Chapter 5
Cause-Effect Pairs in Time Series with a Focus on Econometrics

Nicolas Doremus, Alessio Moneta, and Sebastiano Cattaruzzo

5.1 Introduction

Let us consider two scalar stochastic processes x_t and y_t, $t \in \mathbb{Z}$, each observed for T realizations. We assume that x_t and y_t are covariance stationary or that Δx_t and Δy_t are covariance stationary. Most time series observed in macroeconomics, for example, belong to this class of processes (see e.g. [29]). If we exclude the possibility that the future can cause the past, but we allow contemporaneous feedback loops due for example to temporal aggregation, there are several possibilities as regards the causal structure between x_t and y_t, which we list here below. We denote causal relationships[1] with directed edges ($\rightarrow$), following the graphical causal models terminology [64].

[1]When referring to "causal relationships", we endorse here, in the spirit of Hoover [32], Pearl [55], a structural account of causality: causal relationships are the fundamental, but usually latent, building blocks of the mechanism that has generate the observed data, which we aim at representing through a structural (or causal) model. While a structural model entails probabilistic relations, it contains more information than a statistical model, because it allows us to analyze the effect of interventions (cf. [58]).

N. Doremus (✉)
IUSS Pavia, Pavia, Italy
e-mail: nicolas.doremus@iusspavia.it

A. Moneta
Sant'Anna School of Advanced Studies, Pisa, Italy
e-mail: a.moneta@santannapisa.it

S. Cattaruzzo
Rovira i Virgili University, Tarragona, Spain
e-mail: sebastiano.cattaruzzo@urv.cat

I. Guyon et al. (eds.), *Cause Effect Pairs in Machine Learning*,
The Springer Series on Challenges in Machine Learning,
https://doi.org/10.1007/978-3-030-21810-2_5

(i) The series x_t has a contemporaneous or lagged causal effect on y_t, i.e. $x_i \rightarrow y_{i+s}$ for some i, s such that $i \geq 0, s \geq 0$.
(ii) The series y_t has a contemporaneous or lagged causal effect on x_t, i.e. $y_i \rightarrow x_{i+s}$ for some i, s such that $i \geq 0, s \geq 0$.
(iii) A not-measured series z_t has a contemporaneous or lagged causal effect on both x_t and y_t.
(iv) The causal structure between x_t and y_t can be described by any combination of (i)–(iii).
(v) There is no causal link or path (of any type) linking x_t and y_{t+s}, for any $s \in \mathbb{N}$.

In principle, other, more involute, causal structures are possible between x_t and y_t. For example, the data generating process may have a frequency that is different from the frequency of data collection, so that there are hidden causal structures between the observed variables. This class of structures has been considered in the literature on temporal aggregation in econometrics (see e.g. [17, 18, 50]) and in the literature on subsampling in machine learning (see [10, 36]), but will not be further discussed in this paper. We will also limit our discussion on structures in which variables are well-defined (i.e. they are not aggregate of variables with diverse causal roles) and the causal structures are *time invariant*: i.e. if $x_i \rightarrow w_{i+s}$ given any $s \in \mathbb{Z}$, then this true for all $i \in \mathbb{Z}$, where w can be any variable (included x itself). We will also typically assume that each observed series w_t will be directly causally influenced by its own past, until a certain lag and that each variable at each time unit will be affected, in an additive manner, by one or more independent shock. In other words, we focus on *additive noise models*.

The causal structure between two time series can be represented by a causal graph consisting of nodes for $x_t, \ldots, x_{t-p}, y_t, \ldots, y_{t-p}$, where p is the largest lag by which x_t or y_t can be directly causally influenced. Using the terminology proposed by Chu and Glymour [7], this graph is called a *unit causal graph*. Examples for unit causal graphs are shown in Figs. 5.1 and 5.2, for $p = 2$. Figure 5.1 represents the case in which (i) is true, while Fig. 5.2 represents the case in which (iii) is true. Chu and Glymour [7] notice that a unit causal graph can be extended to *repetitive causal graph* (not shown), including the variables x_t and y_t at a potentially infinite time units. The repetitive causal graph corresponding to the unit causal graph of Fig. 5.1, for example, would include nodes for $x_{t-3}, x_{t-4}, \ldots$, for $y_{t-3}, y_{t-4}, \ldots$ and direct edges from x_{t-s} to y_{t-s}, as well as $x_{t-s-2} \rightarrow y_{t-s}$ and $x_{t-s-1} \rightarrow y_{t-s}$, for any $s \in \mathbb{Z}$.

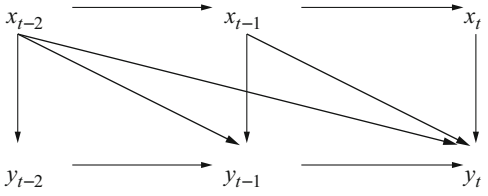


Fig. 5.1 Unit causal graph for bi-variate time series with both lagged and contemporaneous effects

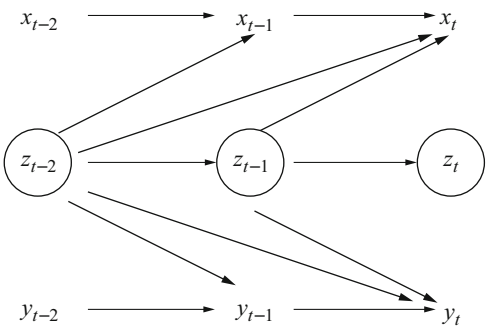


Fig. 5.2 Unit causal graph for bi-variate time series with a latent series z_t

How do we detect which of the five cases listed above (i)–(v) is true? How do we learn which causal graph better represents the data generating process? How do we learn to what extent an intervention on one variable at time t propagates on all the variables at time $t + h$ for any value of $h > 0$? These are the typical questions that concern, for example, the applied macro-econometrician. In this paper we discuss possible manners to address these questions. We review methods that are able to disentangle among different causal structures, under different assumptions.

Some causal discovery methods developed for i.i.d. data cannot be applied, without further modification, to the time series setting, due to the fact that, even in a simple setting of causal pairs, there is the possibility of causal relationships with different effects at different lags. Furthermore, the autocorrelation (or self-dependence) structure underlying the data introduces some complications in standard statistical inference that reduce the efficiency of simple regression estimation or conditional independence testing [26]. Nevertheless, the time series setting is not necessarily a curse, and is actually a blessing in specific contexts of causal inference. Indeed, if one accepts the assumption that the future cannot cause the past (whose acceptance in economics involves a careful taking into account of expectational variables, see [33]), exploiting the arrow of time allows one to solve many *orientation* problems, i.e. problems where it is known that there is a causal dependence between two variables, but not the direction. Moreover, in the case of causal pairs, the possibility of observing past values of the variables allows us to condition on more than two variables, which is not possible in the context of i.i.d. causal pairs.

We shall also notice that if the framework is the one of a causal time-series pair in which only one direction of causal influence is admitted: either $x_t \to y_s$ (for one or more values of s such that $s \geq t$) or $y_t \to x_s$ ($s \geq t$) and one is only interested in the "summary graph" [58], i.e. in ascertaining whether x causes y or y causes x at any time unit, then the problem can be solved in a relatively easy fashion in many settings. Using a simple regression analysis, it will be sufficient to regress x_t on lagged values of itself and of the other variable, as well as to regress y_t on lagged values of itself and of the other variable. Since all the covariates in the two regressions are pre-determined there are no endogeneity problems here and the error terms will be independent of the regressors. Therefore, by simple testing

the hypothesis of non-zero statistical influence of one lagged variable (e.g. x_{t-1}) on another (e.g. y_t) and the hypothesis of a zero statistical influence on the symmetric regression (e.g. of y_{t-1} on x_t), we will be able to detect a genuine causal influence (at some unknown time unit) from one variable to another (e.g. from x to y). This framework is identical to the vector autoregressive framework that we will discuss below and also related to an interpretation of Granger non-causality test that we will also discuss below. Notice, however, than in many fields like economics the assumption of causality running in only one direction between time series, without the possibility of a feedback at a different time unit, is a toy example, with very poor empirical applicability. This is why our discussion framework will be larger, including the possibility of structures like $y_{t-1} \rightarrow x_t \rightarrow y_t$.

In reviewing different methods we distinguish between methods that filter the series through a vector autoregressive model (Sect. 5.2.1) and methods that apply causal search directly to time series data (Sect. 5.3).

5.2 Vector-Autoregressive Framework

5.2.1 The VAR Model

One of the most popular approaches to identify dynamic causal effects in time series econometrics is structural vector autoregressive (VAR) analysis. Structural VAR analysis is based on the assumption that the statistical properties of a data generating process can be well approximated by a reduced-form VAR model.

Let us consider a vector Y_t of k time series variables. For example, $Y_t = (x_t, y_t)'$, in which case $k = 2$. We assume that Y_t follows a stochastic process that can be well approximated by a linear VAR process of the form

$$Y_t = \mu + A_1 Y_{t-1} + \cdots + A_p Y_{t-p} + u_t, \tag{5.1}$$

where μ is a $k \times 1$ vector of constants, A_i $(i = 1, \ldots, p)$ is a $k \times k$ matrix and u_t is a $k \times 1$ vector of white noise, whose elements are referred to as *reduced-form residuals*. Each element of u_t is in turn assumed to be a linear combination of latent structural shocks, $\epsilon_{1t}, \epsilon_{2t}, \ldots$, which are the sources of variation of the system. In macroeconomics these shocks have special meaning such as, for example, the productivity shock, the monetary policy shock, the fiscal policy shock, etc. It is standard in the VAR literature to assume that the number of shocks is equal to the number of measured variables. Another usual assumptions is that $\epsilon_{1t}, \ldots, \epsilon_{kt}$ are mutually independent, although orthogonality is sufficient in many applications. Thus we have:

$$u_t = B\varepsilon_t, \tag{5.2}$$

where B is a $k \times k$ invertible matrix (the *impact* or *mixing* matrix) and $\varepsilon_t = (\epsilon_{1t}, \ldots, \epsilon_{kt})'$ is a vector of independent shocks. Let W be B^{-1}. By pre-multiplying Eq. (5.1) by W we get the structural VAR form:

$$WY_t = \mu' + \Gamma_1 Y_{t-1} + \cdots + \Gamma_p Y_{t-p} + \varepsilon_t, \tag{5.3}$$

where $\mu' = W\mu$ and $\Gamma_i = WA_i$ for $i = 1, \ldots, p$. From Eq. (5.1) it is evident that the matrix W incorporates information about the contemporaneous causal structure, while the matrices Γ_i's incorporate information about the lagged causal structure. Since Sims [62], econometricians have focused their attention on the identification of the effect of ε_t on Y_t over time. These are called *impulse response functions*, and we will be discussed in a subsequent Sect. 5.2.4.

Since Eq. (5.3) cannot be directly estimated because of endogeneity problem, the idea of VAR analysis is to follow a two-step procedure: first Eq. (5.1) is estimated through standard regression methods. From this stage one obtains an estimate of the reduced-form residuals u_t. Second, the parameters of Eq. (5.3) (in particular the coefficients entering in W and Γ_i) can be recovered by analyzing the relationships among the elements of u_t, which, under some conditions, may allow identifying the matrix B entering in Eq. (5.2). Notice that, having estimated (5.1), knowing B is sufficient for identifying (5.3).

For example, Swanson and Granger [68], Bessler and Lee [3], Demiralp and Hoover [11], Moneta [52] propose a two-step identification method, consisting in first estimating the reduced-form VAR residuals, and then applying to the estimated u_t (which should share characteristics of i.i.d. data) conditional independence tests, in the spirit of a causal search based on graphical causal models [64]. This allows them to find out which entries of B are zero.

For $k = 2$, as is the case of causal pairs, independence tests between u_{1t} and u_{2t} can only discriminate between the presence and the absence of a causal link between the contemporaneous variables, but are not of any help in finding causal directions. In other words, they find zero entries in B only in the case when u_{1t} and u_{2t} are mutually independent (corresponding to the absence of contemporaneous causal relations).

5.2.2 ICA-Based Identification

An alternative method to identify B in the same two-step framework is to apply Independent Component Analysis (ICA) to the estimated reduced-form residuals u_t. Since, as shown in (5.2), $u_t = B\varepsilon_t$, it is possible to apply ICA to recover the coefficients that linearly mix the elements of ε_t to produce u_t [9, 37, 39]. ICA has been applied to a VAR setting by Hyvärinen et al. [40], Moneta et al. [53], Gouriéroux et al. [22], among others.

ICA is based on a theorem, see [9, Th. 11], [15, Th. 3], [22, p.112], according to which if B is invertible, and if the components of ε_t $(\epsilon_{1t}, \ldots, \epsilon_{kt})$ are independent,

with at most one Gaussian distribution, then the matrix B is identifiable up to a post multiplication by DP, where P is a permutation matrix and D a diagonal matrix with non zero diagonal elements.

There are many ICA approaches to estimate the mixing matrix B (cfr. [39] for an overview), most popular of which are the fastICA algorithm [38], which is based on minimization of mutual information and maximization of negentropy, the JADE algorithm [5], which maximizes a measure of non-Gaussianity based on the fourth moments, and the product density ICA algorithm [28], which is based on maximum likelihood principle. Alternative approaches have been also recently proposed in econometrics, e.g. the distance covariance approach by Matteson and Tsay [51], the Cramer-von-Mises distance approach by Herwartz [30], the maximum likelihood approach by Lanne et al. [48], and the pseudo ML approach by Gouriéroux et al. [22].

Assuming that B is invertible implies that each observed variable u_{it} is affected by at least one shock ϵ_{it} and that each ϵ_{it} influences at least one variable. In other words, there is always a column-permutation of the mixing matrix $\tilde{B}$ output of ICA such that all the elements in the main diagonal are significantly different from zero. This assumption is in tune with the standard VAR framework.

In the case of causal pairs ($k = 2$), with matrix B of dimension 2×2, it is therefore very useful to test which entries in B are significantly close to zero and check their row position. The significance test can be done with a bootstrap procedure, by performing a nonparametric quantile test in order to decide whether 0 is an outlier, as proposed by Lacerda et al. [47]. Alternatively, one can test a zero restriction in B by exploiting the asymptotic distribution of the pseudo ML estimator of B, as proposed by Gouriéroux et al. [22].

Let us continue to assume that $Y_t = (x_t, y_t)'$. On the basis of tests on zero restrictions in B, one can distinguish among four different cases: (1) If there is only one zero entry in B and this lies in the first row, this means that the first element of u_t, which we call u_{xt}, is affected only by one shock, while the second element of u_t, which we call u_{yt}, is affected by both shocks. This means that x_t causes y_t. (2) Symmetrically, if the only zero entry of B lies in the second row, y_t causes x_t. (3) If there are two zero entries in B, which, by construction, must lie either in its main or anti-diagonal, then x_t and y_t are not (contemporaneously) causally related. (4) If there are no zero entries in B, some other structures are possibilities: there could be a feedback loop between x_t and y_t, or a latent variable z_t affecting both x_t and y_t, possibly also including causal relationships between x_t and y_t.

If there is a latent variable z_t, this means that the shocks affecting the system are potentially three, while the observed variables are still two. Attempting to identify the structural model would bring us outside the VAR framework. It is worth noting, however, that the ICA framework has been extended to the cases where the number of sources is greater than the number of mixtures (*overcomplete ICA*) (see [39, ch.16]). The identification of the rectangular mixing matrix potentially allows distinguishing between the case of feedback loop between x_t and y_t (two shocks affecting the system) and the case of a latent variable (three shocks affecting the system with at least one idiosyncratic shock).

If it is known that, underlying the structural model, there is a recursive contemporaneous structure, that is either x_t causes y_t or y_t causes x_t (equivalently, there is a permutation of the matrices B and W that make them lower triangular), then, a valid and efficient alternative to the test of zero-coefficient suggested above, is performing a LiNGAM (short for Linear Non-Gaussian Acyclic Model) analysis, as proposed by Shimizu et al. [61]. LiNGAM is an algorithm that incorporates ICA in the first step, and then search for the right row-permutation of the unmixing matrix W that yields a lower triangular matrix. Lacerda et al. [47] propose an extension of this algorithm to the cyclic case (in which feedback loops are allowed), called LiNG. Hoyer et al. [34] propose an extension of basic LiNGAM to the case in which latent common cause are allowed, called LvLiNGAM.

5.2.3 Nonlinear Framework

The standard VAR framework, as proposed in the econometric literature, is a linear model. In economics and in many other fields, however, there is no compelling substantive reason why a variable should depend *only linearly* on current values of other variables, on past values of itself and of other variables. Thus, a class of nonlinear structural VAR models has been proposed (see [44, ch. 18]) that allows nonlinear dependence among measured time-series but with an additive white noise error terms. In this case, we can apply a two-step identification procedure similar to linear case: in a first step one estimates a reduced-form nonlinear VAR model, and in a second step one extracts from the estimated additive errors information in order to recover the structural VAR model. A general nonlinear VAR model with additive errors can be written as:

$$Y_t = F_t(Y_{t-1}, \ldots, Y_{t-p}) + u_t, \tag{5.4}$$

where the nonlinear function $F_t(\cdot)$ may depend on t. Most nonlinear VAR models considered in the econometric literature deal with time-varying coefficients (see e.g. [59]) which are able to capture very general nonlinear dynamics, while keeping linear the mixing structure between reduced-form and structural residuals.

We do not review here this literature (see [27, 43], and references therein). Rather, we point out a method to identify the contemporaneous causal direction that exploits the nonlinear dependence among the variables and is based on two assumptions: (i) there is a contemporaneous, nonlinear causal relationship between x_t and y_t in only one direction (either $x_t \longrightarrow y_t$ or $y_t \longrightarrow x_t$), (ii) the structural form model can be written as $Y_t = F(Y_{t-1}, \ldots, Y_{t-p}) + G(Y_t) + \varepsilon_t$, where $F(\cdot)$ and $G(\cdot)$ are two linear functions with $\epsilon_{1t} \perp\!\!\!\perp \epsilon_{2t}$.

The method follows a two-step procedure, as is typical of a VAR-based approach. In the first step the lagged effects are filtered out through nonlinear or nonparametric estimates of the regressions $x_t = f_1(x_{t-1}, \ldots, x_{t-p}, y_{t-1}, \ldots, y_{t-p}) + u_{1t}$ and $y_t = f_2(x_{t-1}, \ldots, x_{t-p}, y_{t-1}, \ldots, y_{t-p}) + u_{2t}$, in order to obtain estimates of u_{1t}

and u_{2t}. In the second step one the contemporaneous causal direction is detected through a nonlinear additive noise model (see [35, 58]). Indeed we will have that if the contemporaneous causal relation is $x_t \rightarrow y_t$

$$u_{2t} = f_y(u_{1t}) + N_y \tag{5.5}$$

where N_y is an unobserved noise term and $N_t^y \perp\!\!\!\perp u_{1t}$. Likewise, if the contemporaneous causal relation is $y_t \rightarrow x_t$

$$u_{1t} = f_x(u_{2t}) + N_x, \tag{5.6}$$

where N_x is an unobserved noise term and $N_x \perp\!\!\!\perp u_{2t}$.

Thus, once u_{1t} and u_{2t} are estimated through a nonlinear or nonparametric VAR model, one regress them on each other, using a nonparametric estimator, and obtains estimated of N_x and N_y. If, on the basis of a nonparametric independence test (see e.g. [25]), the independence between N_y and u_{1t} is not rejected, while the independence between N_x and u_{2t} is rejected, one infer $x_t \rightarrow y_t$. If, the independence between N_x and u_{2t} is not rejected, while the independence between N_y and u_{1t} is rejected, one infer $y_t \rightarrow x_t$.

5.2.4 Impulse Response Functions

Having identified the mixing matrix B and the structural shocks ε_t, econometricians are mostly interested in the responses over time of each element of $Y_t = (x_t, y_t)'$ to a one-time impulse in each element of $\varepsilon_t = (\epsilon_{1t}, \epsilon_{2t})'$. These impulse response functions are defined [44, p. 110] as:

$$\frac{\partial Y_{t+i}}{\partial \varepsilon_t'} = \Theta_i \qquad i = 0, 1, 2, \ldots, H, \tag{5.7}$$

where, in the case of two variables, Θ_i is a 2×2 matrix, whose four elements are: $\frac{\partial x_{t+i}}{\partial \epsilon_{1t}}$, $\frac{\partial y_{t+i}}{\partial \epsilon_{1t}}$ (first column), $\frac{\partial x_{t+i}}{\partial \epsilon_{2t}}$, $\frac{\partial y_{t+i}}{\partial \epsilon_{2t}}$ (second column).

Consider, for simplicity, a linear VAR model with one lag (p=1) and no intercept:

$$Y_t = A_1 Y_{t-1} + u_t. \tag{5.8}$$

By recursive substitution it can be written:

$$Y_{t+i} = A_1^{i+1} Y_{t-1} + \sum_{j=0}^{i} A_1^j u_{t+i-j}. \tag{5.9}$$

The responses of Y_t to reduced-form errors (also referred to as forecast errors) i periods ago[2] are then captured by the matrix $\Phi_i = A_1^i$. If Y_t is a stable process (all eigenvalues of A have modulus less than 1), i.e. each element of Y_t is covariance stationary, Eq. (5.8) can be equivalently expressed according to the moving average (MA) representation (Wold decomposition):

$$Y_t = \sum_{i=0}^{\infty} \Phi_i u_{t-i}, \tag{5.10}$$

where Φ_i is calculated as above (for the one-lag case), with $\Phi_0 = I$. From Eqs. (5.10), (5.2) and (5.7) it follows

$$Y_t = \sum_{i=0}^{\infty} \Phi_i B B^{-1} u_{t-i} = \sum_{i=0}^{\infty} \Phi_i B \varepsilon_{t-i} = \sum_{i=0}^{\infty} \Theta_i \varepsilon_{t-i}. \tag{5.11}$$

If the VAR is not stable, the infinite Wold representation is not allowed, but the same approach to calculate Φ_i and Θ_i will work, because Eq. (5.9) does not depend on stationarity. In case of unstable process, the impulse response functions will not be tied to the MA representation and will not converge to zero for $i \to \infty$. In particular if Δx_t is stationary the impulse response function to Δx_t will converge to a finite number.

This framework to calculate impulse response functions can be easily extended to the case of more lags using a "companion matrix" representation (see [44, p. 25]) and is not substantively affected by the presence of a constant in (5.8). However, it cannot be applied to nonlinear VAR models, due to its reliance on Eq. (5.9).

Thus structural impulse responses in a nonlinear setting are defined in an alternative manner, using the concept of conditional expectation [44, 45, p. 615]. Denoting by Ω_{t-1} the information set available at date $t-1$ and by δ the magnitude of the impulse of which one wants to study the response (e.g. δ = standard deviation (ϵ_{1t})), the structural response of x_{t+i} to the structural shock ϵ_{1t} is defined as

$$I_x(i, \delta, \Omega_{t-1}) = \mathbb{E}(x_{t+i} | \epsilon_{1t} = \delta, \Omega_{t-1}) - \mathbb{E}(x_{t+i} | \Omega_{t-1}) \quad i = 0, \ldots, H. \tag{5.12}$$

Having estimated a nonlinear reduced form VAR model (5.4) and having recovered the structural shocks (for example on the basis of additive noise model framework, see end of Sect. 5.2.3), one can evaluate (5.12) using a Monte Carlo procedure [44, pp. 615–616]. In this procedure, one simulates two time paths: in a first path the shock of interest is set at time 0 to a particular value δ and the subsequent realizations of the variables of interest are estimated; in a second time path the value of the shock of interest is drawn from an empirically estimated marginal distribution.

[2] Or, equivalently, the responses of Y_{t+i} to forecast errors at time t.

Thus, Eq. (5.12) is estimated by subtracting the average outcome of the second path from the first.

5.2.5 *Granger Causality in a VAR Framework*

VAR models have also been used for a type of causal analysis that does not involve the identification of a structural model like Eq. (5.3). This approach is based on a notion of causal relationship proposed by Granger [23, 24], which is referred to as *Granger causality*. Granger's general definition of causality relies on two general principles: (i) the effect does not precede its cause in time; (ii) the causal time series contains unique information about the series being caused that is not available otherwise (see [13]). A corollary of these principles is that x_t Granger causes y_t if x_t is helpful for predicting future values of y_t. Incidentally, these tenets share profound similarities with probabilistic theories of causality proposed in the philosophy of science literature [20, 21, 67] (see also [65]).

Although the definition of Granger causality is more general (see Sect. 5.3.1 below), several empirical studies and statistical software make it operational in a linear VAR framework. Consider a bivariate VAR with p lags:

$$\begin{pmatrix} x_t \\ y_t \end{pmatrix} = \sum_{i=1}^{p} \begin{bmatrix} a_{11,i} & a_{12,i} \\ a_{21,i} & a_{22,i} \end{bmatrix} \begin{pmatrix} x_{t-i} \\ y_{t-i} \end{pmatrix} + u_t. \tag{5.13}$$

In this framework x_t is said to be non-Granger-causal for y_t if and only if $a_{21,i} = 0$ for $i = 1, \ldots, p$ [49, p. 154]. This amounts to say that the information set available until time $t-1$ to forecast y_t comprises only x_{t-1} (with more lagged terms) and y_{t-1} (with more lagged terms), and one wants to check whether excluding or not lagged x_t from the information set makes a difference in predicting y_t. The zero restrictions can be tested with standard Wald χ^2- or F-tests, which have standard asymptotic properties if the series are stationary [49, p. 154].

A main limitation of this framework is that lagged x_t may make a difference in forecasting y_t (so that to infer that x_t Granger-causes y_t) because it contains information that is not contained in the information set comprising lagged y_t and lagged x_t, but it is always possible that if one considered a larger set of information, for example one containing lagged values of a series z_t, x_t would not bring a further contribution for the prediction of y_t. If z_t is a common cause of both x_t and y_t one would have wrongly inferred that x_t causes y_t. Thus, although scholars have worked in this direction, introducing concepts such as conditional independencies and higher-order interactions, causal sufficiency is still a fundamental tenet of this approach; this is particularly true, if the focus on causality goes beyond what sometimes is referred to as "predictive causality."

Granger-causality in causal pairs is a very powerful method in a setting in which, as mentioned in the introduction, the presence of a causal relationship between the

two variables, until some lag $p \geq 0$, is known, but is unknown whether it is x_{t-p} that causes y_t or it is y_{t-p} that causes x_t.

Suppose, for example, that is known that x_{t-p} causes y_t with $p = 0$ or 1 and there are no causal relationships from y_t to x_t at any lag. Then in all the 3 admitted cases in which x_t can cause y_t ((i) $x_{t-1} \rightarrow y_t$; (ii) $x_t \rightarrow y_t$; (iii) i $\cup$ ii), the coefficient $a_{12,1}$, estimated by regressing equation (5.13), is expected to be not significantly different from zero, while the other coefficients of the same matrix will be non-zero. Symmetrically, if y_{t-p} causes x_t with $p = 0$, 1 (and no feedback from x_t to y_t at any lag), then the only coefficient of the same matrix, obtained by regressing the same equation, which is expected to be zero is $a_{21,1}$.

Standard Granger-causality in a VAR framework neglects, by choice, the contemporaneous causal link, which is considered by the structural VAR approach. Geweke [19], however, proposes an extension of the Granger-causality concept to detect linear contemporaneous feedback between two time-series, x_t and y_t.

Jacobs et al. [41] and Hoover [32, pp. 151–152] present examples of bivariate, one-lag structural VAR models in which $x_{t-1} \rightarrow y_t$; $y_{t-1} \rightarrow x_t$; $x_t \rightarrow y_t$, but, for particular configurations of the parameters, in the reduced form VAR the coefficient corresponding to the influence of y_{t-1} on x_t ($a_{11,1}$ in Eq. (5.13)) is zero. One could exclude these types of parameters configuration as "measure-zero." This assumption would be similar to the faithfulness assumption in the graphical causal model literature [64], where configurations of parameters that yield statistical independence actually corresponding to causal dependence are ruled out. Hoover [32] argues further that specific configurations of parameters for which Granger non-causality does not match structural non-causality may correspond to theoretical economic models and thus cannot be easily dismissed.

5.3 Direct Causal Search

In this section we discuss methods for causal pairs search that are applied directly to time series data, without filtering them through a vector autoregressive model. Skipping VAR estimation has the clear advantage of not being tied to the imposition of a functional form (e.g. linear VAR), when estimating the relationship between current and lagged values of the variables of interest. On the other hand, direct causal search deals directly with autocorrelated data.

5.3.1 Granger Causality

As mentioned above (Sect. 5.2.5), the central notion in Granger causality is "incremental predictability" [32, p.150]: if a time series y_{t+1} is better predicted by the set of all information available up to time t than by the same information set less the

series x_t, then x_t *Granger-causes* y_{t+1}. The general definition given by Granger [24, p. 49] is that x_t is said to cause y_{t+1} if

$$P(y_{t+1} \in A|\Omega_t) \neq P(y_{t+1} \in A|\Omega_t - x_t), \tag{5.14}$$

where Ω_t is all the knowledge in the universe available at time t, $\Omega_t - x_t$ is the same information set except the values taken by a x_t up to time t, where $x_t \in \Omega_t$, and A is any set of values that y_{t+1} can take. We can also write that x_t does *not* Granger-causes y_{t+1} if [13]

$$y_{t+1} \perp\!\!\!\perp \Omega_t|\Omega_t - x_t, \tag{5.15}$$

otherwise x_t is said to Granger-cause y_{t+1}. As Granger [24] admits, this general definition of causality is not operational, i.e. it cannot be implemented with actual data. A practical solution is to consider Ω_t as incorporating only current and past values (until certain lags) of x_t, y_t and of a set of observed variables Z_t. Thus we have that x_t is Granger-noncausal for y_{t+1} if [16, 66]

$$y_{t+1} \perp\!\!\!\perp \{x_t, \ldots, x_{t-q}\}|\{y_t, \ldots, y_{t-p}, Z_t, \ldots, Z_{t-r}\}, \tag{5.16}$$

given lags p, q, r, where by $\{x_t, \ldots, x_{t-q}\}$ we denote the σ-field generated by the vector of random variables $(x_t, \ldots, x_{t-q})$, and similarly for $\{y_t, \ldots\}$. The σ-field generated by a random variable is the set of events that may be described in terms of that random variable [16, p. 588]. Let us suppose that the background knowledge available at time t comprises only two time series: x_t and y_t. Then, given lags p and q, x_{t-1} does not Granger causes y_t if

$$y_t \perp\!\!\!\perp \{x_{t-1}, \ldots, x_{t-q}\}|\{y_{t-1}, \ldots, y_{t-p}\}. \tag{5.17}$$

Assuming that x_t and y_t are stationary and ergodic, many studies have proposed nonparametric tests of (5.17), without assuming a linear structure (which could be treated in a linear VAR framework) (see [1, 2, 4, 12, 31, 66, 70]). In case of $p, q = 1$ the proposed tests have high performance, which tends to decline for high p and q for data with limited sample size [6]. The assumption of Ω_t as comprising only two time series is, of course, a strong assumption in empirical contexts where causal sufficiency may fail.

5.3.2 Graphical Models for Time Series

Since Granger-causality faces fundamental hurdles in case of unmeasured causal variables, one possible solution is to rely on causal inference procedures that are designed to perform well in presence of latent variables. One algorithm that is asymptotically correct in the presence of latent variables is the Fast Causal Inference

(FCI) algorithm proposed by Spirtes et al. [64]. This method belongs to the more general approach of graphical causal models based on conditional independence tests, also known as "constraint-based causal search" (see [63]). We have mentioned this approach in Sect. 5.2.1, noticing that it was of little use when applied to pairs of estimated VAR reduced-form residuals. This approach, however, has larger applicability when applied directly to pairs of time series data (not filtered by a VAR model), because it can exploit the possibility of conditioning both on lagged and contemporaneous variables. An interesting method, in this setting, is the adaptation of the FCI algorithm that Entner and Hoyer [14] propose for time series.

In case of causal sufficiency (and no feedback loops), constraint-based causal search moves from the assumption that the data generating process can be described by a directed acyclic graph (DAG) and a joint distribution $P(\mathbf{X})$, where $\mathbf{X} = (X_1, \ldots, X_n)$ is the set of observable variables represented by the set V of n vertices of the DAG. Causal inference is based on two assumptions: *Markov* and *faithfulness* condition. Markov condition states that if vertices i and j of a DAG $\mathscr{G}$ given some subset $W \subseteq V \backslash \{i, j\}$ are *d-separated* (a graphical criterion defined by Pearl [54]), then we have $X_i \perp\!\!\!\perp X_j | \{X_w : w \in W\}$. Faithfulness condition states that all (conditional and unconditional) independence relations in $P(\mathbf{X})$ are entailed by the Markov condition. In this setting, the PC algorithm [64], on the base of these assumptions, starts from a complete graph (all vertices connected by undirected edges) over all variables, and performs a series of independence tests that allows the removal of edges between pairs of variables that are independent conditionally on any set of variables (included the empty set). Then it makes use of some rules which allow us to orient edges among triple of vertices, and in particular to distinguish between collider ($\cdot \rightarrow \cdot \leftarrow \cdot$) structure and fork/chain structures ($\cdot \leftarrow \cdot \rightarrow \cdot$, or $\cdot \leftarrow \cdot \leftarrow \cdot$, or $\cdot \rightarrow \cdot \rightarrow \cdot$). This is also done on the basis of conditional independence tests and the two conditions above. The outcome of the algorithm is a set of DAGs that share the same (conditional) independence relations, i.e. a class of *Markov equivalent* DAGs.

Relaxing the assumption of causal sufficiency, the FCI algorithm [64] moves also from the assumption that the process underlying the data can be described by a DAG, but this DAG may contain vertices that correspond to latent variables. Richardson and Spirtes [60] (see also [8]) introduced a new class of graphs whose vertices are observed variables, but in which the causal relationships may involve latent variables. These graphs, in which a latent cause Z affecting the observed variables X and Y is represented by $X \leftrightarrow Y$, are called *maximal ancestral graphs* (MAGs). The idea is that any DAG whose vertices include latent variables can be transformed in a unique MAG whose vertices comprise only observed variables. Moreover, MAGs encode conditional independence relations among the observed variables through m-separation, a generalization of d-separation [8, 60]. A MAG is a graph $\mathscr{M}$ with the following properties: (i) $\mathscr{M}$ is a *mixed* graph (it contains not only directed ($\rightarrow$), but also undirected ($-$) and bi-directed ($\leftrightarrow$) edges); (ii) $\mathscr{M}$ is an *ancestral* graph (there is no vertex i which is a ancestor of any of its parents nor any

of its spouse[3]); (iii) for every pair of variables $\langle X_i, X_j \rangle$ there is an edge between i and j in $\mathscr{M}$ if and only if there does not exist a set of vertices $W \subseteq V \backslash \{i, j\}$ in $\mathscr{M}$ such that $X_i \perp\!\!\!\perp X_j | \{X_w : w \in W\}$ [14, 60].

Similarly to PC algorithm, the output of the FCI algorithm is a class of MAGs that entail the same set of conditional independence relationships. This class of MAGs is represented by a *partial ancestral graph* (PAG), which is a graph which have a third edge mark, besides arrowtail ($-$) and arrowhead ($>$), namely a circle ($\circ$). Excluding feedback loops or selection bias (hence undirected edges), a PAG can only incorporate these types of edges: $\rightarrow$, $\leftrightarrow$, $\circ\!\!\rightarrow$, and $\circ\!\!-\!\!\circ$. If $X_i \leftrightarrow X_j$ then neither variable is ancestor of the other and there is a latent variable between X_i and X_j. The circle ($\circ$) denotes the case where it is undecided whether in the underlying data generating process there is an arrowtail or an arrowhead next to the vertex where the circle appear. This means that the PAG contains a MAG with ($-$) and a MAG with ($>$) at that location. Like the PC algorithm, the FCI in a first step removes edges from a complete graph on the base of conditional independence tests, and in a second step it orients edges so that the inferred causal structures are in tune with the Markov and faithfulness assumptions (all the conditional independence relations must be derived from m-separation).

Entner and Hoyer [14] adapt the FCI in a time series framework, which they call tsFCI. Suppose the observed time series variables are $\{x_t\} = x_1, \ldots, x_T$ and $\{y_t\} = y_1, \ldots, y_T$. The algorithm starts from a complete graph on a time window of the time series, i.e. the set of vertices are $x_t, x_{t-1}, \ldots, x_{t-p}, y_t, y_{t-1}, \ldots y_{t-p}$. It then remove edges from this complete graph, as in a standard FCI algorithm, on the basis of conditional independence test, but with the addition that if the contemporaneous edge is eliminated, this will be eliminated at all time units $(t, t-1, \ldots, t-p)$. If a lagged edge with lag l is eliminated (for example from x_{t-l} to y_t), this is eliminated at all time units (for example from x_{t-l-1} to y_{t-1}). Orientation makes use not only of the orientation rules of the standard FCI algorithm, but also makes use of the "arrow of time": if there is an undirected edge between two lagged variable, it will be put an arrowtail at the variable coming before and an arrowhead at the variable coming after. Moreover, if an edge is oriented contemporaneously at time t, this will be oriented in the same manner for all time units $(t, t-1, \ldots)$. If a lagged edge with lag l is oriented (for example $x_{t-l} \rightarrow y_t$), this is oriented in the same manner for all time units (for example $x_{t-l-1} \rightarrow y_{t-1}$).

Thus, exploiting the assumption that an effect cannot precede a cause and the assumption of repetition of causal structures over time (time invariance), one can reach a more detailed description of the data generating process than the one that would be provided by a standard application of constraint based algorithm. However, since these methods ultimately rely on conditional independence tests

[3]A vertex i is an ancestor of j if there is a sequence of directed edges ($\rightarrow$) between i and j. A vertex i is a parent of j if $i \rightarrow j$. A vertex i is a spouse of j (and j a spouse of i) if there is a bi-directed edge between i and j.

is crucial that they are designed taking into account the specificity of testing self-dependence in a time series context (see [46]).

5.3.3 *Additive Noise Models*

We consider in this subsection the problem of distinguishing among different causal structures over the time-series pair $\{x_t, y_t\}$, using a specific class of structural equation models. We assume that: (i) there are no latent common causes between x_t and y_t (at any lag); (ii) no contemporaneous causal feedback loops (i.e. either $x_t \to y_t$ or $x_t \leftarrow y_t$, but it is possible that $x_{t-s} \to y_t \to x_{t+h}$, for $s \geq 0$, $h \geq 1$); (iii) each variable x_t and y_t causally depends on its own past (respectively $x_{t-1}, \ldots$ and $y_{t-1}, \ldots$) until a lag p; (iv) both contemporaneous and lagged causal structures recur over time: if $x_{t-i} \to y_t$ then $x_{t-i-s} \to y_{t-s}$, for $i \geq 0$, $s \geq 1$. To simplify the illustration, we also assume here that (v) $\mathrm{p} = 1$. In Fig. 5.3 we show the 12 directed acyclic graphs (DAGs) corresponding to all the possible causal structures related to the data generating process (represented as unit graphs) under these assumptions. We also assume that (vi) x_t and y_t are stationary and ergodic processes. We also assume that the data generating process can be formalized as a specific type of structural equation model (or *functional equation model*, see [55]), namely as an *additive noise model* [35, 56, 57], where

$$x_t = f_x(\mathbf{PA}^x) + N_t^x \tag{5.18}$$

and

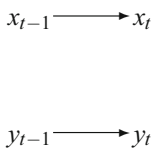
$$y_t = f_y(\mathbf{PA}^y) + N_t^y, \tag{5.19}$$

where $\mathbf{PA}^x$ are the graphical parents of x_t (and $\mathbf{PA}^y$ of y_t) in the DAG representing the data generating process, and N_t^x and N_t^y are independent white noise processes. We assume (vii) $N_t^x \perp\!\!\!\perp \mathbf{PA}^x$, $N_t^y \perp\!\!\!\perp \mathbf{PA}^y$, and $N_t^x \perp\!\!\!\perp N_t^y$; (viii) $f_x(\cdot)$ and $f_y(\cdot)$ are either nonlinear functions or linear but with the additional assumption that N_t^x and N_t^y have non-Gaussian distribution.[4]

In Fig. 5.3, below each DAG it is shown the set of corresponding structural equations and the set of implied (conditional or unconditional) independence relationships. Hoyer et al. [35] (see also Sect. 5.2.3) proposes a procedure to check if a DAG corresponding to a nonlinear additive noise model is consistent with the data: first one constructs a nonlinear regression of each variable on its parents, then one tests whether the estimated residuals are independent of the covariates and among

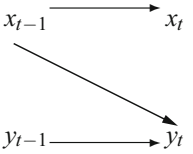
[4]Specific nonlinear functions $f_x(\cdot)$ and distributions of the noise terms have also to be excluded. A precise specification can be found in Peters et al. [57, Proposition 23] and Zhang and Hyvärinen [69].

DAG 1



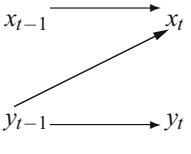
$x_t = f(x_{t-1}) + N_t^x$
$y_t = f(y_{t-1}) + N_t^y$
$N_t^x \perp\!\!\!\perp x_{t-1}$
$N_t^y \perp\!\!\!\perp y_{t-1}$
$N_t^x \perp\!\!\!\perp N_s^y \quad \forall t,s$

DAG 2



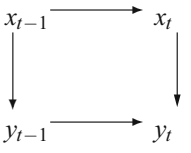
$x_t = f(x_{t-1}) + N_t^x$
$y_t = f(x_{t-1}, y_{t-1}) + N_t^y$
$N_t^x \perp\!\!\!\perp x_{t-1}$
$N_t^y \perp\!\!\!\perp x_{t-1}; N_t^y \perp\!\!\!\perp y_{t-1}$
$N_t^x \perp\!\!\!\perp N_s^y \quad \forall t,s$

DAG 3



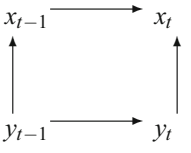
$x_t = f(x_{t-1}, y_{t-1}) + N_t^x$
$y_t = f(y_{t-1}) + N_t^y$
$N_t^x \perp\!\!\!\perp x_{t-1}; N_t^x \perp\!\!\!\perp y_{t-1}$
$N_t^y \perp\!\!\!\perp y_{t-1}$
$N_t^x \perp\!\!\!\perp N_s^y \quad \forall t,s$

DAG 4



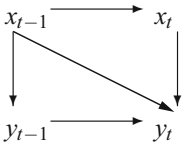
$x_t = f(x_{t-1}) + N_t^x$
$y_t = f(x_t, y_{t-1}) + N_t^y$
$N_t^x \perp\!\!\!\perp x_{t-1}$
$N_t^y \perp\!\!\!\perp x_t$
$N_t^y \perp\!\!\!\perp y_{t-1}$
$N_t^x \perp\!\!\!\perp N_s^y \quad \forall t,s$

DAG 5



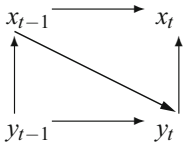
$x_t = f(y_t, x_{t-1}) + N_t^x$
$y_t = f(y_{t-1}) + N_t^y$
$N_t^x \perp\!\!\!\perp x_{t-1}, N_t^x \perp\!\!\!\perp y_t$
$N_t^y \perp\!\!\!\perp y_{t-1}$
$N_t^x \perp\!\!\!\perp N_s^y \quad \forall t,s$

DAG 6



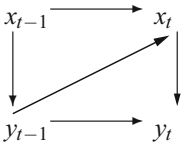
$x_t = f(x_{t-1}) + N_t^x$
$y_t = f(x_t, x_{t-1}, y_{t-1}) + N_t^y$
$N_t^x \perp\!\!\!\perp x_{t-1}, N_t^y \perp\!\!\!\perp x_t$
$N_t^y \perp\!\!\!\perp x_{t-1}, N_t^y \perp\!\!\!\perp y_{t-1}$
$N_t^x \perp\!\!\!\perp N_s^y \quad \forall t,s$

DAG 7



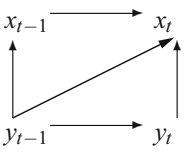
$x_t = f(y_t, x_{t-1}) + N_t^x$
$y_t = f(x_{t-1}, y_{t-1}) + N_t^y$
$N_t^x \perp\!\!\!\perp x_{t-1}$
$N_t^x \perp\!\!\!\perp y_t$
$N_t^y \perp\!\!\!\perp y_{t-1}, N_t^y \perp\!\!\!\perp x_{t-1}$
$N_t^x \perp\!\!\!\perp N_s^y \quad \forall t,s$

DAG 8



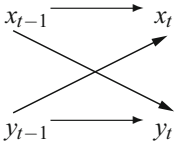
$x_t = f(y_{t-1}, x_{t-1}) + N_t^x$
$y_t = f(y_{t-1}, x_t) + N_t^y$
$N_t^x \perp\!\!\!\perp x_{t-1}, N_t^x \perp\!\!\!\perp y_{t-1}$
$N_t^y \perp\!\!\!\perp y_{t-1}$
$N_t^y \perp\!\!\!\perp x_t$
$N_t^x \perp\!\!\!\perp N_s^y \quad \forall t,s$

DAG 9



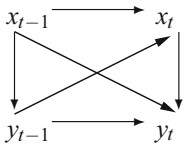
$x_t = f(y_t, x_{t-1}, y_{t-1}) + N_t^x$
$y_t = f(y_{t-1}) + N_t^y$
$N_t^x \perp\!\!\!\perp y_t,$
$N_t^x \perp\!\!\!\perp x_{t-1}$
$N_t^x \perp\!\!\!\perp y_{t-1}$
$N_t^y \perp\!\!\!\perp y_{t-1}$
$N_t^x \perp\!\!\!\perp N_s^y \quad \forall t,s$

DAG 10



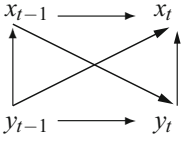
$x_t = f(x_{t-1}, y_{t-1}) + N_t^x$
$y_t = f(x_{t-1}, y_{t-1}) + N_t^y$
$N_t^x \perp\!\!\!\perp x_{t-1}$
$N_t^x \perp\!\!\!\perp y_{t-1}$
$N_t^y \perp\!\!\!\perp x_{t-1}$
$N_t^y \perp\!\!\!\perp y_{t-1}$
$N_t^x \perp\!\!\!\perp N_s^y \quad \forall t,s$

DAG 11



$x_t = f(x_{t-1}, y_{t-1}) + N_t^x$
$y_t = f(x_t, x_{t-1}, y_{t-1}) + N_t^y$
$N_t^x \perp\!\!\!\perp x_{t-1},$
$N_t^x \perp\!\!\!\perp y_{t-1}$
$N_t^y \perp\!\!\!\perp x_t,$
$N_t^y \perp\!\!\!\perp x_{t-1}$
$N_t^y \perp\!\!\!\perp y_{t-1}$
$N_t^x \perp\!\!\!\perp N_s^y \quad \forall t,s$

DAG 12



$x_t = f(y_t, y_{t-1}, x_{t-1}) + N_t^x$
$y_t = f(y_{t-1}, x_{t-1}) + N_t^y$
$N_t^x \perp\!\!\!\perp y_t,$
$N_t^x \perp\!\!\!\perp y_{t-1}$
$N_t^x \perp\!\!\!\perp x_{t-1}$
$N_t^y \perp\!\!\!\perp y_{t-1},$
$N_t^y \perp\!\!\!\perp x_{t-1}$
$N_t^x \perp\!\!\!\perp N_s^y \quad \forall t,s$

Fig. 5.3 Unit graphs of all the possible structural equations models under assumption (i)–(viii)

each other. If any independence test is rejected the DAG is rejected, if none of the independence tests are rejected, the DAG is consistent with the data.

Thus, in principle, one could run the regressions corresponding to the equations indicated below each DAG in Fig. 5.3 to check whether a specific DAG is consistent with the data. Let us analyze some specific cases.

If the data are generated by DAG 1 (see Fig. 5.3), and the data generating process were not known to the observer, by constructing the nonparametric regressions[5]:

$$x_t = f_1(x_{t-1}) + N_t^{x,1} \tag{5.20}$$

$$y_t = f_1(y_{t-1}) + N_t^{y,1} \tag{5.21}$$

and by not rejecting the independence relations:

$$\widehat{N_t^{x,1}} \perp\!\!\!\perp x_{t-1}, \tag{5.22}$$

$$\widehat{N_t^{y,1}} \perp\!\!\!\perp y_{t-1}, \tag{5.23}$$

$$\widehat{N_{t-i}^{x,1}} \perp\!\!\!\perp \widehat{N_{t-j}^{y,1}} \quad \text{for} \langle i, j \rangle = \langle 0, 0 \rangle, \langle 1, 0 \rangle, \langle 0, 1 \rangle, \tag{5.24}$$

one would conclude the DAG 1 is consistent with the data. Are other DAGs consistent with these findings? If we run the same regressions but using data generated by DAG 2, we will not necessarily reject: $\widehat{N_t^{x,1}} \perp\!\!\!\perp x_{t-1}$, $\widehat{N_t^{y,1}} \perp\!\!\!\perp y_{t-1}$. Indeed these regressions may suffer of omitted variable bias, but not of reverse causality. However, we will have that $\widehat{N_{t-1}^{x,1}} \not\perp\!\!\!\perp \widehat{N_t^{y,1}}$. Indeed $\widehat{N_t^{y,1}}$ results from a regression in which it is omitted x_{t-1}. Hence $\widehat{N_t^{y,1}}$ is dependent on x_{t-1}, and since x_{t-1} is in turn dependent on $\widehat{N_{t-1}^{x,1}}$, then $\widehat{N_{t-1}^{x,1}} \not\perp\!\!\!\perp \widehat{N_t^{y,1}}$. If we run the same regressions (Eqs. (5.20), (5.21)) using data generated by any other DAG (from DAG 3 to DAG 12), for analogous lines of reasoning we would reach the same conclusion: $\widehat{N_{t-i}^{x,1}} \not\perp\!\!\!\perp \widehat{N_{t-j}^{y,1}}$ for some $\langle i, j \rangle = \langle 0, 0 \rangle, \langle 1, 0 \rangle, \langle 0, 1 \rangle$.

Let us now suppose that DAG 1 has been found not consistent with the data and one runs the nonparametric regressions (also indicated below DAG 2 in Fig. 5.3):

$$x_t = f_2(x_{t-1}) + N_t^{x,2}, \tag{5.25}$$

$$y_t = f_2(x_{t-1}, y_{t-1}) + N_t^{y,2}. \tag{5.26}$$

By not rejecting:

[5]Here and below the subscript i in the function $f_i(\cdot)$, as well as the superscript i in the noise term $N_t^{\cdot,i}$, indicate that these functions and noise terms enter in the additive noise model associated to DAG i (see Fig. 5.3).

$$\widehat{N_t^{x,2}} \perp\!\!\!\perp x_{t-1}, \tag{5.27}$$

$$\widehat{N_t^{y,2}} \perp\!\!\!\perp x_{t-1}, \tag{5.28}$$

$$\widehat{N_t^{y,2}} \perp\!\!\!\perp y_{t-1}, \tag{5.29}$$

$$\widehat{N_{t-i}^{x,2}} \perp\!\!\!\perp \widehat{N_{t-j}^{y,2}} \quad \text{for} \langle i, j \rangle = \langle 0, 0 \rangle, \langle 1, 0 \rangle, \langle 0, 1 \rangle, \tag{5.30}$$

one would conclude the DAG 2 is consistent with the data. If the data were generated by DAG 3, we would have that $\widehat{N_t^{x,2}} \not\perp\!\!\!\perp \widehat{N_{t-1}^{y,2}}$, because in regressing x_t on x_{t-1} we are omitting y_{t-1}, which is a graphical parent of x_t in DAG 3. If the data were generated by any DAG containing the contemporaneous causal link (DAG 4–DAG 12, except DAG 10), we would have that $\widehat{N_t^{x,2}} \not\perp\!\!\!\perp \widehat{N_t^{y,2}}$. If DAG 10 were generating the data, we would have that $\widehat{N_t^{x,2}} \not\perp\!\!\!\perp \widehat{N_{t-1}^{y,2}}$, because, again, we would omit y_{t-1} in the regression of x_t on x_{t-1}.

Let us suppose now that DAG 4 is the data generating process. By running the nonparametric regressions,

$$x_t = f_4(x_{t-1}) + N_t^{x,4} \tag{5.31}$$

$$y_t = f_4(x_t, y_{t-1}) + N_t^{y,4} \tag{5.32}$$

and not rejecting

$$\widehat{N_t^{x,4}} \perp\!\!\!\perp x_{t-1} \tag{5.33}$$

$$\widehat{N_t^{y,4}} \perp\!\!\!\perp x_t \tag{5.34}$$

$$\widehat{N_t^{y,4}} \perp\!\!\!\perp y_{t-1} \tag{5.35}$$

$$\widehat{N_{t-i}^{x,4}} \perp\!\!\!\perp \widehat{N_{t-j}^{y,4}} \quad \text{for} \langle i, j \rangle = \langle 0, 0 \rangle, \langle 1, 0 \rangle, \langle 0, 1 \rangle, \tag{5.36}$$

we would conclude that DAG 4 is consistent with the data. If the data generating process were any DAG with opposite contemporaneous causal link (DAG 5, 7, 9, 12), running the same regressions ((5.31), (5.32)) and tests ((5.33)–(5.36)), we would get $\widehat{N_t^{y,4}} \not\perp\!\!\!\perp x_t$. If the data generating process were any DAG among DAG 2, 3, 6, 8, 10, 11, there would be no reverse contemporaneous causal link, but an omitted lagged variables in one (or both) of the two regressions. This would imply that $\widehat{N_{t-i}^{x,4}} \not\perp\!\!\!\perp \widehat{N_{t-j}^{y,4}}$ for some $\langle i, j \rangle = \langle 1, 0 \rangle, \langle 0, 1 \rangle$.

These examples should already suggest that, under the framework of the 12 possible DAGs of Fig. 5.3, under the assumptions listed above, with an exhaustive search of independence relationships derived by the possible DAGs, one is able to uniquely identify the model that has generated the data. Based on these considerations, we propose a search procedure formalized in the algorithm described in the Table here

below. The algorithm avoids an exhaustive causal search, but at the same time is able to uniquely identify, among the 12 DAGs represented in Fig. 5.3, the one that has generated the data.

The search algorithm is able to efficiently infer one of the 12 DAGs on the base of a limited number of nonparametric regressions and tests of unconditional independence. Once the algorithms outputs DAG number i, however, we suggest to check its consistency with the data through the nonparametric regressions and (conditional and unconditional) independence tests indicated in Fig. 5.3 under the inferred DAG number.

For a more general framework in which there are k possible time series and p lags of causal influence, Peters et al. [56] propose a search procedure based on additive noise models called TiMINO, i.e. time series models with independent noise. The

Search Algorithm

1. **Input**: Samples from a 2-dimensional time series of length T, maximal order $p = 1$.
2. Run nonpar. regressions: $x_t = f_1(x_{t-1}) + N_t^{x,1}$; $y_t = f_1(y_{t-1}) + N_t^{y,1}$, get $\widehat{N_t^{x,1}}$, $\widehat{N_t^{y,1}}$
3. Test: $\widehat{N_t^{x,1}} \perp\!\!\!\perp \widehat{N_t^{y,1}}$
4. If $\widehat{N_t^{x,1}} \perp\!\!\!\perp \widehat{N_t^{y,1}}$
5. Test: $\widehat{N_{t-i}^{x,1}} \perp\!\!\!\perp \widehat{N_{t-j}^{y,1}}$ for $\langle i, j\rangle = \langle 1, 0\rangle, \langle 0, 1\rangle$
6. If $\widehat{N_{t-i}^{x,1}} \perp\!\!\!\perp \widehat{N_{t-j}^{y,1}}$ for $\langle i, j\rangle = \langle 1, 0\rangle, \langle 0, 1\rangle$, break, output DAG 1
7. If $\widehat{N_t^{x,1}} \perp\!\!\!\perp \widehat{N_{t-1}^{y,1}}$ and $\widehat{N_{t-1}^{x,1}} \not\perp\!\!\!\perp \widehat{N_t^{y,1}}$, then break, output DAG 2
8. If $\widehat{N_{t-1}^{x,1}} \perp\!\!\!\perp \widehat{N_t^{y,1}}$ and $\widehat{N_t^{x,1}} \not\perp\!\!\!\perp \widehat{N_{t-1}^{y,1}}$, then break, output DAG 3
9. Else break, output DAG 10
10. If $\widehat{N_t^{x,1}} \not\perp\!\!\!\perp \widehat{N_t^{y,1}}$
11. Run nonp. reg.: $x_t = f_4(x_{t-1}) + N_t^{x,4}$; $y_t = f_4(x_t, y_{t-1}) + N_t^{y,4}$, get $\widehat{N_t^{x,4}}$, $\widehat{N_t^{y,4}}$
12. Test: $\widehat{N_t^{y,4}} \perp\!\!\!\perp x_t$
13. If $\widehat{N_t^{y,4}} \perp\!\!\!\perp x_t$
14. Test: $\widehat{N_{t-i}^{x,4}} \perp\!\!\!\perp \widehat{N_{t-j}^{y,4}}$ for $\langle i, j\rangle = \langle 0, 0\rangle, \langle 1, 0\rangle, \langle 0, 1\rangle$
15. If $\widehat{N_{t-i}^{x,4}} \perp\!\!\!\perp \widehat{N_{t-j}^{y,4}}$ for $\langle i, j\rangle = \langle 0, 0\rangle, \langle 1, 0\rangle, \langle 0, 1\rangle$, break, output DAG 4
16. If $\widehat{N_{t-i}^{x,4}} \perp\!\!\!\perp \widehat{N_{t-j}^{y,4}}$ *only* for $\langle i, j\rangle = \langle 0, 0\rangle, \langle 0, 1\rangle$, break, output DAG 6
17. If $\widehat{N_{t-i}^{x,4}} \perp\!\!\!\perp \widehat{N_{t-j}^{y,4}}$ *only* for $\langle i, j\rangle = \langle 0, 0\rangle, \langle 1, 0\rangle$, break, output DAG 8
18. Else break, output DAG 11
19. If $\widehat{N_t^{y,4}} \not\perp\!\!\!\perp x_t$
20. Run $x_t = f_5(x_{t-1}, y_t) + N_t^{x,5}$; $y_t = f_5(y_{t-1}) + N_t^{y,5}$, get $\widehat{N_t^{x,5}}$, $\widehat{N_t^{y,5}}$
21. Test: $\widehat{N_{t-i}^{x,5}} \perp\!\!\!\perp \widehat{N_{t-j}^{y,5}}$ for $\langle i, j\rangle = \langle 0, 0\rangle, \langle 1, 0\rangle, \langle 0, 1\rangle$
22. If $\widehat{N_{t-i}^{x,5}} \perp\!\!\!\perp \widehat{N_{t-j}^{y,5}}$ for $\langle i, j\rangle = \langle 0, 0\rangle, \langle 1, 0\rangle, \langle 0, 1\rangle$, break, output DAG 5
23. If $\widehat{N_{t-i}^{x,5}} \perp\!\!\!\perp \widehat{N_{t-j}^{y,5}}$ *only* for $\langle i, j\rangle = \langle 0, 0\rangle, \langle 0, 1\rangle$, break, output DAG 7
24. If $\widehat{N_{t-i}^{x,5}} \perp\!\!\!\perp \widehat{N_{t-j}^{y,5}}$ *only* for $\langle i, j\rangle = \langle 0, 0\rangle, \langle 1, 0\rangle$, break, output DAG 9
25. Else break, output DAG 12
26. **Output**: One DAG among DAG 1 - DAG 12.

output of TiMINO is however, a *summary* graph. This means that it is not possible to disentangle between contemporaneous and lagged causal effects. The advantage of our search algorithm is that it is possible to distinguish between these two types of effects, but only under the specific framework of time series pairs.

5.3.4 Local Projections

Local projections were introduced by Jorda [42] to compute impulse responses (see Sect. 5.2.4) without specifying and estimating a VAR model. Furthermore, any attempt of representing the data generating process through a multivariate time series structural system is eschewed in local projections. The idea here is to focus on the estimation of impulse responses through regression methods that are applied at each period of interest, without hinging on a pre-specified or pre-estimated time series model.

Let be $Y_t = (x_t, y_t)'$, as in Sect. 5.2.1. Jorda [42] considered projecting Y_{t+s} onto the linear space generated by $(Y_{t-1}, \ldots, Y_{t-p})'$ for a certain choice of lag p, namely

$$Y_{t+s} = \alpha^s + P_1^{s+1} Y_{t-1} + P_2^{s+1} Y_{t-2} + \ldots + P_p^{s+1} Y_{t-p} + u_{t+s}^s, \tag{5.37}$$

where α^s is a (2×1) vector of constant, P_i^{s+1} are (2×2) matrices of coefficients, and u_{t+s}^s is a (2×1) vector of errors by construction uncorrelated with the regressors. Superscripts here are meant to denote the time window where the regression is performed.

Impulse response functions are defined as the difference between two forecasts, which is an idea consistent with Eq. (5.12). More specifically, we have that the impulse response of x_{t+s} to a shock at time t, $s \in Z$ is

$$IR(t, s, \delta) = E(x_{t+s} | v_{1t} = \delta, Y_t) - E(x_{t+s} | v_{1t} = 0, Y_t) \quad i = 0, \ldots, H. \tag{5.38}$$

where $E(\cdot|\cdot)$ denotes the best, mean squared predictor, v_{1t} is a disturbance shock, and d is the magnitude of the shock the impact of which one wants to measure.

The impulse responses estimated from (5.37) are

$$IR(t, s, \delta) = \hat{P}_1^s \delta. \tag{5.39}$$

As noted by Kilian and Lütkepohl [44, chapter 12], these impulse responses will be relative to a reduced-form error ($v_{it} = u_{it}$) and not to the true shock affecting the system, if they are estimated directly through a least square regression of Eq. (5.38). Thus, it is fundamental in this context to transform the reduced-form residuals in a mixture of structural shocks. But here the problem is analogous to the problem of identification of the structural VAR model and the literature on local projections seems not to have found a method yet that bypasses this step.

5.4 Conclusions

In this paper we have addressed the problem of causal inference from data that are realizations of bivariate time series processes. We have focused on the setting typically encountered in econometrics, namely stationary or difference-stationary autoregressive processes with additive noises. The standard approach in econometrics to address this problem is structural vector autoregressive analysis. This allows the researcher to filter the time-series data, in order to apply causal search algorithms to the i.i.d. filtered data. Since the time structure is filtered out, the output of this causal search is a contemporaneous causal structure, which, in a second step, gives the possibility of recovering the entire structural autoregressive model. In a causal pair setting, however, causal search in this framework is limited. For example, in the case of Gaussian data, the linear causal structure between the two filtered time series is not identifiable. We have shown that identification is possible under non-Gaussianity (exploiting independent component analysis) or under non-linearity (exploiting non-linear additive noise model). But we have also shown that in a setting of bivariate time series, an alternative valid approach is to address the problem of causal inference by avoiding the vector autoregressive framework. This is possible by applying graphical models algorithms (like FCI) or nonlinear additive noise models algorithms (like the one presented in this paper) directly to the data, without filtering them. We have also shown the possibility of applications of Granger non-causality testing and local projections in a framework in which VAR models are not necessarily estimated. The latter two techniques, however, deviate for many aspects, from a structural interpretation of causality (see footnote 1), i.e. from a framework which allows intervention, while they are closer to a notion of predictability. A study of the relative merits of the different methods presented above with empirical and simulated data is left to future research.

Acknowledgements The authors want to thank Isabelle Guyon, Alexander Statnikov, and Daniele Marinazzo for very valuable comments on a first draft.

References

1. Ehung Baek and William Brock. A general test for nonlinear Granger causality: Bivariate model. *Iowa State University and University of Wisconsin at Madison Working Paper*, 1992.
2. David Bell, Jim Kay, and Jim Malley. A non-parametric approach to non-linear causality testing. *Economics Letters*, 51(1):7–18, 1996.
3. David A Bessler and Seongpyo Lee. Money and prices: US data 1869–1914 (a study with directed graphs). *Empirical Economics*, 27(3):427–446, 2002.
4. Taoufik Bouezmarni, Jeroen VK Rombouts, and Abderrahim Taamouti. Nonparametric copula-based test for conditional independence with applications to Granger causality. *Journal of Business & Economic Statistics*, 30(2):275–287, 2012.
5. Jean-François Cardoso and Antoine Souloumiac. Blind beamforming for non-Gaussian signals. In *IEE proceedings F (radar and signal processing)*, volume 140, pages 362–370. IET, 1993.

6. Nadine Chlaß and Alessio Moneta. Can graphical causal inference be extended to nonlinear settings? In *EPSA Epistemology and Methodology of Science*, pages 63–72. Springer, 2009.
7. Tianjiao Chu and Clark Glymour. Search for additive nonlinear time series causal models. *Journal of Machine Learning Research*, 9(May):967–991, 2008.
8. Diego Colombo, Marloes H Maathuis, Markus Kalisch, and Thomas S Richardson. Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics*, pages 294–321, 2012.
9. Pierre Comon. Independent component analysis, a new concept? *Signal processing*, 36(3):287–314, 1994.
10. David Danks and Sergey Plis. Learning causal structure from undersampled time series. *JMLR: Workshop and Conference Proceedings (NIPS Workshop on Causality)*, 2013.
11. Selva Demiralp and Kevin D Hoover. Searching for the causal structure of a vector autoregression. *Oxford Bulletin of Economics and statistics*, 65(s1):745–767, 2003.
12. Cees Diks and Valentyn Panchenko. A new statistic and practical guidelines for nonparametric Granger causality testing. *Journal of Economic Dynamics & Control*, 30:1647–1669, 2006.
13. Michael Eichler. Causal inference in time series analysis. *Causality: Statistical Perspectives and Applications*, pages 327–354, 2012.
14. Doris Entner and Patrik O Hoyer. On causal discovery from time series data using FCI. *Probabilistic graphical models*, pages 121–128, 2010.
15. Jan Eriksson and Visa Koivunen. Identifiability, separability, and uniqueness of linear ICA models. *IEEE signal processing letters*, 11(7):601–604, 2004.
16. Jean-Pierre Florens and Michel Mouchart. A note on noncausality. *Econometrica*, pages 583–591, 1982.
17. Claudia Foroni, Eric Ghysels, and Massimiliano Marcellino. Mixed-frequency vector autoregressive models. In *VAR Models in Macroeconomics–New Developments and Applications: Essays in Honor of Christopher A. Sims*, pages 247–272. Emerald Group Publishing Limited, 2013.
18. Andrea Gazzani and Alejandro Vicondoa. Proxy-svar as a bridge between mixed frequencies. *Unpublished Manuscript*, 2016.
19. John Geweke. Measurement of linear dependence and feedback between multiple time series. *Journal of the American Statistical Association*, 77(378):304–313, 1982.
20. Irving John Good. A causal calculus (i). *The British Journal for the Philosophy of Science*, 11(44):305–318, 1961a.
21. Irving John Good. A causal calculus (ii). *The British Journal for the Philosophy of Science*, 12(45):43–51, 1961b.
22. Christian Gouriéroux, Alain Monfort, and Jean-Paul Renne. Statistical inference for independent component analysis: Application to structural VAR models. *Journal of Econometrics*, 196(1):111–126, 2017.
23. Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, pages 424–438, 1969.
24. Clive WJ Granger. Testing for causality: a personal viewpoint. *Journal of Economic Dynamics and control*, 2:329–352, 1980.
25. Arthur Gretton, Kenji Fukumizu, Choon H Teo, Le Song, Bernhard Schölkopf, and Alex J Smola. A kernel statistical test of independence. In *Advances in neural information processing systems*, pages 585–592, 2008.
26. James Douglas Hamilton. *Time Series Analysis*, volume 2. Princeton university press Princeton, NJ, 1994.
27. Wolfgang Härdle, Helmut Lütkepohl, and Rong Chen. A review of nonparametric time series analysis. *International Statistical Review*, 65(1):49–72, 1997.
28. Trevor Hastie and Rob Tibshirani. Independent components analysis through product density estimation. In *Advances in neural information processing systems*, pages 665–672, 2003.
29. David F Hendry. *Dynamic Econometrics*. Oxford University Press on Demand, 1995.
30. Helmut Herwartz. Hodges–Lehmann detection of structural shocks–an analysis of macroeconomic dynamics in the Euro area. *Oxford Bulletin of Economics and Statistics*, 2018.

31. Craig Hiemstra and Jonathan D Jones. Testing for linear and nonlinear Granger causality in the stock price-volume relation. *The Journal of Finance*, 49(5):1639–1664, 1994.
32. Kevin D Hoover. *Causality in Macroeconomics*. Cambridge University Press, 2001.
33. Kevin D Hoover. Economic theory and causal inference. In Uskali Mäki, editor, *Philosophy of Economics*. North Holland, 2012.
34. Patrik O Hoyer, Shohei Shimizu, Antti J Kerminen, and Markus Palviainen. Estimation of causal effects using linear non-gaussian causal models with hidden variables. *International Journal of Approximate Reasoning*, 49(2):362–378, 2008.
35. Patrik O Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in neural information processing systems*, pages 689–696, 2009.
36. Antti Hyttinen, Sergey Plis, Matti Järvisalo, Frederick Eberhardt, and David Danks. Causal discovery from subsampled time series data by constraint optimization. *JMLR: Workshop and Conference Proceedings (PGM)*, 2016.
37. Aapo Hyvärinen. Independent component analysis: recent advances. *Phil. Trans. R. Soc. A*, 371(1984):20110534, 2013.
38. Aapo Hyvärinen and Erkki Oja. A fast fixed-point algorithm for independent component analysis. *Neural computation*, 9(7):1483–1492, 1997.
39. Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. *Independent Component Analysis. Series on Adaptive and Learning Systems for Signal Processing, Communications, and Control*. Wiley, 2001.
40. Aapo Hyvärinen, Kun Zhang, Shohei Shimizu, and Patrik O Hoyer. Estimation of a structural vector autoregression model using non-gaussianity. *Journal of Machine Learning Research*, 11(May):1709–1731, 2010.
41. Rodney L Jacobs, Edward E Leamer, and Michael P Ward. Difficulties with testing for causation. *Economic Inquiry*, 17(3):401–413, 1979.
42. Oscar Jorda. Estimation and inference of impulse responses by local projections. *American Economic Review*, 95(1):161–182, 2005.
43. Maria Kalli and Jim E Griffin. Bayesian nonparametric vector autoregressive models. *Journal of Econometrics*, 203(2):267–282, 2018.
44. Lutz Kilian and Helmut Lütkepohl. *Structural Vector Autoregressive Analysis*. Cambridge University Press, 2017.
45. Gary Koop, M Hashem Pesaran, and Simon M Potter. Impulse response analysis in nonlinear multivariate models. *Journal of Econometrics*, 74(1):119–147, 1996.
46. RJ Kulperger and RA Lockhart. Tests of independence in time series. *Journal of Time Series Analysis*, 19(2):165–185, 1998.
47. Gustavo Lacerda, Peter L Spirtes, Joseph Ramsey, and Patrik O Hoyer. Discovering cyclic causal models by independent components analysis. In *Proc. Conf. on Uncertainty in Artificial Intelligence (UAI-08)*, pages 366–374, 2008.
48. Markku Lanne, Mika Meitz, and Pentti Saikkonen. Identification and estimation of non-gaussian structural vector autoregressions. *Journal of Econometrics*, 196(2):288–304, 2017.
49. Helmut Lütkepohl. Vector autoregressive models. In Nigar Hashimzade and Michael A. Thornton, editors, *Handbook of Research Methods and Applications in Empirical Macroeconomics*, pages 139–164. Edward Elgar, 2013.
50. Massimiliano Marcellino. Some consequences of temporal aggregation in empirical analysis. *Journal of Business & Economic Statistics*, 17(1):129–136, 1999.
51. David S Matteson and Ruey S Tsay. Independent component analysis via distance covariance. *Journal of the American Statistical Association*, 112(518):623–637, 2017.
52. Alessio Moneta. Graphical causal models and vars: an empirical assessment of the real business cycles hypothesis. *Empirical Economics*, 35(2):275–300, 2008.
53. Alessio Moneta, Doris Entner, Patrik O Hoyer, and Alex Coad. Causal inference by independent component analysis: Theory and applications. *Oxford Bulletin of Economics and Statistics*, 75(5):705–730, 2013.

54. Judea Pearl. *Probabilistic reasoning in intelligent systems: Networks of plausible reasoning*. Morgan Kaufmann Publishers, Los Altos, 1988.
55. Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2009.
56. Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Causal inference on time series using restricted structural equation models. In *Advances in Neural Information Processing Systems*, pages 154–162, 2013.
57. Jonas Peters, Joris M Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *The Journal of Machine Learning Research*, 15(1):2009–2053, 2014.
58. Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT press, 2017.
59. Giorgio E Primiceri. Time varying structural vector autoregressions and monetary policy. *The Review of Economic Studies*, 72(3):821–852, 2005.
60. Thomas Richardson and Peter Spirtes. Ancestral graph markov models. *The Annals of Statistics*, 30(4):962–1030, 2002.
61. Shohei Shimizu, Patrik O. Hoyer, Aapo Hyvärinen, and Antti Kerminen. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7:2003–2030, 2006.
62. Christopher A Sims. Macroeconomics and reality. *Econometrica*, pages 1–48, 1980.
63. Peter Spirtes and Kun Zhang. Causal discovery and inference: concepts and recent methodological advances. In *Applied informatics*, volume 3, page 3. SpringerOpen, 2016.
64. Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT press, 2000.
65. Wolfgang Spohn. Probabilistic causality: From Hume via Suppes to Granger. In M. Galvotti and G. Gambetta, editors, *Causalità e modelli probabilistici*, pages 69–87. Clueb, 1983.
66. Liangjun Su and Halbert White. A nonparametric Hellinger metric test for conditional independence. *Econometric Theory*, 24(4):829–864, 2008.
67. Patrick Suppes. *A probabilistic theory of causality*. North-Holland, Amsterdam, 1970.
68. Norman R Swanson and Clive WJ Granger. Impulse response functions based on a causal approach to residual orthogonalization in vector autoregressions. *Journal of the American Statistical Association*, 92(437):357–367, 1997.
69. Kun Zhang and Aapo Hyvärinen. On the identifiability of the post-nonlinear causal model. In *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*, pages 647–655. AUAI Press, 2009.
70. Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, pages 804–813. AUAI Press, 2011.