves with the sample size T. On the other hand, the power of test of condition (iii) in Theorem

2 is high and reaches one even when the sample size is small. The difference in the power performance for testing different conditions of Definition (1) is mainly due to the particular design of our DGPs specified either in one auxiliary variable case or ten auxiliary variables case. For example, in the latter case we specified, Tables 8 and 9 show the results for checking condition (iii) of Definition (1), in which Table 8 has significantly higher power when testing the first regression of condition (iii). This is expected, as the process for $\{Z_{t,j}\}_{j=1}^{10}$ has large coefficients for Y_{t-1}. On the other hand, in Table 9 when testing the second regression of condition (iii), the process for X_t (expressed as an equation of $Z_{t-1,j}$) has ten very small coefficients of $Z_{t-1,j}$, leading to the less powerful results. Note that as the magnitude of coefficients of $Z_{t-1,j}$ increases, the power increases as expected, see DGPs 11, 12 and 13 in Table 9. By designing other types of DGPs, we may have different power performance. However, our theory predicts that the testing procedure should work for other general situations. Finally, we find that both empirical size and power are quite stable when N changes.

7.2 Spurious causality

We describe the DGPs that we use to assess the performance of tests of spurious causality of type I in Theorem 3. These DGPs are constructed based on the Example 2 of Section 3 and such that our assumptions are satisfied. We first consider a set of DGPs in which the spurious causality of type I is induced by one and only one auxiliary variable Z_1, among a total of N variables $\{Z_1, ..., Z_N\}$ indicating the richness of the data environment. In particular, we consider the following processes:

$$\begin{aligned}
X_t &= \mu_X^{(1)} + \phi_X^{(1)} X_{t-1} + \varepsilon_{X_t}^{(1)} \text{ with } \varepsilon_{X_t}^{(1)} \sim N(0,1) && (43)\\
X_t &= \mu_X^{(2)} + \phi_X^{(2)} X_{t-1} + \phi_Z^{(2)} Z_{t-1,1} + \varepsilon_{X_t}^{(2)} \text{ with } \varepsilon_{X_t}^{(2)} \sim N(0,1) && (44)\\
Z_{t,1} &= \mu_Z + \phi_Z^{(3)} Z_{t-1,1} + \phi_Y^{(3)} Y_{t-1} + \phi_X^{(3)} X_{t-1} + \varepsilon_{Z_1,t} \text{ with } \varepsilon_{Z_1,t} \sim N(0,1), && (45)\\
Z_{t,1} &= \varepsilon_{Z_i,t} \sim N(0,1) \text{ for } i = 2, ..., N, \text{ with } \varepsilon_{Z_i,t} \text{ mutually independent,}
\end{aligned}$$

where the error terms $\varepsilon_{X_t}^{(1)}$, $\varepsilon_{X_t}^{(2)}$, and $\varepsilon_{Z_i,t}$, for $i = 1, ..., N$, are assumed to be mutually independent and X has to satisfy both (43) and (44). Numerical values for the coefficients $\mu_X^{(1)}$, $\mu_X^{(2)}$, μ_Z, $\phi_X^{(1)}$, $\phi_X^{(2)}$, $\phi_X^{(3)}$, $\phi_Z^{(2)}$, $\phi_Z^{(3)}$, $\phi_Y^{(3)}$ will be specified later. The functional forms of DGPs of X and Z are selected such that there is a spurious causality of type I from Y to X. According to Definition (2), spurious causality of type I occurs if: (i) Y Granger causes X with respect to the information set $I(t) - I_Y(t)$; (ii) Y does not Granger cause X with respect to the information set $I_X(t)$; and (iii) Y Granger causes Z and Z Granger causes X with respect to the information sets $I(t) - I_Y(t)$ and $I(t) - I_Z(t)$, respectively. Thus, conditions (i) and (iii) will be satisfied if we choose the coefficients $\phi_Z^{(2)}$ and $\phi_Y^{(3)}$ in (44) and (45) to be different from zero. Furthermore, condition (ii) is also satisfied according to (43). To see why condition (i) is satisfied, note that by combining (44) and (45) we obtain

$$X_t = \left(\mu_X^{(2)} + \phi_Z^{(2)}\mu_Z\right) + \phi_X^{(2)}X_{t-1} + \phi_Z^{(2)}\phi_X^{(3)}X_{t-2} + \phi_Z^{(2)}\phi_Y^{(3)}Y_{t-2} + \phi_Z^{(2)}\phi_Z^{(3)}Z_{t-2,1} + \phi_Z^{(2)}\varepsilon_{Z_1,t-1} + \varepsilon_{X_t}^{(2)},$$

where Y does cause X in the presence of Z.

Depending on whether or not $\phi_Z^{(2)}$ and $\phi_Y^{(3)}$ are taken to be equal to zero, the following steps can be performed to simulate a sample of T observations on X, Y and Z under the absence/presence of spurious causality of type I from Y to X induced by the auxiliary variable Z_1:

(1) Choose the initial value X_1 and generate X_2 and X_3 using (43):

$$X_2 = \mu_X^{(1)} + \phi_X^{(1)}X_1 + \varepsilon_{X_2}^{(1)}, \text{ for } \varepsilon_{X_2}^{(1)} \sim N(0,1);$$
$$X_3 = \mu_X^{(1)} + \phi_X^{(1)}X_2 + \varepsilon_{X_3}^{(1)}, \text{ for } \varepsilon_{X_3}^{(1)} \sim N(0,1) \text{ and } \varepsilon_{X_2}^{(1)} \perp \varepsilon_{X_3}^{(1)};$$

(2) Generate $Z_{1,1}$ and $Z_{2,1}$ using (44):

$$Z_{1,1} = \left(X_2 - \mu_X^{(2)} - \phi_X^{(2)}X_1 - \varepsilon_{X_2}^{(2)}\right)/\phi_Z^{(2)}, \text{ for } \phi_Z^{(2)} \neq 0,\ \varepsilon_{X_2}^{(2)} \sim N(0,1), \text{ and } \varepsilon_{X_2}^{(2)} \perp \left(\varepsilon_{X_2}^{(1)}, \varepsilon_{X_3}^{(1)}\right)$$
$$Z_{2,1} = X_3 - \mu_X^{(2)} - \phi_X^{(2)}X_2 - \varepsilon_{X_3}^{(2)}/\phi_Z^{(2)}, \text{ for } \phi_Z^{(2)} \neq 0,\ \varepsilon_{X_3}^{(2)} \sim N(0,1), \text{ and } \varepsilon_{X_3}^{(2)} \perp \left(\varepsilon_{X,2}^{(2)}, \varepsilon_{X,2}^{(1)}, \varepsilon_{X,3}^{(1)}\right);$$

(3) Generate Y_1 using (45):

$$Y_1 = \left(Z_{2,1} - \mu_Z - \phi_Z^{(3)}Z_{1,1} - \phi_X^{(3)}X_1 - \varepsilon_{Z_1,2}\right)/\phi_Y^{(3)}, \text{ for } \varepsilon_{Z_1,2} \sim N(0,1) \text{ and } \varepsilon_{Z_1,2} \perp \left(\varepsilon_{X_2}^{(2)}, \varepsilon_{X_2}^{(1)}, \varepsilon_{X_3}^{(1)}\right);$$

(4) Generate $\{Z_{t,i}\}_{i=2}^{N}$ with $Z_{t,i}$ mutually independent, using $Z_{t,i} = \varepsilon_{Z_i,t} \sim N(0,1)$;

(5) Repeat steps (1)-(4) $T+500$ times and discard the first 500 observations to eliminate the effects of initial values.

To examine the size of tests in Theorem 3, Table 10 of the companion Appendix summarizes the DGPs, when there is no causality from Y to X, that we use in our simulations. Regarding the assessment of the power, we consider the DGPs in Table 11 of the companion Appendix that correspond to equations (43), (44), and (45) with four different sets of parameters that represent different scenarios of spurious causality of type I. As in Section 7.1, three sample sizes are studied; $T = 100$, 200, 400, and N is chosen to be varying according to the sample sizes T. The nominal level 5% is studied and results for other levels are omitted. All the results are based on 2000 replications.

Tables 12 to 14 of the companion Appendix report the empirical size and power of tests of conditions of spurious causality of Type I under the DGPs in tables 10 and 11. On the one hand, the results, using DGPs 14 to 17, show that the proposed tests control the size reasonably whatever the sample size T and the number of auxiliary variables N. The size control is achieved by all the tests in Theorem 3 and under all the DGPs, except DGP 16 for which the empirical size is slightly higher than the nominal level of 5%. On the other hand, the empirical power of the tests reaches one for all DGPs [DGP18 to DGP21], even when the sample size is small and whatever the number of auxiliary variables. Finally, the case of multiple auxiliary variable Zs is omitted for the sake of brevity. Simulations for spurious causality of type II are also omitted.

8 Empirical application

We use the tests proposed in the above sections to test for the presence of an indirect causality from credit/money to real activity. Studying the interaction between real activity (income) and monetary policy measures (money) and credit is of great importance to economists, because of its role in stabilizing the economy and for economic welfare; see e.g. Friedman (1981), Friedman and Kuttner (1992), and Balke (2000), Uhlig (2005) and references therein. In this section we use our methodology to re-examine this old relationship and identify the channels behind its existence.

Fackler (1985) was among the first to examine the channels behind the impact of credit/money on real activity. As pointed out by this paper [see Section 3], in studying the relationship between money and income, empirical evidence suggests that important information may be lost by ignoring some variables such as the one that comes from credit market. He argued that empirical results on examining money-income causal relationship might differ depending on the information set one has at hands. His analysis shows that the results based on bivariate causality analysis are misleading and often overturned when one extends the information set and includes other key variables such as the ones related to the credit. In particular, he found that interest rates play the role of an auxiliary variable that transmits the causality between financial and real sectors. In other words, money/credit does not directly influence real output; but it plays at most an indirect role in income determination. To obtain his results, Fackler (1985) applied an ad hoc approach in which the auxiliary variables were predetermined or specified at the beginning of the analysis and not selected by any statistical method. Furthermore, his tests were run in the presence of only few variables, thus this excluded the hundred of economic variables that might play a role in income determination.

Our objective is to use the tests proposed in Section 5.1 to re-examine the existence of an indirect causality from money/credit to income using a recent dataset that contains more than 130 economic variables. In particular, we would like to confirm whether or not there is an indirect causality from money/credit to income. Thereafter, if this indirect causality exists, we would like to use the algorithm discussed in Section 6 to identify the auxiliary variable(s) that transmit this causality and compare them with those used in Fackler (1985). In this application income is measured by Industrial Production Index (IPI), money is measured by M1 Money Stock [hereafter M1SL using the FRED], and we consider two measures of credit: Commercial and Industrial Loans [hereafter BUSLOANS] and Securities in Bank Credit at All Commercial Banks [hereafter INVEST].

8.1 Data

We consider a big data set that consists of monthly observations on 135 economic variables from Federal Reserve Bank of St. Louis (FRED). The sample runs from January 1959 to May 2016 for a total of 689 observations. All the variables are reported in Tables 15-20 of the companion Appendix. In particular, we consider 8 groups of variables: (1) Output and income with 17 variables; (2) Labor market with 32 variables; (3) Housing with 10 variables; (4) Consumption, orders, and inventories with 14 variables; (5) Money and credit with 14 variables; (6) Interest and exchange rates with 22 variables; (7) Prices with 21 variables; and (8) Stock market with 5 variables. This big data mimic the coverage of datasets already used in the literature and it is updated in real-time through the FRED database. A detailed description of the dataset can be found in McCracken and Ng (2015).

8.2 Results

First, using Akaike information criterion, our results show that regressions (7), (15), (17), and (18) with 3 or 4 lagged terms suffice to test the conditions of a possibly indirect causality from money/credit to income. Table 21 of the companion Appendix reports the p-values for testing the conditions in Definition (1). On one hand, we find that there is no indirect causality from money to income, as the first condition [money Granger causes income without the presence of other variables] is not satisfied. Consequently, this renders the subsequent testing procedure unnecessary, even though all the following conditions are satisfied. Hence, we conclude that money is not Granger indirectly causing income, which goes against the findings in Fackler (1985). In Table 21, we only include the results from one measure of money, i.e. M1SL, for illustration. In fact all other money measures fail to pass the indirect causality tests and demonstrate quantitatively similar results.

On the other hand, the measures of credit [BUSLOANS and INVEST] fit to our testing paradigm well. In particular, the p-value for the first condition [BUSLOANS Granger causing income without the presence of other variables] is equal to 0.0728. However, once the auxiliary variable f [extracted from 135 economic variables] is included, BUSLOANS does not Granger cause income any more with a high p-value of 0.5618. Lastly, for the third condition, BUSLOANS appears to Granger cause the auxiliary variable f and f furthermore Granger causes income with p-values of 0.0385 and 0.0005, respectively. This leads us to believe that Commercial and Industrial Loans serves as an indirect source of income. For the credit measure INVEST, the same argument applies and the four p-values again help us to conclude that INVEST Granger causes income, but only in an indirect way. These results are in line with the findings in Fackler (1985).

Now that we have found that there is an indirect causality from credit to income, we next use the procedure outlined in Section 6 to identify the auxiliary variables that transmit this causality. The results are summarized in Table 22 of the companion Appendix, where the first and second columns report the auxiliary variables that transmit the indirect causality from Commercial and Industrial Loans and Securities in Bank Credit at All Commercial Banks to income, respectively. On one hand, we see that 18 auxiliary variables are responsible for the transmission of indirect causality from BUSLOANS to income. These variables belong to five groups: (i) Labor market; (ii) Housing; (iii) Consumption, orders, and inventories; (iv) Interest and exchange rates; and (v) Stock market. The variables from the other groups are found to be silent. We also find that most of the auxiliary variables [11 over a total of 18] belong to the group on interest and exchange rates. These variables are: Effective Federal Funds Rate, 3-Month AA Financial Commercial Paper Rate, 1-Year Treasury Rate, Moody's Seasoned Baa Corporate Bond Yield, 3-Month Treasury C Minus FEDFUNDS, 6-Month Treasury C Minus FEDFUNDS, 1-Year Treasury C Minus FEDFUNDS, 5-Year Treasury C Minus FEDFUNDS, 10-Year Treasury C Minus FEDFUNDS, Moody's Aaa Corporate Bond Minus FEDFUNDS, and Moody's Baa Corporate Bond Minus FEDFUNDS. Thus, it seems that the short and long-term interest rates are the main auxiliary variables that transmit the indirect causality from BUSLOANS to income, which is in line with the findings in Fackler (1985). Fackler (1985) wrote: "*What is presumably relevant for income determination, and especially for the investment component of income, is the long-term interest rate.*"

Regarding the indirect causality from INVEST to income, column 2 of Table 22 shows that 14 auxiliary variables are responsible for the transmission of this causality. These variables belong to five groups: (i) Output and income; (ii) Housing; (iii) Consumption, ordered inventories; (iv) Interest and exchange rates; and (v) Stock market. The dominant groups with highest numbers of auxiliary variables are Housing and Interest and exchange rates. Thus, in addition to the short and long-term interest rates, the housing sector is essential for transmitting the causality from credit to income, which is different from the findings in Fackler (1985).

9 Conclusion

We introduced several statistical procedures for testing indirect and spurious causal effects. In practice, detecting indirect/spurious causality is a complicated task, since the pertinent auxiliary variables that transmit/induce the indirect/spurious causality are very often unknown. The availability of hundreds of economic variables makes this task even harder as it is generally infeasible to find the appropriate auxiliary variable(s) among all the available ones. We proposed several statistical procedures to test for the presence of an indirect/spurious causality using big data analysis. A diffusion index was included in the regression equation to represent all the variables that are available to practitioners. We derived the asymptotic distributions of the tests in the presence of an estimated index. Furthermore, we conducted a Monte Carlo simulation to evaluate the performance of the proposed statistical procedures. The results showed that our procedures have good power for detecting indirect/spurious causality. Finally, we provided an empirical application where hundreds of variables are used to study a possible indirect causality from money/credit to income.

References

[1] Bai, J. (2003). 'Inferential theory for factor models of large dimensions', *Econometrica*, Vol. 71, pp. 135–171.

[2] Bai, J., Ng, S. (2002). 'Determining the number of factors in approximate factor models', *Econometrica*, Vol. 70, pp. 191–221.

[3] Bai, J., Ng, S. (2005). 'Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions', *Working paper*, University of Michigan.

[4] Bai, J., Ng, S. (2006). 'Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions', *Econometrica*, Vol. 74, pp. 1133–1150.

[5] Balke, N.S. (2000). 'Credit and economic activity: credit regimes and nonlinear propagation of shocks', *The Review of Economics and Statistics*, Vol. 82(2), pp. 344–349.

[6] Bouezmarni, T., Rombouts, J., Taamouti, A. (2012). 'A nonparametric copula based test for conditional independence with applications to granger causality', *Journal of Business & Economic Statistics*, Vol. 30, pp. 275–287.

[7] Dufour, J.M., Renault, E. (1998). 'Short run and long run causality in time series: theory', *Econometrica*, pp. 1099–1125.

[8] Dufour, J.-M., Taamouti, A. (2010). 'Short and long run causality measures: theory and inference', *Journal of Econometrics,* Vol. 154, pp. 42–58.

[9] Eichler, M. (2007). 'Granger causality and path diagrams for multivariate time series', *Journal of Econometrics*, Vol. 137(2), pp. 334–353.

[10] Eichler, M. (2012). Causal inference in time series analysis. In: C. Berzuini, A.P. Dawid, L. Bernardinelli (eds), Causality: Statistical Perspectives and Applications, Wiley, Chichester.

[11] Fackler, J.S. (1985). 'An empirical analysis of the markets for goods, money, and credit', *Journal of Money, Credit and Banking*, Vol. 17, pp. 28–42.

[12] Florens, J.P., Mouchart, M. (1985a). 'A linear theory for noncausality', *Econometrica*, pp.157–175.

[13] Florens, J.P., Mouchart, M. (1985b). 'Conditioning in dynamic models', *Journal of Time Series Analysis*, Vol. 6, pp. 15–34.

[14] Friedman, B.M. (1981). 'The roles of money and credit in macroeconomic analysis', *Mimeographed.*

[15] Friedman, B.M., Kuttner, K. (1992). 'Money, income, prices, and interest rates', *American Economic Review,* pp. 472–492.

[16] Gonçalves, S., Perron, P. (2014). 'Bootstrapping factor-augmented regression models', *Journal of Econometrics*, Vol. 182, pp. 156–173.

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

[18] Hsiao, C. (1980). 'Autoregressive modeling and causal ordering of economic variables', Princeton University Working Paper.

[19] Hsiao, C. (1982). 'Autoregressive modeling and causal ordering of economic variables', *Journal of Economic Dynamics and Control,* Vol. 4, pp. 243–259.

[20] Ludvigson, S., Ng, S. (2009). 'Macro factors in bond risk premia', *Review of Financial Studies*, Vol. 22, pp. 5027–5067.

[21] Ludvigson, S., Ng, S. (2011). A factor analysis of bond risk premia. In: Ullah, A., Giles, D. (Eds.), Handbook of Empirical Economics and Finance. Chapman and Hall, pp. 313–372.

[22] Lütkepohl, H., (1982). 'Non-causality due to omitted variables', *Journal of Econometrics*, Vol. 19, pp. 367–378.

[23] Lütkepohl, H., (1993). 'Introduction to Multiple Time Series Analysis', second edn. Springer-Verlag, Berlin.

[24] McCracken, M.W., Ng, S. (2015). 'FRED-MD: A monthly database for macroeconomic research', Federal Reserve Bank of St. Louis Working Paper Series.

[25] Sims, C.A. (1980). 'Macroeconomics and reality', *Econometrica*, pp. 1–48.

[26] Song, X., Taamouti, A. (2018). 'Measuring nonlinear Granger causality in mean', *Journal of Business & Economic Statistics*, Vol. 36(2), pp. 321–333.

[27] Stock, J.H., Watson, M.W. (2002). 'Forecasting using principal components from a large number of predictors', *Journal of the American Statistical Association,* Vol. 97, pp. 1167–1179.

[28] Uhlig, H. (2005). 'What are the effects of monetary policy on output? Results from an agnostic identification procedure', *Journal of Monetary Economics*, Vol. 52(2), pp. 381–419.

[29] Wiener, N. (1956): 'The theory of prediction', *in E. F. Beckenback, ed., The Theory of Prediction*, McGraw-Hill, New York, chapter 8.

A better understanding of Granger causality analysis: A big data environment

Online Appendix

October 9, 2018

ABSTRACT

This online appendix contains the proofs of the theoretical results derived in the main paper. It also provides the parameter values of the data generating processes (DGPs) used in the simulation study in Section 7 of the main paper, the simulation results for the empirical size and power of the test procedures developed in the main paper, and the data and the empirical results discussed in Section 8 of the main paper.

1 Proofs

This appendix provides the proofs of Theorems 1 to 4 in the main text of the main paper. We first introduce some notations, which are adapted from Bai and Ng (2006). Let $\hat{V}$ be the $(k \times k)$ diagonal matrix consisting of the k largest eigenvalues of $WW'/(TN)$ and let $H = \hat{V}^{-1}(\hat{f}'f/T)(\Lambda'\Lambda/N)$ be the rotation matrix, due to the fact that $\hat{f}$ can only consistently estimate Hf, the space spanned by the true factors f. Let $\Phi_0 = \text{diag}(I_{1+\bar{p}+\bar{q}}, V^{-1}Q\Sigma_\Lambda)$ being block diagonal, where $V = \text{plim}\,\hat{V}$, $Q = \text{plim}\,\hat{f}'f/T$ and Σ_Λ is defined in Assumption **A** of the main paper.

Three auxiliary lemmas are first given below. The first one is due to Bai and Ng (2005).

Lemma 1: *Take $\hat{z}_t$ to be $\hat{z}_{tj}$, or $\hat{z}_{tj}^{sii}$, or $\hat{z}_{tj}^{si}$ for any $j = 1, 2, 3$, which are defined in Section* 5 of the main paper. *Let z_t be the corresponding infeasible regressors and $\bar{e}_{t+1}$ be any of the error terms in the corresponding regression. Let $\delta_{NT}^2 = \min[N, T]$. Then under Assumption **A**, we have:* **(i)** $\frac{1}{T}\sum_{t=1}^{T}||\hat{f}_t - Hf_t||^2 = O_p\left(\delta_{NT}^{-2}\right)$; **(ii)** $\frac{1}{T}\sum_{t=1}^{T}(\hat{f}_t - Hf_t)z_t' = O_p\left(\delta_{NT}^{-2}\right)$; **(iii)** $\frac{1}{T}\sum_{t=1}^{T}(\hat{f}_t - Hf_t)\hat{z}_t' = O_p\left(\delta_{NT}^{-2}\right)$; and **(iv)** $\frac{1}{T}\sum_{t=1}^{T}(\hat{f}_t - Hf_t)\bar{e}_{t+1}' = O_p\left(\delta_{NT}^{-2}\right)$.

Lemma 2: *Let $\hat{z}_t$ and z_t be the feasible and infeasible regressors defined in Lemma 1. Let $\delta_{NT}^2 = \min[N, T]$.*

*Then under Assumption **A**, we have:* **(i)** $\frac{1}{T}\sum_{t=1}^{T}(\hat{f}_{t+1} - Hf_{t+1})z_t' = O_p\left(\delta_{NT}^{-2}\right)$ *and* **(ii)** $\frac{1}{T}\sum_{t=1}^{T}(\hat{f}_{t+1} - Hf_{t+1})\hat{z}_t' = O_p\left(\delta_{NT}^{-2}\right)$.

Proof of Lemma 2: They are similar to the proofs of results (ii) and (iii) in Lemma 1.■

Lemma 3: *Consider the infeasible regression (15) or (18) of the main paper. Let τ be the parameter to be estimated and $\hat{\tau}$ its ordinary least squares estimate obtained from a regression of X_{t+1} on the vector of regressors $\hat{z}_t$, with $\hat{z}_t$ includes the intercept, lagged values of X_t, Y_t, and the estimated factors $\hat{f}_t$ and its lags. Suppose Assumption **A** hold. If $\sqrt{T}/N \to 0$, then*

$$\sqrt{T}(\hat{\tau} - \tau) \to_d N(0, \Sigma_\tau),$$

where $\Sigma_\tau = \Phi_0'^{-1}\Sigma_{zz}^{-1}\Sigma_{zz,e}\Sigma_{zz}^{-1}\Phi_0^{-1}$. Moreover, a consistent estimator of the variance covariance matrix Σ_τ is given by

$$\hat{\Sigma}_\tau = \left(\frac{1}{T}\sum_{t=1}^{T-1}\hat{z}_t\hat{z}_t'\right)^{-1}\left(\frac{1}{T}\sum_{t=1}^{T-1}\hat{e}_{t+1}^2\hat{z}_t\hat{z}_t'\right)\left(\frac{1}{T}\sum_{t=1}^{T-1}\hat{z}_t\hat{z}_t'\right)^{-1},$$

where $\hat{e}_{t+1} = X_{t+1} - \hat{z}_t'\hat{\tau}$ are the least squares residuals.

Proof of Lemma 3: For Lemma 3, we will only prove the case for regression (15) in the main paper as the proof for the limiting distribution of $\hat{\tau}$ in regression (18) of the main paper is identical and hence it is omitted. The proof is much similar to that of Theorem 1 in Bai and Ng (2006). Without loss of generality, assume $\bar{p} = \bar{q} = \bar{h} = 1$, and define $z_t = (1, X_t, Y_t, f_t)'$ and $\hat{z}_t = (1, X_t, Y_t, \hat{f}_t)'$ so that $\tau = (\eta, \gamma, \lambda, \theta H^{-1})'$ are the parameters from the infeasible regression when f_t is observed.

From the infeasible regression (15) in the main paper, adding and subtracting terms, we obtain

$$\begin{aligned} X_{t+1} &= \eta + \gamma X_t + \lambda Y_t + \theta f_t + e_{t+1} = \eta + \gamma X_t + \lambda Y_t + \theta H^{-1}\hat{f}_t + e_{t+1} + \theta H^{-1}(Hf_t - \hat{f}_t) \\ &= \hat{z}_t'\tau + e_{t+1} + \theta H^{-1}(Hf_t - \hat{f}_t). \end{aligned}$$

In matrix notation, $X = \hat{z}\tau + e + (fH' - \hat{f})H^{-1'}\theta$, where $X = (X_2, \ldots, X_T)'$, $\hat{z} = (\hat{z}_1, \ldots, \hat{z}_{T-1})'$, $e = (e_2, \ldots, e_T)'$, $f = (f_1, \ldots, f_{T-1})'$ and $\hat{f} = (\hat{f}_1, \ldots, \hat{f}_{T-1})'$. Therefore, the ordinary least squares estimator of τ is given by

$$\hat{\tau} = (\hat{z}'\hat{z})^{-1}\hat{z}'X = \tau + (\hat{z}'\hat{z})^{-1}\hat{z}'e + (\hat{z}'\hat{z})^{-1}\hat{z}'(fH' - \hat{f})H^{-1'}\theta.$$

Thus,

$$\sqrt{T}(\hat{\tau} - \tau) = (T^{-1}\hat{z}'\hat{z})^{-1}T^{-1/2}\hat{z}'e + (T^{-1}\hat{z}'\hat{z})^{-1}[T^{-1/2}\hat{z}'(fH' - \hat{f})]H^{-1'}\theta.$$

By the result (iii) of Lemma 1, the second term on the right-hand side of the above equation is $O_p(T^{1/2}/\min(N,T)) = o_p(1)$ if $\sqrt{T}/N \to 0$. Define $W_t = (1, X_t, Y_t)'$ so that $T^{-1/2}\hat{z}'e = T^{-1/2}(e'W, e'\hat{f})'$. Due to the fact that $T^{-1/2}\hat{f}'e = T^{-1/2}Hf'e + T^{-1/2}(\hat{f} - fH')'e$, we have

$$T^{-1/2}\hat{f}'e = T^{-1/2}Hf'e + o_p(1)$$

by the result (iv) of Lemma 1 and $\sqrt{T}/N \to 0$. Thus, we get that $T^{-1/2}\hat{f}'e = T^{-1/2}(e'W, e'fH')' + o_p(1) = T^{-1/2}\Phi z'e + o_p(1)$, with $\Phi = \text{diag}(I, H)$ a block diagonal matrix. Therefore,

$$\sqrt{T}(\hat{\tau} - \tau) = (T^{-1}\hat{z}'\hat{z})^{-1}T^{-1/2}\hat{z}'e + o_p(1) = (T^{-1}\hat{z}'\hat{z})^{-1}\Phi T^{-1/2}z'e + o_p(1).$$

Under standard assumptions, we have $T^{-1/2}\sum_{t=1}^{T-1} z_t e_{t+1} \to_d N(0, \Sigma_{zz,e})$ with $\Sigma_{zz,e} = \text{plim} T^{-1}\sum_{t=1}^{T-1} e_{t+1}^2 z_t z_t'$. Therefore, $\sqrt{T}(\hat{\tau} - \tau) \to_d N(0, \Sigma_\tau)$ with the asymptotic variance covariance matrix given by

$$\Sigma_\tau = \text{plim}\left(\frac{\hat{z}'\hat{z}}{T}\right)^{-1}\Phi\left(\frac{1}{T}\sum_{t=1}^{T-1} e_{t+1}^2 z_t z_t'\right)\Phi'\left(\frac{\hat{z}'\hat{z}}{T}\right)^{-1},$$

where $\Phi = diag(I, H)$ is a block diagonal matrix with the probability limit Φ_0. Following Bai and Ng (2006), $\Sigma_\tau = \Phi_0'^{-1}\Sigma_{zz}^{-1}\Sigma_{zz,e}\Sigma_{zz}^{-1}\Phi_0^{-1}$.

In addition, by Bai and Ng (2006), $\hat{\Sigma}_\tau$ is a consistent estimator for Σ_τ.■

Proof of Theorem 1: We focus on the case where $\bar{p} = \bar{q} = \bar{h} = 1$. Under heteroskedasticity, the corresponding Wald-statistic is defined by

$$W_T^{Ind,\lambda} = \left(\sqrt{T}R^{Ind,\lambda}\hat{\tau}\right)\left(R^{Ind,\lambda}\hat{\Sigma}_\tau R^{Ind,\lambda\prime}\right)^{-1}\left(\sqrt{T}R^{Ind,\lambda}\hat{\tau}\right)',$$

where $R^{Ind,\lambda} = (0, 0, 1, 0)$. Since $R^{Ind,\lambda}\tau = 0$ under the null hypothesis of $\lambda = 0$, we can apply the central limit theorem in Lemma 3 for regression (15) in the main paper to obtain $\sqrt{T}R^{Ind,\lambda}\hat{\tau} = \sqrt{T}R^{Ind,\lambda}(\hat{\tau} - \tau) \to_d N(0, R^{Ind,\lambda}\Sigma_\tau R^{Ind,\lambda\prime})$. Furthermore, by the consistency of $\hat{\Sigma}_\tau$ we have $W_T^{Ind,\lambda} \to_d \chi_1^2$. Note that if e_{t+1} is homoskedastic, then the proof is analogous to the heteroskedastic case, except $\hat{\Sigma}_\tau \to_p \sigma_e^2\Sigma_{zz}$ which can be consistently estimated by $\hat{\sigma}_e^2\left(\frac{1}{T}\sum_{t=1}^{T-1}\hat{z}_t\hat{z}_t'\right)^{-1}$, with $\hat{\sigma}_e^2 = (1/T)\sum_{t=1}^{T-1}\hat{e}_{t+1}^2$ a consistent estimator of σ_e^2 and $\hat{e}_{t+1}$ the least squares residuals.■

Proof of Theorem 2: The proof of $W_T^{Ind,\varsigma} \to_d \chi_{\ddot{h}}^2$ for testing $\varsigma_1 = \ldots = \varsigma_{\ddot{q}} = 0$ in regression (18) of the main paper follows immediately from Lemma 3 and the proof of Theorem 1. We now establish the asymptotic chi-squared distribution for the test of the null hypothesis $\dot{H}_0 : \psi_1 = \ldots = \psi_{\dot{q}} = 0$ in regression (17) of the main paper. For simplicity of exposition, let $\dot{p} = \dot{q} = \dot{h} = 1$. Following the steps in the proof of Theorem 1, we note that the infeasible regression (17) in the main paper can be rewritten as

$$H^{-1}\hat{f}_{t+1} = \hat{z}_t'\tau + u_{t+1} + \rho H^{-1}(Hf_t - \hat{f}_t) - H^{-1}(Hf_{t+1} - \hat{f}_{t+1}).$$

In the following, we denote $\tau = (\nu, \kappa, \psi, \rho H^{-1})'$ and $\hat{z}_t = (1, X_t, Y_t, \hat{f}_t)'$. It is important to remark that, comparing with the standard set up in Theorem 1, the above expression has an extra term $H^{-1}(Hf_{t+1} - \hat{f}_{t+1})$, because the dependent variable also has to be replaced by the estimated factors $\hat{f}_{t+1}$ in the feasible regression. We now write the model in matrix form

$$\hat{f}^1 H^{-1'} = \hat{z}\tau + u + (fH' - \hat{f})H^{-1'}\rho - (f^1H' - \hat{f}^1)H^{-1'}, \tag{1}$$

where $\hat{f}^1 = (\hat{f}_2, \ldots, \hat{f}_T)'$, $f^1 = (f_2, \ldots, f_T)'$, $\hat{f} = (\hat{f}_1, \ldots, \hat{f}_{T-1})'$ and $f = (f_1, \ldots, f_{T-1})'$, $u = (u_2, \ldots, u_T)'$ and $\hat{z} = (\hat{z}_1, \ldots, \hat{z}_{T-1})'$. Hence, least squares estimation of (1) yields

$$\sqrt{T}\,(\hat{\tau} - \tau) = \left(T^{-1}\hat{z}'\hat{z}\right)^{-1} T^{-1/2}\hat{z}'u + \left(T^{-1}\hat{z}'\hat{z}\right)^{-1}\left[T^{-1/2}\hat{z}'\left(fH' - \hat{f}\right)\right] H^{-1'}\rho$$
$$- \left(T^{-1}\hat{z}'\hat{z}\right)^{-1}\left[T^{-1/2}\hat{z}'\left(f^1 H' - \hat{f}^1\right)\right] H^{-1'}.$$

The second term on the right-hand side of last equation is $O_p(T^{1/2}/\min(N,T)) = o_p(1)$ by the result (iii) of Lemma 1 if $\sqrt{T}/N \to 0$. In addition, the third term is $T^{-1/2}\sum_{t=1}^{T}(\hat{f}_{t+1} - Hf_{t+1})\hat{z}_t' = O_p\left(T^{1/2}/\min(N,T)\right) = o_p(1)$ when $\sqrt{T}/N \to 0$ according to (ii) of Lemma 2.

By the result (iv) of Lemma 1 and $\sqrt{T}/N \to 0$, $T^{-1/2}\hat{z}'^{-1/2}\Phi z'u + o_p(1)$, with $\Phi = \mathrm{diag}(I, H)$ a block diagonal matrix, we obtain that $\sqrt{T}(\hat{\tau} - \tau) = (T^{-1}\hat{z}'\hat{z})^{-1}\Phi T^{-1/2} z'u + o_p(1)$ and $\sqrt{T}(\hat{\tau} - \tau) \to_d N(0, \Sigma_\tau)$, where Σ_τ can be consistently estimated using

$$\hat{\Sigma}_\tau = \left(\frac{1}{T}\sum_{t=1}^{T-1}\hat{z}_t\hat{z}_t'\right)^{-1}\left(\frac{1}{T}\sum_{t=1}^{T-1}\hat{u}_{t+1}^2\hat{z}_t\hat{z}_t'\right)\left(\frac{1}{T}\sum_{t=1}^{T-1}\hat{z}_t\hat{z}_t'\right)^{-1}$$

with $\hat{u}_{t+1}$ the OLS residuals. Finally, let $R^{Ind,\psi} = (0,0,1,0)$. The rest of the proof for the Wald-statistic $W_T^{Ind,\psi} = \left(\sqrt{T}R^{Ind,\psi}\hat{\tau}\right)\left(R^{Ind,\psi}\hat{\Sigma}_\tau R^{Ind,\psi\prime}\right)^{-1}\left(\sqrt{T}R^{Ind,\psi}\hat{\tau}\right)' \to_d \chi_1^2$ follows straightforwardly from the results on the asymptotic normality of $\hat{\tau}$, $\sqrt{T}R^{Ind,\psi}\hat{\tau} = \sqrt{T}R^{Ind,\psi}(\hat{\tau} - \tau) \to_d N(0, R^{Ind,\psi}\Sigma_\tau R^{Ind,\psi\prime})$ under the null hypothesis of $\psi = 0$, and the consistency of $\hat{\Sigma}_\tau$ to Σ_τ.$\blacksquare$

2 DGPs and simulation results

This section describes the parameter values of the data generating processes (DGPs) used in the simulation study in Section 7 of the main paper. It also provides the simulation results for the empirical size and power of the test procedures developed in the main paper.

Table 1: Data-generating processes: Direct causality

DGPs	Variables of interest		
	$X_t =$	$Y_t =$	$Z_{t,j}$ for $j = 1, \ldots, N$
DGP1	$0.5 + 0.2X_{t-1} + 0.3Y_{t-1} + \varepsilon_{t,1}$	$0.5 + 0.5Y_{t-1} + \varepsilon_{t,2}$	$Z_{t,1} = \varepsilon_{t,3},\ Z_{t,j} = \varepsilon_{t,j+2}$
DGP2	$0.5 + 0.2X_{t-1} + 0.7Y_{t-1} + \varepsilon_{t,1}$	$0.5 + 0.5Y_{t-1} + \varepsilon_{t,2}$	$Z_{t,1} = \varepsilon_{t,3},\ Z_{t,j} = \varepsilon_{t,j+2}$
DGP3	$0.5 + 0.2X_{t-1} + 0.3Y_{t-1} + \varepsilon_{t,1}$	$0.5 + 0.5Y_{t-1} + \varepsilon_{t,2}$	$Z_{t,1} = 0.5 + 0.5Z_{t-1,1} + \varepsilon_{t,3},\ Z_{t,j} = \varepsilon_{t,j+2}$
DGP4	$0.5 + 0.2X_{t-1} + 0.7Y_{t-1} + \varepsilon_{t,1}$	$0.5 + 0.5Y_{t-1} + \varepsilon_{t,2}$	$Z_{t,1} = 0.5 + 0.5Z_{t-1,1} + \varepsilon_{t,3},\ Z_{t,j} = \varepsilon_{t,j+2}$

Note: This table summarizes the DGPs, with a direct causality from Y to X, considered in the simulation study in Section 7 of the main paper to examine the performance of the tests in Theorem 1 of the main paper for testing the nulls outlined in (16), (19) and (20) of the main paper. The error terms $\varepsilon_{t,i}$, for $i = 1, 2, 3$, and $\varepsilon_{t,j+2}$ for $j = 2, \ldots, N$ are $N + 2$ mutually independent standard normal random variables, where N can be large indicating the richness of the data environment. Notice that DGP1 and DGP3 are exhibiting a relatively weaker extent of direct causality from Y to X compared to DGP2 and DGP4, in terms of the coefficients in front of Y_{t-1}; i.e. 0.3 versus 0.7.

Table 2: Data-generating processes: Indirect causality transmitted by one auxiliary variable

DGPs	Coefficients			
	Constants	X	Y	Z
DGP5	$\mu_X^{(1)} = \mu_X^{(2)} = \mu_Z = 0.5$	$\phi_X^{(1)} = \phi_X^{(2)} = 0.2$	$\phi_Y^{(1)} = \phi_Y^{(2)} = 0.3$	$\phi_Z^{(2)} = \phi_Z = 0.4$
DGP6	$\mu_X^{(1)} = \mu_X^{(2)} = \mu_Z = 0.5$	$\phi_X^{(1)} = \phi_X^{(2)} = 0.2$	$\phi_Y^{(1)} = \phi_Y^{(2)} = 0.7$	$\phi_Z^{(2)} = \phi_Z = 0.8$
DGP7	$\mu_X^{(1)} = \mu_X^{(2)} = \mu_Z = 0.5$	$\phi_X^{(1)} = \phi_X^{(2)} = 0.3$	$\phi_Y^{(1)} = \phi_Y^{(2)} = 0.2$	$\phi_Z^{(2)} = \phi_Z = 0.4$
DGP8	$\mu_X^{(1)} = \mu_X^{(2)} = \mu_Z = 0.5$	$\phi_X^{(1)} = \phi_X^{(2)} = 0.7$	$\phi_Y^{(1)} = \phi_Y^{(2)} = 0.2$	$\phi_Z^{(2)} = \phi_Z = 0.8$

Note: This table summarizes the DGPs with an indirect causality from Y to X, considered in the simulation study in Section 7 of the main paper to examine the performance of the tests in Theorem 1 of the main paper for testing the nulls outlined in (16), (19) and (20) of the main paper. The coefficients in this table are the coefficients of regression equations in (37)-(39) of the main paper. The error terms $\varepsilon_{t,i}$ for $i = 1, 2, 3$ and $\varepsilon_{t,j+2}$ for $j = 2, \ldots, N$ are $N + 2$ mutually independent standard normal random variables, where N can be large indicating the richness of the data environment.

Table 3: Empirical rejection rates of the proposed test in regression (15) based on one Z

	DGPs							
	DGP1	DGP2	DGP3	DGP4	DGP5	DGP6	DGP7	DGP8
				$T = 100$				
$N = 100$	32.9	68.2	34.3	66.3	5.4	5.4	5.0	5.5
$N = 200$	30.9	65.8	31.9	67.2	4.9	5.2	5.2	5.2
$N = 400$	32.4	65.4	29.3	65.2	5.0	5.0	5.3	8.0
				$T = 200$				
$N = 200$	56.3	93.9	57.6	91.8	4.4	6.0	4.8	5.1
$N = 400$	55.9	91.7	57.9	93.1	5.0	5.7	4.6	6.3
$N = 600$	55.1	93.2	57.5	93.1	5.0	6.5	5.0	7.5
				$T = 400$				
$N = 400$	87.4	99.9	85.8	99.9	7.3	5.2	6.1	6.9
$N = 600$	86.3	99.7	84.9	99.9	5.0	5.0	4.5	7.3
$N = 800$	86.4	99.9	85.4	99.9	5.3	4.6	4.6	8.6

Note: This table reports the empirical size and power of the test stated in Theorem 1 of the main paper for testing condition (ii) of Definition 1 of indirect causality from Y to X at $\alpha = 5\%$ significance level in regression (15) of the main paper. The number of simulations is equal to 2000 replications. The indirect causality is transmitted by one auxiliary variable.

Table 4: Empirical rejection rates of the proposed test in regression (17) based on one Z

	DGPs							
	DGP1	DGP2	DGP3	DGP4	DGP5	DGP6	DGP7	DGP8
				$T = 100$				
$N = 100$	5.3	5.8	5.0	5.3	100.0	100.0	100.0	100.0
$N = 200$	4.6	6.0	4.8	5.2	100.0	100.0	100.0	100.0
$N = 400$	4.7	5.2	5.0	5.5	100.0	100.0	100.0	100.0
				$T = 200$				
$N = 200$	5.6	4.7	5.1	5.4	100.0	100.0	100.0	100.0
$N = 400$	4.8	4.7	4.6	4.7	100.0	100.0	100.0	100.0
$N = 600$	5.9	4.9	5.6	5.4	100.0	100.0	100.0	100.0
				$T = 400$				
$N = 400$	4.5	4.4	5.0	5.0	100.0	100.0	100.0	100.0
$N = 600$	4.7	4.9	6.1	4.9	100.0	100.0	100.0	100.0
$N = 800$	5.2	5.6	5.1	4.9	100.0	100.0	100.0	100.0

Note: This table reports the empirical size and power of the test stated in Theorem 2 of the main paper for testing the null hypothesis (19) in the main paper for the condition (iii) of Definition 1 of indirect causality from Y to X at $\alpha = 5\%$ significance level in regression (17) of the main paper. The number of simulations is equal to 2000 replications. The indirect causality is transmitted by one auxiliary variable.

Table 5: Empirical rejection rates of the proposed test in regression (18) based on one Z

	DGPs							
	DGP1	DGP2	DGP3	DGP4	DGP5	DGP6	DGP7	DGP8
				$T = 100$				
$N = 100$	5.2	5.5	4.5	4.8	48.9	99.7	38.8	41.6
$N = 200$	4.6	4.8	5.0	4.8	47.6	99.7	37.3	41.2
$N = 400$	5.5	5.4	5.9	4.8	47.5	98.8	37.4	39.2
				$T = 200$				
$N = 200$	5.5	4.4	6.1	4.5	79.2	100.0	65.4	74.5
$N = 400$	5.0	5.8	5.5	5.7	79.4	100.0	65.9	74.1
$N = 600$	5.1	4.6	5.2	4.8	76.7	100.0	65.6	72.5
				$T = 400$				
$N = 400$	5.0	4.3	5.3	4.4	97.5	100.0	92.7	96.6
$N = 600$	5.0	5.0	4.8	4.7	97.7	100.0	92.7	97.1
$N = 800$	5.2	4.4	4.9	5.0	96.7	100.0	92.5	96.4

Note: This table reports the empirical size and power of the test stated in Theorem 2 of the main paper for testing the null hypothesis (20) in the main paper for the condition (iii) of Definition 1 of indirect causality from Y to X at $\alpha = 5\%$ significance level in regression (18). The number of simulations is equal to 2000. The indirect causality is transmitted by one auxiliary variable.

Table 6: Data-generating processes: Direct causality and indirect causality transmitted by many auxiliary variables

DGPs	Variables of Interest		
	$X_t =$	$Y_t =$	Z_t
		Direct Causality	
DGP9	$0.5 + 0.5X_{t-1} + 0.1Y_{t-1} + \varepsilon_{1t}$	$0.5 + 0.5Y_{t-1} + \varepsilon_{2t}$	$Z_{t,j} = \varepsilon_{j+2,t}$, for $j = 1, \ldots, 10$, $Z_{t,j} = \varepsilon_{j+2,t}$, for $j = 11, \ldots, N$
DGP10	$0.5 + 0.5X_{t-1} + 0.1Y_{t-1} + \varepsilon_{1t}$	$0.5 + 0.5Y_{t-1} + \varepsilon_{2t}$	$Z_{t,j} = 0.5 + 0.5Z_{j,t-1} + \varepsilon_{j+2,t}$, for $j = 1, .., 10$, $Z_{t,j} = \varepsilon_{j+2,t}$, for $j = 11, .., N$
		Indirect Causality	
DGP11	$\phi_X^{(1)} = \phi_X^{(2)} = 0.1$	$\phi_Y^{(1)} = \phi_Y^{(2)} = 0.05$	$\phi_{Z_1}^{(2)} = \cdots = \phi_{Z_{10}}^{(2)} = 0.01$, $\phi_{Z_1} = \cdots = \phi_{Z_{10}} = 0.2$
DGP12	$\phi_X^{(1)} = \phi_X^{(2)} = 0.2$	$\phi_Y^{(1)} = \phi_Y^{(2)} = 0.3$	$\phi_{Z_1}^{(2)} = \cdots = \phi_{Z_{10}}^{(2)} = 0.05$, $\phi_{Z_1} = \cdots = \phi_{Z_{10}} = 0.4$
DGP13	$\phi_X^{(1)} = \phi_X^{(2)} = 0.2$	$\phi_Y^{(1)} = \phi_Y^{(2)} = 0.7$	$\phi_{Z_1}^{(2)} = \cdots = \phi_{Z_{10}}^{(2)} = 0.1$, $\phi_{Z_1} = \cdots = \phi_{Z_{10}} = 0.8$

Note: This table summarizes the DGPs, with direct and indirect causalities from Y to X, considered in the simulation study in Section 7 of the main paper to examine the performance of the tests for the hypothesis testing problems outlined in (16), (19) and (20) of the main paper. The error terms $\varepsilon_{t,i}$ for $j = 1, \ldots, N+2$ are mutually independent standard normal random variables, where N can be large indicating the richness of the data environment. The constant terms $\mu_X^{(1)}$, $\mu_X^{(2)}$, $\mu_{Z_1}, \cdots, \mu_{Z_{10}}$ in the equations (40), (41) and (42) of the main paper for DGP11-DGP13 are all equal to $0.5 : \mu_X^{(1)} = \mu_X^{(2)} = \mu_{Z_1} = \cdots = \mu_{Z_{10}} = 0.5$.

Table 7: Empirical rejection rates of the proposed test in regression (15) based on ten Zs

	DGPs				
	DGP9	DGP10	DGP11	DGP12	DGP13
			$T = 100$		
$N = 100$	7.0	7.9	4.2	5.0	4.8
$N = 200$	9.0	9.7	5.6	4.8	5.8
$N = 400$	9.3	9.6	4.3	5.1	4.4
			$T = 200$		
$N = 200$	12.0	13.3	4.4	4.8	4.8
$N = 400$	11.9	12.0	5.3	5.0	5.0
$N = 600$	12.7	14.0	5.2	4.3	5.0
			$T = 400$		
$N = 400$	21.1	20.4	4.8	5.7	4.9
$N = 600$	20.7	19.2	5.0	5.4	4.6
$N = 800$	19.4	21.2	5.3	4.8	5.3

Note: This table reports the empirical size and power of the test stated in Theorem 1 of the main paper for testing condition (ii) of Definition 1 of indirect causality from Y to X at $\alpha = 5\%$ significance level in regression (15) of the main paper. The number of simulations is equal to 2000. The indirect causality is transmitted by ten auxiliary variables.

Table 8: Empirical rejection rates of the proposed test in regression (17) based on ten Zs

	DGPs				
	DGP9	DGP10	DGP11	DGP12	DGP13
		$T = 100$			
$N = 100$	5.9	5.5	100.0	100.0	100.0
$N = 200$	5.9	6.6	100.0	100.0	100.0
$N = 400$	5.6	5.7	100.0	100.0	100.0
		$T = 200$			
$N = 200$	4.5	5.4	100.0	100.0	100.0
$N = 400$	6.2	4.6	100.0	100.0	100.0
$N = 600$	5.8	5.2	100.0	100.0	100.0
		$T = 400$			
$N = 400$	4.6	4.9	100.0	100.0	100.0
$N = 600$	5.2	5.0	100.0	100.0	100.0
$N = 800$	5.2	6.7	100.0	100.0	100.0

Note: This table reports the empirical size and power of the test stated in Theorem 2 of the main paper for testing the null hypothesis (19) in the main paper of the condition (iii) of Definition 1 of indirect causality from Y to X at $\alpha = 5\%$ significance level in regression (17) of the main paper. The number of simulations is equal to 2000. The indirect causality is transmitted by ten auxiliary variables.

Table 9: Empirical rejection rates of the proposed test in regression (18) based on ten Zs

	DGPs				
	DGP9	DGP10	DGP11	DGP12	DGP13
			$T = 100$		
$N = 100$	5.3	5.6	8.0	32.7	99.5
$N = 200$	4.5	5.3	7.9	33.9	99.3
$N = 400$	4.9	5.1	8.6	31.8	99.0
			$T = 200$		
$N = 200$	5.4	4.6	10.5	57.4	100.0
$N = 400$	5.5	5.9	10.2	58.6	100.0
$N = 600$	4.9	5.0	9.1	57.2	100.0
			$T = 400$		
$N = 400$	4.9	4.3	14.4	85.4	100.0
$N = 600$	5.2	5.6	15.4	85.0	100.0
$N = 800$	4.7	4.9	14.8	85.4	100.0

Note: This table reports the empirical size and power of the test stated in Theorem 2 of the main paper for testing the null hypothesis (20) in the main paper of the condition (iii) of Definition 1 of indirect causality from Y to X at $\alpha = 5\%$ significance level in regression (18) of the main paper. The number of simulations is equal to 2000. The indirect causality is transmitted by ten auxiliary variables.

Table 10: Data-generating processes: Non-causality cases

DGPs	Variables of Interest		
	$X_t =$	$Y_t =$	$Z_{t,j}$ for $j = 1, \ldots, N$
DGP14	$0.5 + 0.8X_{t-1} + \varepsilon_{t,1}$	$0.5 + 0.8Y_{t-1} + \varepsilon_{t,2}$	$Z_{t,1} = \varepsilon_{t,3},\ Z_{t,j} = \varepsilon_{t,j+2}$
DGP15	$0.5 + 0.2X_{t-1} + \varepsilon_{t,1}$	$0.5 + 0.2Y_{t-1} + \varepsilon_{t,2}$	$Z_{t,1} = \varepsilon_{t,3},\ Z_{t,j} = \varepsilon_{t,j+2}$
DGP16	$0.5 + 0.8X_{t-1} + \varepsilon_{t,1}$	$0.5 + 0.8Y_{t-1} + \varepsilon_{t,2}$	$Z_{t,1} = 0.5 + 0.8Z_{t-1,1} + \varepsilon_{t,3},\ Z_{t,j} = \varepsilon_{t,j+2}$
DGP17	$0.5 + 0.2X_{t-1} + \varepsilon_{t,1}$	$0.5 + 0.2Y_{t-1} + \varepsilon_{t,2}$	$Z_{t,1} = 0.5 + 0.2Z_{t-1,1} + \varepsilon_{t,3},\ Z_{t,j} = \varepsilon_{t,j+2}$

Note: This table summarizes the DGPs, with no spurious causality of type I from Y to X, considered in the simulation study in Section 7 of the main paper to examine the size of the tests for the hypothesis testing problems outlined in (22), (26), and (27) of the main paper. The error terms $\varepsilon_{t,i}$ for $j = 1, \ldots, N + 2$ are mutually independent standard normal random variables, where N can be large indicating the richness of the data environment.

Table 11: Data-generating processes: Spurious causality of type I

DGPs	Coefficients			
	Constants	X	Y	Z
DGP18	$\mu_X^{(1)} = \mu_X^{(2)} = \mu_Z = 0.5$	$\phi_X^{(1)} = \phi_X^{(2)} = \phi_X^{(3)} = 0.1$	$\phi_Y^{(3)} = 0.1$	$\phi_Z^{(2)} = \phi_Z^{(3)} = 0.1$
DGP19	$\mu_X^{(1)} = \mu_X^{(2)} = \mu_Z = 0.5$	$\phi_X^{(1)} = \phi_X^{(2)} = \phi_X^{(3)} = 0.2$	$\phi_Y^{(3)} = 0.2$	$\phi_Z^{(2)} = \phi_Z^{(3)} = 0.2$
DGP20	$\mu_X^{(1)} = \mu_X^{(2)} = \mu_Z = 0.5$	$\phi_X^{(1)} = \phi_X^{(2)} = \phi_X^{(3)} = 0.3$	$\phi_Y^{(3)} = 0.3$	$\phi_Z^{(2)} = \phi_Z^{(3)} = 0.3$
DGP21	$\mu_X^{(1)} = \mu_X^{(2)} = \mu_Z = 0.5$	$\phi_X^{(1)} = \phi_X^{(2)} = \phi_X^{(3)} = 0.4$	$\phi_Y^{(3)} = 0.4$	$\phi_Z^{(2)} = \phi_Z^{(3)} = 0.4$

Note: This table summarizes the DGPs, with a spurious causality of type I from Y to X, considered in the simulation study in Section 7 of the main paper to examine the power of the tests for the hypothesis testing problems outlined in (22), (26), and (27) of the main paper. The error terms $\varepsilon_{t,i}$ for $j = 1, \ldots, N + 2$ are mutually independent standard normal random variables, where N can be large indicating the richness of the data environment.

Table 12: Empirical rejection rates of the proposed test in regression (21) based on one Z

	DGPs							
	DGP14	DGP15	DGP16	DGP17	DGP18	DGP19	DGP20	DGP21
				$T = 100$				
$N = 100$	6.6	4.7	5.9	5.3	100.0	100.0	100.0	100.0
$N = 200$	6.6	5.1	7.2	5.8	100.0	100.0	100.0	100.0
$N = 400$	6.7	5.7	6.5	5.4	100.0	100.0	100.0	100.0
				$T = 200$				
$N = 200$	6.4	5.0	6.9	5.6	100.0	100.0	100.0	100.0
$N = 400$	5.7	5.1	6.6	5.2	100.0	100.0	100.0	100.0
$N = 600$	6.2	5.5	6.9	5.4	100.0	100.0	100.0	100.0
				$T = 400$				
$N = 400$	5.2	5.2	5.8	5.3	100.0	100.0	100.0	100.0
$N = 600$	5.2	4.9	5.6	4.7	100.0	100.0	100.0	100.0
$N = 800$	5.6	5.5	5.3	4.6	100.0	100.0	100.0	100.0

Note: This table reports the empirical size and power of the test stated in Theorem 3 of the main paper for testing condition (i) of Definition 2 of spurious causality of type I from Y to X at $\alpha = 5\%$ significance level in regression (22) of the main paper. The number of simulations is equal to 2000.

Table 13: Empirical rejection rates of the proposed test in regression (24) based on one Z

	DGPs							
	DGP14	DGP15	DGP16	DGP17	DGP18	DGP19	DGP20	DGP21
				$T = 100$				
$N = 100$	5.3	5.4	7.3	5.4	100.0	100.0	100.0	100.0
$N = 200$	5.3	4.8	8.2	5.1	100.0	100.0	100.0	100.0
$N = 400$	5.6	5.3	7.4	4.1	100.0	100.0	100.0	100.0
				$T = 200$				
$N = 200$	4.6	5.0	7.6	5.1	100.0	100.0	100.0	100.0
$N = 400$	4.8	4.6	7.1	5.1	100.0	100.0	100.0	100.0
$N = 600$	5.5	4.7	6.7	5.2	100.0	100.0	100.0	100.0
				$T = 400$				
$N = 400$	5.0	5.3	8.0	5.1	100.0	100.0	100.0	100.0
$N = 600$	6.2	3.9	8.0	5.4	100.0	100.0	100.0	100.0
$N = 800$	5.0	4.6	7.6	5.0	100.0	100.0	100.0	100.0

Note: This table reports the empirical size and power of the test stated in Theorem 3 of the main paper for testing the null hypothesis (26) in the main paper of the condition (iii) of Definition 2 of spurious causality of type I from Y to X at $\alpha = 5\%$ significance level in regression (24) of the main paper. The number of simulations is equal to 2000.

Table 14: Empirical rejection rates of the proposed test in regression (25) based on one Z

	DGPs							
	DGP14	DGP15	DGP16	DGP17	DGP18	DGP19	DGP20	DGP21
				$T = 100$				
$N = 100$	5.0	5.4	7.0	4.8	100.0	100.0	100.0	100.0
$N = 200$	4.2	4.6	5.7	5.7	100.0	100.0	100.0	100.0
$N = 400$	4.7	4.4	6.4	5.6	100.0	100.0	100.0	100.0
				$T = 200$				
$N = 200$	4.2	5.7	5.1	5.0	100.0	100.0	100.0	100.0
$N = 400$	5.1	5.1	5.6	4.1	100.0	100.0	100.0	100.0
$N = 600$	4.7	5.0	4.8	4.9	100.0	100.0	100.0	100.0
				$T = 400$				
$N = 400$	5.5	4.5	5.5	4.5	100.0	100.0	100.0	100.0
$N = 600$	4.8	5.0	4.5	4.7	100.0	100.0	100.0	100.0
$N = 800$	4.3	5.3	5.9	5.3	100.0	100.0	100.0	100.0

Note: This table reports the empirical size and power of the test stated in Theorem 3 of the main paper for testing the null hypothesis (27) in the main paper of the condition (iii) of Definition 2 of spurious causality of type I from Y to X at $\alpha = 5\%$ significance level in regression (25) of the main paper. The number of simulations is equal to 2000.

3 Data and empirical results

This section describes the dataset used in the empirical application of the main paper [see their Section 8], and it provides the empirical results obtained using this data.

The dataset consists of monthly observations on 135 economic variables from Federal Reserve Bank of St. Louis (FRED). The sample runs from January 1959 to May 2016 for a total of 689 observations. All the variables are reported in Tables 15-20 below. In particular, the following 8 groups of variables are considered: (1) Output and income with 17 variables; (2) Labor market with 32 variables; (3) Housing with 10 variables; (4) Consumption, orders, and inventories with 14 variables; (5) Money and credit with 14 variables; (6) Interest and exchange rates with 22 variables; (7) Prices with 21 variables; and (8) Stock market with 5 variables. This big data mimic the coverage of datasets already used in the literature and it is updated in real-time through the FRED database. A detailed description of the dataset can be found in McCracken and Ng (2015).

Table 15: Description of the variables of the big data

	id	tcode	fred	description
			Group 1: Output and income	
1	1	5	RPI	Real Personal Income
2	2	5	W875RX1	Real personal income ex transfer receipts
3	6	5	INDPRO	IP Index
4	7	5	IPFPNSS	IP: Final Products and Nonindustrial Supplies
5	8	5	IPFINAL	IP: Final Products (Market Group)
6	9	5	IPCONGD	IP: Consumer Goods
7	10	5	IPDCONGD	IP: Durable Consumer Goods
8	11	5	IPNCONGD	IP: Nondurable Consumer Goods
9	12	5	IPBUSEQ	IP: Business Equipment
10	13	5	IPMAT	IP: Materials
11	14	5	IPDMAT	IP: Durable Materials
12	15	5	IPNMAT	IP: Nondurable Materials
13	16	5	IPMANSICS	IP: Manufacturing (SIC)
14	17	5	IPB51222s	IP: Residential Utilities
15	18	5	IPFUELS	IP: Fuels
16	19	1	NAPMPI	ISM Manufacturing: Production Index
17	20	2	CUMFNS	Capacity Utilization: Manufacturing
			Group 2: Labor market	
1	21	2	HWI	Help-Wanted Index for United States
2	22	2	HWIURATIO	Ratio of Help Wanted/No. Unemployed
3	23	5	CLF16OV	Civilian Labor Force
4	24	5	CE16OV	Civilian Employment
5	25	2	UNRATE	Civilian Unemployment Rate
6	26	2	UEMPMEAN	Average Duration of Unemployment (Weeks)
7	27	5	UEMPLT5	Civilians Unemployed - Less Than 5 Weeks
8	28	5	UEMP5TO14	Civilians Unemployed for 5-14 Weeks
9	29	5	UEMP15OV	Civilians Unemployed - 15 Weeks & Over
10	30	5	UEMP15T26	Civilians Unemployed for 15-26 Weeks

Note: This table presents the variables included in the groups "Output and income" and "Labor market".

Table 16: Description of the variables of big data (Cont.)

	id	tcode	fred	description
			Group 2: Labor market (Cont.)	
11	31	5	UEMP27OV	Civilians Unemployed for 27 Weeks and Over
12	32	5	CLAIMSx	Initial Claims
13	33	5	PAYEMS	All Employees: Total nonfarm
14	34	5	USGOOD	All Employees: Goods-Producing Industries
15	35	5	CES1021000001	All Employees: Mining and Logging: Mining
16	36	5	USCONS	All Employees: Construction
17	37	5	MANEMP	All Employees: Manufacturing
18	38	5	DMANEMP	All Employees: Durable goods
19	39	5	NDMANEMP	All Employees: Nondurable goods
20	40	5	SRVPRD	All Employees: Service-Providing Industries
21	41	5	USTPU	All Employees: Trade, Transportation & Utilities
22	42	5	USWTRADE	All Employees: Wholesale Trade
23	43	5	USTRADE	All Employees: Retail Trade
24	44	5	USFIRE	All Employees: Financial Activities
25	45	5	USGOVT	All Employees: Government
26	46	1	CES0600000007	Avg Weekly Hours : Goods-Producing
27	47	2	AWOTMAN	Avg Weekly Overtime Hours : Manufacturing
28	48	1	AWHMAN	Avg Weekly Hours : Manufacturing
29	49	1	NAPMEI	ISM Manufacturing: Employment Index
30	127	6	CES0600000008	Avg Hourly Earnings : Goods-Producing
31	128	6	CES2000000008	Avg Hourly Earnings : Construction
32	129	6	CES3000000008	Avg Hourly Earnings : Manufacturing
			Group 3: Housing	
1	50	4	HOUST	Housing Starts: Total New Privately Owned
2	51	4	HOUSTNE	Housing Starts, Northeast
3	52	4	HOUSTMW	Housing Starts, Midwest
4	53	4	HOUSTS	Housing Starts, South
5	54	4	HOUSTW	Housing Starts, West

Note: This table presents the variables included in the groups “Labor market” and “Housing”.

Table 17: Description of the variables of big data (Cont.)

	id	tcode	fred	description
			Group 3: Housing (Cont.)	
6	55	4	PERMIT	New Private Housing Permits (SAAR)
7	56	4	PERMITNE	New Private Housing Permits, Northeast (SAAR)
8	57	4	PERMITMW	New Private Housing Permits, Midwest (SAAR)
9	58	4	PERMITS	New Private Housing Permits, South (SAAR)
10	59	4	PERMITW	New Private Housing Permits, West (SAAR)
			Group 4: Consumption, orders, and inventories	
1	3	5	DPCERA3M086SBEA	Real personal consumption expenditures
2	4	5	CMRMTSPLx	Real Manu. and Trade Industries Sales
3	5	5	RETAILx	Retail and Food Services Sales
4	60	1	NAPM	ISM : PMI Composite Index
5	61	1	NAPMNOI	ISM : New Orders Index
6	62	1	NAPMSDI	ISM : Supplier Deliveries Index
7	63	1	NAPMII	ISM : Inventories Index
8	64	5	ACOGNO	New Orders for Consumer Goods
9	65	5	AMDMNOx	New Orders for Durable Goods
10	66	5	ANDENOx	New Orders for Nondefense Capital Goods
11	67	5	AMDMUOx	Unfilled Orders for Durable Goods
12	68	5	BUSINVx	Total Business Inventories
13	69	2	ISRATIOx	Total Business: Inventories to Sales Ratio
14	130	2	UMCSENTx	Consumer Sentiment Index
			Group 5: Money and credit	
1	70	6	M1SL	M1 Money Stock
2	71	6	M2SL	M2 Money Stock
3	72	5	M2REAL	Real M2 Money Stock
4	73	6	AMBSL	St. Louis Adjusted Monetary Base
5	74	6	TOTRESNS	Total Reserves of Depository Institutions
6	75	7	NONBORRES	Reserves Of Depository Institutions
7	76	6	BUSLOANS	Commercial and Industrial Loans

Note: This table presents the variables included in the groups "Housing", "Consumption, orders, and inventories" and "Money and credit".

Table 18: Description of the variables of big data (Cont.)

	id	tcode	fred	description
			Group 5: Money and credit (Cont.)	
8	77	6	REALLN	Real Estate Loans at All Commercial Banks
9	78	6	NONREVSL	Total Nonrevolving Credit
10	79	2	CONSPI	Nonrevolving consumer credit to Personal Income
11	131	6	MZMSL	MZM Money Stock
12	132	6	DTCOLNVHFNM	Consumer Motor Vehicle Loans Outstanding
13	133	6	DTCTHFNM	Total Consumer Loans and Leases Outstanding
14	134	6	INVEST	Securities in Bank Credit at All Commercial Banks
			Group 6: Interest and exchange rates	
1	84	2	FEDFUNDS	Effective Federal Funds Rate
2	85	2	CP3Mx	3-Month AA Financial Commercial Paper Rate
3	86	2	TB3MS	3-Month Treasury Bill:
4	87	2	TB6MS	6-Month Treasury Bill:
5	88	2	GS1	1-Year Treasury Rate
6	89	2	GS5	5-Year Treasury Rate
7	90	2	GS10	10-Year Treasury Rate
8	91	2	AAA	Moody's Seasoned Aaa Corporate Bond Yield
9	92	2	BAA	Moody's Seasoned Baa Corporate Bond Yield
10	93	1	COMPAPFFx	3-Month Commercial Paper Minus FEDFUNDS
11	94	1	TB3SMFFM	3-Month Treasury C Minus FEDFUNDS
12	95	1	TB6SMFFM	6-Month Treasury C Minus FEDFUNDS
13	96	1	T1YFFM	1-Year Treasury C Minus FEDFUNDS
14	97	1	T5YFFM	5-Year Treasury C Minus FEDFUNDS
15	98	1	T10YFFM	10-Year Treasury C Minus FEDFUNDS
16	99	1	AAAFFM	Moody's Aaa Corporate Bond Minus FEDFUNDS
17	100	1	BAAFFM	Moody's Baa Corporate Bond Minus FEDFUNDS
18	101	5	TWEXMMTH	Trade Weighted U.S. Dollar Index: Major Currencies
19	102	5	EXSZUSx	Switzerland / U.S. Foreign Exchange Rate
20	103	5	EXJPUSx	Japan / U.S. Foreign Exchange Rate

Note: This table presents the variables included in the groups "Money and credit" and "Interest and exchange rates".

Table 19: Description of the variables of big data (Cont.)

	id	tcode	fred	description
			Group 6: Interest and exchange rates (Cont.)	
21	104	5	EXUSUKx	U.S. / U.K. Foreign Exchange Rate
22	105	5	EXCAUSx	Canada / U.S. Foreign Exchange Rate
			Group 7: Prices	
1	106	6	WPSFD49207	PPI: Finished Goods
2	107	6	WPSFD49502	PPI: Finished Consumer Goods
3	108	6	WPSID61	PPI: Intermediate Materials
4	109	6	WPSID62	PPI: Crude Materials
5	110	6	OILPRICEx	Crude Oil, spliced WTI and Cushing
6	111	6	PPICMM	PPI: Metals and metal products:
7	112	1	NAPMPRI	ISM Manufacturing: Prices Index
8	113	6	CPIAUCSL	CPI : All Items
9	114	6	CPIAPPSL	CPI : Apparel
10	115	6	CPITRNSL	CPI : Transportation
11	116	6	CPIMEDSL	CPI : Medical Care
12	117	6	CUSR0000SAC	CPI : Commodities
13	118	6	CUUR0000SAD	CPI : Durables
14	119	6	CUSR0000SAS	CPI : Services
15	120	6	CPIULFSL	CPI : All Items Less Food
16	121	6	CUUR0000SA0L2	CPI : All items less shelter
17	122	6	CUSR0000SA0L5	CPI : All items less medical care
18	123	6	PCEPI	Personal Cons. Expend.: Chain Index
19	124	6	DDURRG3M086SBEA	Personal Cons. Exp: Durable goods
20	125	6	DNDGRG3M086SBEA	Personal Cons. Exp: Nondurable goods

Note: This table presents the variables included in the groups “Interest and exchange rates” and “Prices”.

Table 20: Description of the variables of big data (Cont.)

	id	tcode	fred	description
			Group 7: Prices (Cont.)	
21	126	6	DSERRG3M086SBEA	Personal Cons. Exp: Services
			Group 8: Stock market	
1	80	5	S&P 500	S&P's Common Stock Price Index: Composite
2	81	5	S&P: indust	S&P's Common Stock Price Index: Industrials
3	82	2	S&P div yield	S&P's Composite Common Stock: Dividend Yield
4	83	5	S&P PE ratio	S&P's Composite Common Stock: Price-Earnings Ratio
5	135	1	VXOCLSx	VXO

Note: This table presents the variables included in the groups "Prices" and "Stock market".

Table 21: Testing Indirect Causality between Income and Money, Credit

Tested conditions		Causal variable (Y)		
		Money	BUSLOANS	INVEST
(i):	$Y \to X$	0.8655	0.0728	0.0717
(ii):	$Y \to X\,\vert f$	0.8098	0.5618	0.4222
(iii).a:	$Y \to f\,\vert X$	0.0195	0.0385	0.0011
(iii).b:	$f \to X\,\vert Y$	0.0012	0.0005	0.0009

Note: This table summarizes the results of testing conditions (i)-(iii) of Definition 1 of an indirect causality from Y =Money, Commercial and Industrial Loans [BUSLOANS] and Securities in Bank Credit at All Commercial Banks [INVEST] to X =Income. f represents the factors extracted from the big data set [see Section 4.1 of the main paper]. Money is measured by M1 Money Stock (M1SL) in Table 17.

Table 22: Identification of the auxiliary variables responsible for the transmission of indirect causality from Credit, Investment to Income

Auxiliary variables (Z)	
Indirect causality from BUSLOANS to Income	**Indirect causality from INVEST to Income**
All Employees: Mining and Logging: Mining	IP: Nondurable Materials
Avg Weekly Overtime Hours : Manufacturing	Housing Starts: Total New Privately Owned
New Private Housing Permits (SAAR)	Housing Starts, West
New Private Housing Permits, South (SAAR)	New Private Housing Permits (SAAR)
New Private Housing Permits, West (SAAR)	New Private Housing Permits, Midwest (SAAR)
Total Business: Inventories to Sales Ratio	New Private Housing Permits, South (SAAR)
Effective Federal Funds Rate	New Private Housing Permits, West (SAAR)
3-Month AA Financial Commercial Paper Rate	Real Manu. and Trade Industries Sales
1-Year Treasury Rate	Total Business: Inventories to Sales Ratio
Moody's Seasoned Baa Corporate Bond Yield	5-Year Treasury C Minus FEDFUNDS
3-Month Treasury C Minus FEDFUNDS	10-Year Treasury C Minus FEDFUNDS
6-Month Treasury C Minus FEDFUNDS	Moody's Aaa Corporate Bond Minus FEDFUNDS
1-Year Treasury C Minus FEDFUNDS	Moody's Baa Corporate Bond Minus FEDFUNDS
5-Year Treasury C Minus FEDFUNDS	S&P's Composite Common Stock: Dividend Yield
10-Year Treasury C Minus FEDFUNDS	-
Moody's Aaa Corporate Bond Minus FEDFUNDS	-
Moody's Baa Corporate Bond Minus FEDFUNDS	-
S&P's Composite Common Stock: Dividend Yield	-

Note: This table summarizes the results of identifying the auxiliary variables responsible for the transmission of indirect causality from credit measured by Commercial and Industrial Loans and Securities in Bank Credit at All Commercial Banks to Income. The results are obtained using the statistical procedure described in Section 6 of the main paper and based on the big data described in tables 15-20.

References

[1] Bai, J., Ng, S. (2005). "Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions," *Working paper*, University of Michigan.

[2] Bai, J., Ng, S. (2006). "Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions," *Econometrica*, vol. 74, pp. 1133–1150.

[3] McCracken, M.W., Ng, S. (2015). "FRED-MD: A monthly database for macroeconomic research," Federal Reserve Bank of St. Louis Working Paper Series.