

CAUSAL INFERENCE IN TIME SERIES ANALYSIS

MICHAEL EICHLER

*Department of Quantitative Economics, Maastricht University
P.O. Box 616, 6200 MD Maastricht, The Netherlands*

November 11, 2011

1. INTRODUCTION

The identification of causal relationships is an important part of scientific research and essential for understanding the consequences when moving from empirical findings to actions. At the same time, the notion of causality has shown to be evasive when trying to formalize it. Among the many properties a general definition of causality should or should not have, there are two important aspects that are of practical relevance:

Temporal precedence: causes precede their effects;

Physical influence: manipulation of the cause changes the effects.

The second aspect is central to most of the recent literature on causal inference (e.g. Pearl 2000, Lauritzen 2001, Dawid 2002, Spirtes et al. 2001), which is also demonstrated by the previous chapters in this book. Here, causality is defined in terms of the effect of interventions, which break the symmetry of association and thus give a direction to the association between two variables.

In time series analysis, most approaches to causal inference make use of the first aspect of temporal precedence. On the one hand, controlled experiments are often not feasible in many time series applications and researchers may be reluctant to think in these terms. On the other hand, temporal precedence is readily available in time series data.

Among these approaches, the definition introduced by Granger (1969, 1980, 1988) is probably the most prominent and most widely used concept. This concept of causality does not rely on the specification of a scientific model and thus is particularly suited for empirical investigations of cause-effect relationships. On the other hand, it is commonly known that Granger causality basically is a measure of association between the variables and thus can lead to so-called spurious causalities if important relevant variables are not included in the analysis (Hsiao 1982). Since in most analyses involving time series data the presence of latent variables that affect the measured components cannot be ruled out, this raises the question whether and how the causal structure can be recovered from time series data.

The objective of this chapter is to embed the concept of Granger causality in the broader framework of modern graph-based causal inference. It is based on a series of papers by the author (Eichler 2007, 2006, 2005, 2009, 2010, Eichler and Didelez 2007, 2010). We start in Section 2 by comparing four possible definitions of

E-mail address: m.eichler@maastrichtuniversity.nl (M. Eichler)

causality that have been used in the context of time series. In Section 3, we sketch approaches for representing the dependence structure of a time series graphically. Furthermore, we discuss in more detail Markov properties associated with Granger causality as these are most relevant for the purpose of this chapter. This is continued in Section 4, where graphical representation of systems with latent variables are considered. Section 5 covers the identification of causal effects from observational data while Sections 6 and 7 present approaches for learning causal structures based on Granger-causality. Section 8 concludes.

2. CAUSALITY FOR TIME SERIES

Suppose that $X = (X_t)_{t \in \mathbb{Z}}$ and $Y = (Y_t)_{t \in \mathbb{Z}}$ are two stationary time series that are statistically dependent on each other. When is it justified to say that the one series X causes the other series Y ? Questions of this kind are important when planning to devise actions, implementing new policies, or subjecting patients to a treatment. Nonetheless, the notion of causality has been evasive and formal approaches to define causality have been much debated and criticised. In this section, we review four approaches to formalize causality in the context of multivariate time series. We start by introducing some notation that we will use throughout this chapter.

Consider a multivariate stationary time series $X = (X_t)_{t \in \mathbb{Z}}$ consisting of random vectors $X_t = (X_{1,t}, \dots, X_{n_x,t})'$ defined on a joint probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The history of X up to time t will be denoted by $X^t = (X_s, s \leq t)$; furthermore, $\sigma\{X^t\}$ denotes the corresponding σ -algebra generated by X^t . The σ -algebra $\sigma\{X^t\}$ represents the information that is obtained by observing the time series X . Finally, for every $1 \leq a \leq n_x$, we decompose X_t into its a -th component $X_{a,t}$ and all other components $X_{\cdot,a,t}$.

2.1. Intervention causality

The previous chapters of this book have shown how important and fruitful the concept of interventions has been for the understanding of causality. This approach formalizes the understanding that causal relationships are fundamental and should persist if certain aspects of the system are changed. In the context of time series, this idea of defining a causal effect as the effect of an intervention in such a system as has first been proposed by Eichler and Didelez (2007).

We start by introducing the concept of intervention indicators already seen in Chapters 4, 14 and 16 of this book. Such indicators allow us to distinguish formally between the “natural” behaviour of a system and its behaviour under an intervention (e.g. Pearl 1993, 2000, Lauritzen 2001, Dawid 2002, Spirtes et al. 2001).

Definition 2.1 (Regimes). Consider a set of indicators $\sigma = \{\sigma_t; t \in \tau\}$ denoting interventions in X_t at points $t \in \tau$ in time. Each σ_t takes values in some set \mathcal{S} augmented by one additional state \emptyset . Different values of σ indicate different distributions of the time series $V_t = (X_t, Y_t, Z_t)$ in the following way.

- (i) *Idle Regime:* if $\sigma_t = \emptyset$ no intervention is performed and the process X_t arise naturally. The corresponding probability measure will be denoted as $\mathbb{P}_{\sigma_t=\emptyset}$ (often abbreviated to \mathbb{P}_\emptyset or even just \mathbb{P}). This regime is also called the *observational regime*.

- (ii) *Atomic interventions*: here σ_t takes values in the domain of X_t such that $\sigma_t = x^*$ means we intervene and force X_t to assume the value x^* . Hence $\mathbb{P}_{\sigma_t=x^*}$ (or shorter \mathbb{P}_{x^*}) denotes the probability measure under such an atomic intervention with

$$\mathbb{P}_{\sigma_t=x^*}(X_t = x | V^{t-1}) = \delta_{\{x^*\}}(x),$$

where $\delta_A(x)$ is one if $x \in A$ and zero otherwise.

- (iii) *Conditional intervention*: here \mathcal{S} consists of functions $g_{x,t}(C^{t-1})$, where C is a subseries of the process V , such that $\sigma_t = g$ means X_t is forced to take on a value that depends on past observations C^{t-1} . With $\mathbb{P}_{\sigma_t=g}$ denoting the distribution of the time series V under such a conditional intervention, we have

$$\mathbb{P}_{\sigma_t=g}(X_t = x | V^{t-1}) = \mathbb{P}_{\sigma_t=g}(X_t = x | C^{t-1}) = \delta_{\{g_{x,t}(C^{t-1})\}}(x).$$

- (iv) *Random intervention*: here \mathcal{S} consists of distributions meaning that X_t is forced to arise from such a distribution, that is, the conditional distribution $\mathbb{P}_{\sigma_t=s}(X_t | V^{t-1})$ is *known* and possibly a function of C^{t-1} for a subseries C of V .

With this definition, we are considering a family of probability measures \mathbb{P}_σ on (Ω, \mathcal{F}) indexed by the possible values that σ can take. While $\mathbb{P} = \mathbb{P}_{\sigma=\emptyset}$ describes the natural behaviour of the time series under the observational regime, any implementation of an intervention strategy $\sigma = s$ will change the probability measure to $\mathbb{P}_{\sigma=s}$. We note that the assumption of stationarity is only made for the natural probability measure \mathbb{P} . In the following, we write \mathbb{E}_\emptyset (or shorter just \mathbb{E}) and $\mathbb{E}_{\sigma=s}$ to distinguish between expectations with respect to \mathbb{P}_\emptyset and $\mathbb{P}_{\sigma=s}$; the shorthand \mathbb{P}_s and \mathbb{E}_s is used when it is clear from the context what variables are intervened in.

One common problem in time series analysis is that controlled experiments cannot be carried out. Therefore, in order to assess the possible success of a meditated intervention, the effect of this intervention must be predicted from data collected under the observational regime. This is only possible if the intervention affects only the distribution of the target variable X_t at times $t \in \tau$ according to Definition 2.1 (ii)–(iv) whereas all other conditional distributions remain the same as under the idle regime. To formalize this invariance under different regimes, we say that a random variable Y is independent of σ_t conditionally on some σ -algebra $\mathcal{H} \subseteq \mathcal{F}$ if the conditional distributions of Y given \mathcal{H} under the probability measures $\mathbb{P}_{\sigma_t=s}$ and \mathbb{P}_\emptyset are almost surely equal for all $s \in \mathcal{S}$. With this notation, the required link between the probability measures \mathbb{P} and \mathbb{P}_σ under the observational and the interventional regime, respectively, is established by the following assumptions, which are analogous to those of (extended) stability in Dawid and Didelez (2005). See also Chapter 8 by Berzuini, Dawid and Didelez for a discussion of this issue.

Assumptions (Stability). Let $V = (X, Y, Z)$ be a multivariate time series that is stationary under the idle regime. The interventions $\sigma = \{\sigma_t, t \in \tau\}$ from Definition 2.1 are assumed to have the following properties.

- (I1) for all $t \notin \tau$: $V_t \perp\!\!\!\perp \sigma | V^{t-1}$;
- (I2) for all $t \in \tau$: $V_t \perp\!\!\!\perp \{\sigma_{t'} | t' \in \tau, t' \neq t\} | V^{t-1}, \sigma_t$;
- (I3) for all $t \in \tau$: $Y_t, Z_t \perp\!\!\!\perp \sigma_t | V^{t-1}$;
- (I4) for all $t \in \tau$ and all $a = 1, \dots, n_x$: $X_{a,t} \perp\!\!\!\perp X_{-a,t}, \sigma_{-a,t} | V^{t-1}, \sigma_{a,t} = s$.

When working with such (conditional) independence relations, it is important to remember that the intervention indicators σ_t are not random variables and that the above notion of independence—although being one of statistical independence—is not symmetric. For a more detailed discussion of conditional independence involving non-random quantities we refer to Dawid (2002) and Chapter 4 of this book.

With the above assumptions the distribution of the time series $V = (X, Y, Z)$ under the interventional regime is fully specified by its natural distribution described by $\mathbb{P} = \mathbb{P}_\emptyset$ and the conditional distributions given in Definition 2.1 (ii)–(iv) under the chosen intervention. As it is usually not possible or feasible to collect data under the interventional regimes, all model assumptions such as stationarity are supposed to apply to the idle regime.

In the case of a single intervention, that is, in one variable at one point in time, the above assumptions (I1) to (I3) simplify to

$$X^{t-1}, Y^t, Z^t \perp\!\!\!\perp \sigma_t \quad (2.1)$$

and

$$\{X_{t'}, Y_{t'}, Z_{t'} | t' > t\} \perp\!\!\!\perp \sigma_t | X^t, Y^t, Z^t. \quad (2.2)$$

Let us now consider effects of interventions. In general, this can be any function of the post-intervention distribution of $\{V_{t'} | t' > t\}$ given an individual intervention $\sigma_t = s$, for instance. It will often involve the comparison of setting X_t to different values, e.g. setting $X_t = x^*$ as compared to setting it to $X_t = x^0$ which could be a baseline value in some sense. One may also want to use the idle case as baseline for the comparison. Typically, one is interested in the mean difference between interventions or between an intervention and the idle case. This leads to the following definition of the average causal effect.

Definition 2.2 (Average causal effect). The *average causal effect (ACE)* of interventions in X_{t_1}, \dots, X_{t_m} according to strategy s on the response variable $Y_{t'}$ with $t' > t$ is given by

$$\text{ACE}_s = \mathbb{E}_{\sigma_t=s} Y_{t'} - \mathbb{E} Y_{t'}.$$

In the stationary case, we may assume without loss of generality that $\mathbb{E} Y_{t'} = 0$ and thus $\text{ACE}_s = \mathbb{E}_{\sigma_t=s} Y_{t'}$. Furthermore, different intervention strategies can be compared by considering the difference $\text{ACE}_{s_1} - \text{ACE}_{s_2}$. Finally, we note that the effect of an intervention need not be restricted to the mean. For example, in financial time series, an intervention might aim at reducing the volatility of the stock market. In general, one can consider any functional of the post-intervention distribution $\mathbb{P}_s(Y_{t'})$.

2.2. Structural causality

In a recent article, White and Lu (2010) proposed a new concept of so-called direct structural causality for the discussion of causality in dynamic structural systems. The approach is based on the assumption that the data-generating process (DGP) has a recursive dynamic structure in which predecessors structurally determine successors. For ease of notation, the following definitions are slightly modified and simplified.

Suppose that we are interested in the causal relationship between two processes, the “cause of interest” X and the “response of interest” Y . We assume that X and

Y are structurally generated as

$$\begin{aligned} X_t &= q_{x,t}(X^{t-1}, Y^{t-1}, Z^{t-1}, U_x^t) \\ Y_t &= q_{y,t}(X^{t-1}, Y^{t-1}, Z^{t-1}, U_y^t) \end{aligned} \quad (2.3)$$

for all $t \in \mathbb{Z}$. Here, the process Z includes all relevant observed variables while the realizations of $U = (U_x, U_y)$ are assumed to be unobserved. The functions $q_{x,t}$ and $q_{y,t}$ are also assumed to be unknown.

Definition 2.3. The process X does not directly structurally cause the process Y if the function $q_{y,t}(x^{t-1}, y^{t-1}, z^{t-1}, u_y^t)$ is constant in x^{t-1} for all admissible values for y^{t-1} , z^{t-1} , and u_y^t . Otherwise, X is said to directly structurally cause Y .

We note that Similar approaches of defining causality by assuming a set of structural equations have been consider before by a number of authors (e.g. Pearl and Verma 1991, Pearl 2000 and also Shpitser 2011 in this book). However, in contrast to this strand of literature, White and Lu (2010) make no reference to interventions or to graphs.

It is clear from the definition that an intervention σ_t on X_t results in replacing the generating equation by the corresponding equation under the interventional regime. For instance, in case of a conditional intervention, we have $X_t = g_{x,t}(C^{t-1})$, where the subprocess C denotes the set of conditioning variables. Consequently, as the generating equation for the response variable $Y_{t'}$, $t' > t$, is unaffected by the intervention, we immediately obtain the following result.

Corollary 2.4. Suppose that (X, Y) is generated by (2.3). If X does not directly structurally cause Y , then

$$\mathbb{E}_{\sigma_t=s}(h(Y_{t+1})) = \mathbb{E}_{\emptyset}(h(Y_{t+1}))$$

for all measurable functions h .

White and Lu (2010) also propose a definition of total structural causality. With this, the above corollary can be generalized to responses at arbitrary times $t' > t$. We omit the details and refer the reader to White and Lu (2010). Furthermore, we note that the converse of the above result is generally not true. This is due to the fact that the function $q_{y,t}$ might depend on x^{t-1} only on a set having probability zero. In that case, the dependence on x^{t-1} will not show up in the ACE.

2.3. Granger causality

In time series analysis, inference about cause-effect relationships is commonly based on the concept of Granger causality Granger (1969, 1980). Unlike the two previous approaches, this probabilistic concept of causality does not rely on the specification of a scientific model and thus is particularly suited for empirical investigations of cause-effect relationships. For his general definition of causality, Granger (1969, 1980) evokes the following two fundamental principles:

- (i) the effect does not precede its cause in time;
- (ii) the causal series contains unique information about the series being caused that is not available otherwise.

The first principle of temporal precedence of causes is commonly accepted and has been also the basis for other probabilistic theories of causation (e.g., Good 1961,

1962, Suppes 1970). In contrast, the second principle is more subtle as it requires the separation of the special information provided by the former series X from any other possible information. To this end, Granger considers two information sets:

- (i) $\mathcal{I}^*(t)$ is the set of all information in the universe up to time t ;
- (ii) $\mathcal{I}_{-X}^*(t)$ contains the same information set except for the values of series X up to time t .

Here it is assumed that all variables in the universe are measured at equidistant points in time, namely $t \in \mathbb{Z}$. Now, if the series X causes series Y , we expect by the above principles that the conditional probability distributions of Y_{t+1} given the two information sets $\mathcal{I}^*(t)$ and $\mathcal{I}_{-X}^*(t)$ differ from each other. The following equivalent formulation is chosen to avoid measure-theoretic subtleties.

Granger's definition of causality (1969, 1980). The series X *does not cause* the series Y if

$$Y_{t+1} \perp\!\!\!\perp \mathcal{I}^*(t) \mid \mathcal{I}_{-X}^*(t) \quad (2.4)$$

for all $t \in \mathbb{Z}$; otherwise the series X is said to *cause* the series Y .

Besides measure-theoretic subtleties and the obviously abstract nature of the set $\mathcal{I}^*(t)$, a problem of the above definition is whether $\mathcal{I}_{-X}^*(t)$ contains truly less information than $\mathcal{I}^*(t)$. Implicitly, such a separation of the two information sets $\mathcal{I}^*(t)$ and $\mathcal{I}_{-X}^*(t)$ seems to be based on the assumption that the universe considered is discretized not only in time (as we consider time-discrete processes) but also in space.

Leaving aside such theoretical problems, it is interesting to note that Granger's definition is also specified in terms of the DGP (cf Granger 2003) and is closely related to the direct structural causality. Like direct structural causality, Granger's definition covers only direct causal relationships. For example, if X affects Y only via a third series Z , then $\mathcal{I}_{-X}^*(t)$ comprises the past values of Z and Y_{t+1} is independent from the past values of X given $\mathcal{I}_{-X}^*(t)$. The following result by White and Lu (2010) shows that in the absence of latent variables Granger causality can be interpreted as direct structural causality.

Proposition 2.5. *Suppose that X and Y are generated by the DGP in (2.3) and assume additionally that $U_{y,t} \perp\!\!\!\perp X^t \mid Y^t, U_x^t$. If X does not directly structurally cause Y , then $Y_{t+1} \perp\!\!\!\perp X^t \mid Y^t, Z^t$.*

Since the process Y_{t+1} is supposed to be generated solely from the variables in Y^t , X^t , and Z^t , the conditioning set coincides with $\mathcal{I}_{-X}^*(t)$. Thus, direct structural noncausality implies (2.4). Like in the case of intervention causality, the converse implication is generally untrue. As an example, suppose that the process Y is generated by

$$Y_t = q_t(X_{t-1}, U_{1,t}, U_{2,t}) = \frac{X_{t-1}}{\sqrt{1 + X_{t-1}^2}} U_{1,t} + \frac{1}{\sqrt{1 + X_{t-1}^2}} U_{2,t},$$

while $U_{1,t}$, $U_{2,t}$, and X_t are all independent and normally distributed with mean zero and variance σ^2 . Furthermore, we assume that $U_{1,t}$ and $U_{2,t}$ are unobservable. Then $Y_t \mid X^{t-1} \sim \mathcal{N}(0, \sigma^2)$, which implies that X does not cause Y in the meaning of Granger. However, this argument raises the question what information should belong to $\mathcal{I}^*(t)$. Namely, if we add at least one of the variables U_1 and U_2 to the information set, Y_t becomes dependent on X^t . Since Y_t is thought to be generated

by its equation, the variables $U_{1,t}$ and $U_{2,t}$ must exist and therefore should belong to $\mathcal{I}^*(t)$.

If we—like White and Lu (2010)—not willing to accept $U_{1,t}$ and $U_{2,t}$ as separate entities that in principle should belong to $\mathcal{I}^*(t)$, does the difference between the two definitions of causality matter in practice? We think not as the difference basically can be viewed as counterfactual. More precisely, let $Y_t^* = q_t(x^*, U_{1,t}, U_{2,t})$ be the value we would have got for Y_t if we had set the value of X_{t-1} to x^* (so we use the same realizations of $U_{1,t}$ and $U_{2,t}$). Then $Y_t - Y_t^* \neq 0$ almost surely. However, this cannot be tested from data as Y^* is counterfactual and cannot be observed together with Y_t ; furthermore, the average $\mathbb{E}(Y_t) - \mathbb{E}(Y_t^*)$ is zero.

We end our discussion of the relationship between the two concepts by remarking that White and Lu (2010) showed that Granger’s notion of causality is equivalent to a slightly weaker notion of direct almost sure causality. For further details, we refer to White and Lu (2010).

It is clear that the above definition of causality usually cannot be used with actual data. In practice, only the background knowledge available at time t can be incorporated into an analysis. Therefore, the definition must be modified to become operational. Suppose that the process $V = (X, Y, Z)$ has been observed. Substituting the new information sets $\sigma\{X^t, Y^t, Z^t\}$ and $\sigma\{Y^t, Z^t\}$ for $\mathcal{I}_V(t)$ and $\mathcal{I}_{-X}^*(t)$, respectively, we obtain the following modified version of the above definition (Granger 1980, 1988).

Definition 2.6 (Granger causality). The series X is *Granger-noncausal* for the series Y with respect to $V = (X, Y, Z)$ if

$$Y_{t+1} \perp\!\!\!\perp X^t \mid Y^t, Z^t. \quad (2.5)$$

Otherwise we say that X Granger-causes Y with respect to V .

Granger (1980, 1988) used the term “ X is a prima facie cause of Y ” to emphasize the fact that a cause in the sense of Granger causality must be considered only as a potential cause. This, however, is not to say that the concept of Granger causality is completely useless for causal inference. Indeed, as we discuss in Section 6, it can be an essential tool for recovering (at least partially) the causal structure of a time series X if used in the right way.

Furthermore, we note that the above definition in terms of conditional independence is usually referred to as *strong Granger causality*; other existing notions are *Granger causality in mean* (Granger 1980, 1988) and *linear Granger causality* (Hosoya 1977, Florens and Mouchart 1985).

It is clear from the general definition given above that Granger intended the information to be chosen as large as possible including all available and possibly relevant variables. Despite of this, most (econometric) textbooks (e.g., Lütkepohl 1993) introduce Granger causality only in the bivariate case. This has led to some confusion about a multivariate definition of Granger causality (e.g., Kamiński et al. 2001).

With the above definition of a causal effect in terms of interventions a first connection between Granger noncausality and the effect of an intervention can be established as follows.

Corollary 2.7. Consider a multivariate time series $V = (X, Y, Z)$ and an individual intervention $\sigma_a(t) = s$ satisfying (2.1) and (2.2). If X is Granger-noncausal for

Y with respect to V , that is, $Y_{t+1} \perp\!\!\!\perp X^t \mid Y^t, Z^t$, then there is no causal effect of intervening in X_t on Y_{t+1} .

The proof of the corollary, which can be found in Eichler and Didelez (2010), relies on both conditions (2.1) and (2.2), underpinning that Granger-noncausality on its own is not enough to make statements about the effect of interventions. However, we do not need the whole V to be observable in the above corollary; the system V with respect to which X is Granger-noncausal for Y can therefore include latent time series if this helps to justify the stability assumptions.

2.4. Sims causality

The econometric literature features other less well known probabilistic notions of causality that are related to Granger causality. Among these, the concept introduced by Sims (1972) seems of most interest. In contrast to Granger causality, it takes in account not only direct but also indirect causal effects. Thus it can be seen as a concept for total causality. The following definition is a slight variation proposed by Florens and Mouchart (1982).

Definition 2.8 (Sims noncausality). The process X is *Sims-noncausal* for another process Y with respect to the process $V = (X, Y, Z)$ if

$$\{Y_{t'} \mid t' > t\} \perp\!\!\!\perp X_t \mid X^{t-1}, Y^t, Z^t$$

for all $t \in \mathbb{Z}$.

Now suppose that we are interested in the causal effect of an intervention $\sigma = s$ in X_t on $Y_{t'}$ for some $t' > t$. Let $V = (X, Y, Z)$ be a process such that the stability assumptions (I1) to (I4) in Section 2.1 are satisfied. If X is Sims-noncausal for Y with respect to V , it follows from (I1) that

$$Y_{t+h} \perp\!\!\!\perp \sigma_t, X_t \mid X^{t-1}, Y^t, Z^t$$

and furthermore by (I1) and (I3)

$$\begin{aligned} \mathbb{E}_s(g(Y_{t+h})) &= \mathbb{E}_s[\mathbb{E}_s(g(Y_{t+h}) \mid X^t, Y^t, Z^t)] = \mathbb{E}_s[\mathbb{E}_\emptyset(g(Y_{t+h}) \mid X^{t-1}, Y^t, Z^t)] \\ &= \mathbb{E}_\emptyset[\mathbb{E}_\emptyset(g(Y_{t+h}) \mid X^{t-1}, Y^t, Z^t)] = \mathbb{E}_\emptyset(g(Y_{t+h})). \end{aligned}$$

This suggests the following result which relates the concepts of intervention causality and Sims causality. The details of the proof are omitted.

Proposition 2.9. *Consider a multivariate time series $V = (X, Y, Z)$ and an individual intervention $\sigma_a(t) = s$ satisfying (2.1) and (2.2). Then X is Sims-noncausal for Y with respect to V if and only if the average causal effect of $\sigma = s$ on $g(Y_{t'})$ is zero for all measurable functions g and all $t' > t$.*

3. GRAPHICAL REPRESENTATIONS FOR TIME SERIES

In this section, we briefly review the two main approaches for representing dependences among multiple time series by graphs. For simplicity, we consider only the case of stationary Gaussian processes. Therefore, throughout this section, we assume that $X = (X_t)_{t \in \mathbb{Z}}$ with $X_t = (X_{1,t}, \dots, X_{d,t})'$ is a stationary Gaussian process with mean zero and covariances $\Gamma(u) = \mathbb{E}(X_t X_t')$; furthermore we make the following assumption.

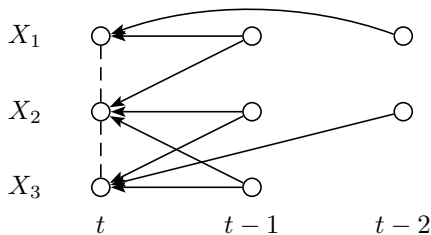


FIGURE 3.1. Graphical representation of the conditional distribution of X_t given its past X^{t-1} for the process in Example 3.2.

Assumption 3.1. The spectral density matrix

$$f(\lambda) = \frac{1}{2\pi} \sum_{u=-\infty}^{\infty} \Gamma(u) e^{-i\lambda u}$$

of X exists, and its eigenvalues are bounded and bounded away from zero uniformly for all $\lambda \in [-\pi, \pi]$.

This technical assumption ensures that the process has a mean-square convergent autoregressive representation

$$X_t = \sum_{u=1}^{\infty} \Phi(u) X_{t-u} + \varepsilon_t, \quad (3.1)$$

where $\Phi(u)$ is a square summable sequence of $d \times d$ matrices and $\varepsilon = (\varepsilon_t)_t$ is a Gaussian white noise process with non-singular covariance matrix Σ .

3.1. Conditional distributions and chain graphs

The simplest approach to visualize the dependence structure of a vector autoregressive process is to construct a graph from the conditional distribution of X_t given its past values X^{t-1} for fixed t . More precisely, representing the variables $X_{a,t}$ by vertices a_t , two distinct vertices a_t and b_t are connected by an undirected (dashed) edge whenever $\Sigma_{ab} \neq 0$. Additionally, vertices b_{t-k} represent the lagged instances $X_{b,t-k}$ of the variable X_b and are connected to nodes a_t whenever $\Phi_{ab}(k) \neq 0$. Any vertices b_{t-k} for which $\Phi_{ab}(k) = 0$ for all a are omitted from the graph. The resulting *conditional distribution graph* gives a concise picture of the autoregressive structure of the process X .

Example 3.2. For an illustration, we consider the trivariate process X given by

$$\begin{aligned} X_{1,t} &= \phi_{11}(1) X_{1,t-1} + \phi_{11}(2) X_{1,t-2} + \varepsilon_{1,t}; \\ X_{2,t} &= \phi_{22}(1) X_{2,t-1} + \phi_{21}(1) X_{1,t-1} + \phi_{23}(1) X_{3,t-1} + \varepsilon_{2,t}; \\ X_{3,t} &= \phi_{33}(1) X_{3,t-1} + \phi_{32}(1) X_{2,t-1} + \phi_{32}(2) X_{2,t-2} + \varepsilon_{3,t}, \end{aligned}$$

where ε_1 , ε_2 , and ε_3 are white noise processes with $\text{corr}(\varepsilon_{1,t}, \varepsilon_{3,t}) = 0$. The corresponding conditional distribution graph is depicted in Figure 3.1.

As can be seen from the example, graphs of this type give a concise picture of the autoregressive structure of the process X . The disadvantage of such graphs is that they encode only a very limited set of conditional independences. In fact, the only conditional dependences that can be derived from such graphs are concerned with the relationships among the variables $X_{a,t}$ conditionally on the complete history

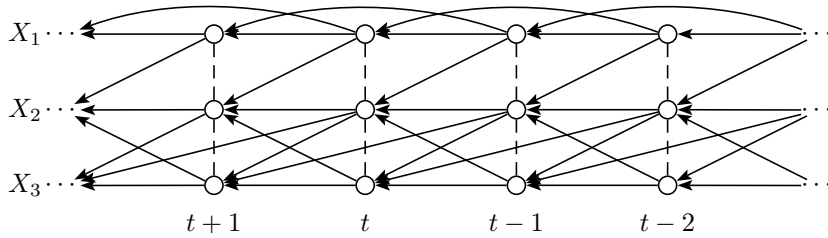


FIGURE 3.2. Time series chain graph for the vector autoregressive process X_V in Example 3.2.

X^{t-1} of the process. Therefore, such or similar graphs have been used mainly for investigating the causal ordering of the variables $X_{1,t}, \dots, X_{d,t}$ (Swanson and Granger 1997, Reale and Tunnicliffe Wilson 2001, Demiralp and Hoover 2003, Moneta and Spirtes 2006, Moneta 2007, Hyvärinen et al. 2010).

In general, we also want to be able to investigate, for instance, the conditional independences that hold among the variables of a subprocess $Y \subseteq X$. For this, we must also include the dependences among lagged variables in the graph. This leads to graphs with an infinite number of variables a_t with $a \in \{1, \dots, d\}$ and $t \in \mathbb{Z}$. Because of stationarity, edges in this graph are translation invariant and can be obtained by using the above conditional distribution graph as a blueprint. The resulting graph is called *time series chain graph* (Lynggaard and Walther 1993, Dahlhaus and Eichler 2003).

While time series chain graphs (or time series DAGs in the case of structural vector autoregressions) allow a full discussion of the Markov properties that are entailed by the autoregressive structure of the process, the graphs become complicated and infeasible even for small numbers of variables unless the dependence structure is very sparse. Figure 3.2 shows the time series chain graph for the process in Example 3.2.

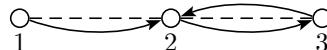
3.2. Path diagrams and Granger causality graphs

In order to avoid the complexity of time series chain graphs, Eichler (2007) proposed to consider path diagrams $G = (V, E)$ with $V = \{1, \dots, d\}$ in which every vertex $v \in V$ corresponds to one complete component series $(X_{v,t})$ while edges—arrows and lines—between vertices indicate non-zero coefficients in the autoregressive representation of the process.

Definition 3.3 (Path diagram). Let $X = X_V$ be a stationary Gaussian process with autoregressive representation (3.1). Then the *path diagram* associated with X is the graph $G = (V, E)$ with vertex set V and edge set E such that for distinct vertices $a, b \in V$

- (i) $a \rightarrow b \notin E \Leftrightarrow \Phi_{ba}(u) = 0 \quad \forall u \in \mathbb{N}$;
- (ii) $a \text{ --- } b \notin E \Leftrightarrow \Sigma_{ab} = \Sigma_{ba} = 0$.

This reduction in complexity compared to time series chain graphs is paid for by a loss of information. While path diagrams only encode whether or not one variable X_a depends on the past of another variables X_b , the time series chain graphs additionally yields the lags at which past instances of X_b have an effect on $X_{a,t}$. However, when concerned with problems of causal inference, algorithms for identifying causal relationships become infeasible if the level of detail is too high.

FIGURE 3.3. Path diagram for the vector autoregressive process X_V in Example 3.2.

Therefore, path diagrams provide a reasonable compromise between time series chain graphs and even simpler graphs such as partial correlation graphs Dahlhaus (2000).

The use of path diagrams for causal inference is based on the following lemma that links edges in the graph with Granger-causal relationships between the corresponding components.

Lemma 3.4. *Let $G = (V, E)$ be the path diagram associated with a stationary Gaussian process X satisfying Assumption 3.1. Then*

- (i) *the directed edge $a \rightarrow b$ is absent in the graph if and only if X_a is Granger-noncausal for X_b with respect to the full process X , that is, $X_{b,t+1} \perp\!\!\!\perp X_a^t \mid X_{-a}^t$.*
- (ii) *the undirected edge $a --- b$ is absent in the graph if and only if X_a and X_b are not contemporaneously correlated with respect to the full process, that is, $X_{a,t+1} \perp\!\!\!\perp X_{b,t+2} \mid X^t$.*

Because of this result, we will call the path diagram associated with a process X also the *Granger causality graph* of X . We note that in the more general case of non-linear non-Gaussian processes X , the definition of such graphs is entirely based on the concept of Granger causality. For details we refer to Eichler (2011).

3.3. Markov properties for Granger causality graphs

Under the assumptions imposed on the process X_V , more general Granger-causal relationships than those in Lemma 3.4 can be derived from the path diagram associated with X_V . This global Markov interpretation is based on a path-oriented concept of separating subsets of vertices in a mixed graph, which has been used previously to represent the Markov properties of linear structural equation systems (e.g. Spirtes et al. 1998, Koster 1999). Following Richardson (2003) we will call this notion of separation in mixed graphs *m-separation*.

More precisely, let $G = (V, E)$ be a mixed graph and $a, b \in V$. A *path* π in G is a sequence $\pi = \langle e_1, \dots, e_n \rangle$ of edges $e_i \in E$ with an associated sequence of nodes v_0, \dots, v_n such that e_i is an edge between v_{i-1} and v_i . The vertices v_0 and v_n are the *endpoints* while v_1, \dots, v_{n-1} are the *intermediate vertices* of the path. Notice that paths may be self-intersecting since we do not require that the vertices v_i are distinct.

An intermediate vertex c on a path π is said to be an *m-collider* on π if the edges preceding and succeeding c both have an arrowhead or a dashed tail at c (i.e. $\rightarrow c \leftarrow$, $\rightarrow c ---$, $--- c \leftarrow$, $--- c ---$); otherwise c is said to be an *m-noncollider* on π . A path π between a and b is said to be *m-connecting* given a set C if

- (i) every *m-noncollider* on π is not in C and
- (ii) every *m-collider* on π is in C ;

otherwise we say that π is *m-blocked* given C . If all paths between a and b are *m-blocked* given C , then a and b are said to be *m-separated* given C . Similarly, two

sets A and B are said to be m -separated given C if for every pair $a \in A$ and $b \in B$, a and b are m -separated given C .

With this notion of separation, it can be shown that path diagrams for multivariate time series have a similar Markov interpretation as path diagrams for linear structural equation systems (cf Koster 1999). But since each vertex $v \in V$ corresponds to a complete process X_v , separation in the path diagram encodes a conditional independence relation among complete subprocesses of X_V .

Proposition 3.5. *Let X_V be a stationary Gaussian process that satisfies Assumption 3.1, and let G be its path diagram. Then, for all disjoint $A, B, S \subseteq V$,*

$$A \bowtie_m B \mid S \text{ in } G \Rightarrow X_A \perp\!\!\!\perp X_B \mid X_S.$$

Derivation of such conditional independence statements requires that all paths between two sets are m -blocked. For the derivation of Granger-causal relationships, it suffices to consider only a subset of these paths, namely those having an arrowhead at one endpoint. For a formal definition, we say that a path π between a and b is b -pointing if it has an arrowhead at the endpoint b ; furthermore, a path between sets A and B is said to be B -pointing if it is b -pointing for some $b \in B$. Then, to establish Granger noncausality from X_A to X_B , it suffices to consider only all B -pointing paths between A and B . Similarly, a graphical condition for contemporaneous correlation can be obtained based on bi -pointing path, which have an arrowhead at both endpoints.

Definition 3.6. A stationary Gaussian process X_V is Markov for a graph $G = (V, E)$ if, for all disjoint subsets $A, B, C \subseteq V$, the following two conditions hold:

- (i) if every B -pointing path between A and B is m -blocked given $B \cup C$, then X_A is Granger-noncausal for X_B with respect to $X_{A \cup B \cup C}$;
- (ii) if the sets A and B are not connected by an undirected edge ($---$) and every bi -pointing path between A and B is m -blocked given $A \cup B \cup C$, then X_A and X_B are contemporaneously uncorrelated with respect to $X_{A \cup B \cup C}$.

With this definition, it can be shown that path diagrams for vector autoregressions can be interpreted in terms of such global Granger-causal relationships.

Theorem 3.7. *Let X_V be a stationary Gaussian process that satisfies Assumption 3.1, and let G be the associated path diagram. Then X_V is Markov for G .*

4. REPRESENTATION OF SYSTEMS WITH LATENT VARIABLES

The notion of Granger causality is based on the assumption that all relevant information is included in the analysis (Granger 1969, 1980). It is well known, that the omission of important variables can lead to temporal correlations among the observed components that are falsely detected as causal relationships. The detection of such so-called spurious causalities (Hsiao 1982) becomes a major problem when identifying the structure of systems that may be affected by latent variables.

Of particular interest will be spurious causalities of type I, where a Granger-causal relationship with respect to the complete process vanishes when only a subprocess is considered. Since Granger causality graphs are defined in terms of the pairwise Granger-causal relationships with respect to the complete process, they provide no

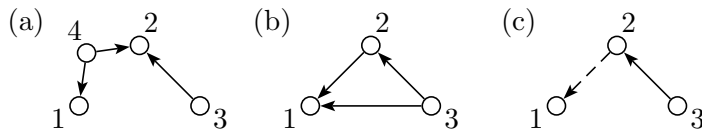


FIGURE 4.1. Graphical representations of the four-dimensional VAR(2) process in (4.1): (a) path diagram associated with $X_{\{1,2,3,4\}}$; (b) path diagram associated with $X_{\{1,2,3\}}$; (c) general path diagram for $X_{\{1,2,3\}}$.

means to distinguish such spurious causalities of type I from true causal relationships. To illustrate this remark, we consider the four-dimensional vector autoregressive process X with components

$$\begin{aligned} X_{1,t} &= \alpha X_{4,t-2} + \varepsilon_{1,t}, & X_{3,t} &= \varepsilon_{3,t}, \\ X_{2,t} &= \beta X_{4,t-1} + \gamma X_{3,t-1} + \varepsilon_{2,t}, & X_{4,t} &= \varepsilon_{4,t}, \end{aligned} \quad (4.1)$$

where $\varepsilon_i(t)$, $i = 1, \dots, 4$ are uncorrelated white noise processes with mean zero and variance one. The true dynamic structure of the process is shown in Fig. 4.1(a). In this graph, the 1-pointing path $3 \rightarrow 2 \leftarrow 4 \rightarrow 1$ is m -connecting given $S = \{2\}$, but not given the empty set. By Theorem 3.7, we conclude that X_3 is Granger-noncausal for X_1 in a bivariate analysis, but not necessarily in an analysis based on $X_{\{1,2,3\}}$.

Now suppose that variable X_4 is latent. Simple derivations show (cf Eichler 2005) that the autoregressive representation of $X_{\{1,2,3\}}$ is given by

$$\begin{aligned} X_{1,t} &= \frac{\alpha\beta}{1+\beta^2} X_{2,t-1} + \frac{\alpha\beta\gamma}{1+\beta^2} X_{3,t-2} + \tilde{\varepsilon}_{1,t}, \\ X_{2,t} &= \gamma X_{3,t-1} + \tilde{\varepsilon}_{2,t}, \\ X_{3,t} &= \varepsilon_{3,t}, \end{aligned}$$

where $\tilde{\varepsilon}_{2,t} = \varepsilon_{2,t} + \beta X_{4,t-1}$ and

$$\tilde{\varepsilon}_{1,t} = \varepsilon_{1,t} - \frac{\alpha\beta}{1+\beta^2} \varepsilon_{2,t-1} + \frac{\alpha}{1+\beta^2} X_{4,t-2}.$$

The path diagram associated with $X_{\{1,2,3\}}$ is depicted in Fig. 4.1(b). In contrast to the graph in Fig. 4.1(a), this path diagram contains an edge $3 \rightarrow 1$ and, thus, does not entail that X_3 is Granger-noncausal for X_1 in a bivariate analysis.

As a response to such situations, two approaches have been considered in the literature. One approach suggests to include all latent variables explicitly as additional nodes in the graph (e.g., Pearl 2000, Eichler 2007); this leads to models with hidden variables, which can be estimated, for example, by application of the EM algorithm (e.g., Boyen et al. 1999). For a list of possible problems with this approach, we refer to Richardson and Spirtes (2002, §1).

The alternative approach focuses on the conditional independence relations among the observed variables; examples of this approach include linear structural equations with correlated errors (e.g. Pearl 1995, Koster 1999) and the maximal ancestral graphs by Richardson and Spirtes (2002). In the time series setting, this approach has been discussed by Eichler (2005), who considered path diagrams in which dashed

TABLE 4.1. Creation of edges by marginalizing over i .

Subpath π in G	Associated edge e_π in $G^{\{i\}}$
$a \rightarrow i \rightarrow b$	$a \rightarrow b$
$a \dashrightarrow i \rightarrow b$	$a \dashrightarrow b$
$a \dashdash i \rightarrow b$	$a \dashrightarrow b$
$a \leftarrow i \rightarrow b$	$a \leftrightarrow b$
$a \leftdash i \rightarrow b$	$a \leftrightarrow b$

edges represent associations due to latent variables. For the trivariate subprocess $X_{\{1,2,3\}}$ in the above example, such a path diagram is depicted in Fig. 4.1(c).

Following this latter approach, we consider mixed graphs that may contain three types of edges, namely undirected edges (\dashdash), directed edges (\rightarrow), and dashed directed edges (\dashrightarrow). For the sake of simplicity, we also use $a \leftrightarrow b$ as an abbreviation for the triple edge $a \dashdash \leftrightarrow b$. Unlike path diagrams for autoregressions, these graphs in general are not defined in terms of pairwise Granger-causal relationships, but only through the global Markov interpretation according to Definition 3.6. To this end, we simply extend the concept of m -separation introduced in the previous section by adapting the definition of m -noncolliders and m -colliders. Let π be a path in a mixed graph G . Then an intermediate vertex n is called an m -noncollider on π if at least one of the edges preceding and succeeding c on the path is a directed edge (\rightarrow) and has its tail at c . Otherwise, c is called an m -collider on π . With this extension, we leave all other definition such as m -separation or pointing paths unchanged.

4.1. Marginalization

The main difference between the class of mixed graphs with directed (\rightarrow) and undirected (\dashdash) edges and the more general class of mixed graphs that has been just introduced is that the latter class is closed under marginalization. This property makes it suitable for representing systems with latent variables.

Let $G = (V, E)$ be a mixed graph and $i \in V$. For every subpath $\pi = \langle e_1, e_2 \rangle$ of length 2 between vertices $a, b \in V \setminus \{i\}$ such that i as an intermediate vertex and an m -noncollider on π , we define an edge e_π according to Tab. 4.1. Let $A^{\{i\}}$ the set of all such edges e_π . Furthermore, let $E^{\{i\}}$ be the subset of edges in E that have both endpoints in $V \setminus \{i\}$. Then we define $G^{\{i\}} = (V \setminus \{i\}, E^{\{i\}} \cup A^{\{i\}})$ as the graph obtained by marginalizing over $\{i\}$. Furthermore, for $L = \{i_1, \dots, i_n\}$ we set $G^L = ((G^{\{i_1\}})^{\{i_2\}} \dots)^{\{i_n\}}$, that is, we proceed iteratively by marginalizing over i_j , for $j = 1, \dots, n$. Similarly as in Koster (1999), it can be shown that the order of the vertices does not matter and that the graph G^L is indeed well defined.

We note that the graph G^L obtained by marginalizing over the set L in general contains self-loops. Simple considerations, however, show that G^L is Markov-equivalent to a graph \tilde{G}^L with all subpaths of the form $a \dashdash b \dashrightarrow b$ and $a \leftdash b \dashrightarrow$ replaced by $a \dashdash b$ and $a \leftrightarrow b$, respectively, and all self-loops deleted. It therefore suffices to consider mixed graphs without self-loops. We omit the details.

Now suppose that, for some subsets $A, B, C \subseteq V \setminus L$, π is an m -connecting path between A and B given S . Then all intermediate vertices on π that are in L must be m -noncolliders. Removing these vertices according to Table 4.1, we obtain a path

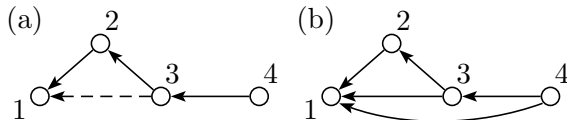


FIGURE 4.2. Two Markov equivalent graphs: (a) non-ancestral graph and (b) corresponding ancestral graph.

π' in G^L that is still m -connecting. Since the converse is also true, we obtain the following result.

Proposition 4.1. *Let $G = (V, E)$ be a mixed graph, and $L \subseteq V$. Then it holds that, for all distinct $a, b \in V \setminus L$ and all $C \subseteq V \setminus L$, every path between a and b in G is m -blocked given C if and only if the same is true for the paths in G^L . Furthermore, the same equivalence holds for all pointing path and for all bi-pointing paths.*

It follows that, if a process X_V is Markov for a graph G , the subprocess $X_{V \setminus L}$ is Markov for G^L , which encodes all relationships about $X_{V \setminus L}$ that are also encoded in G .

We note that insertion of edges according to Tab. 4.1 is sufficient but not always necessary for representing the relations in the subprocess $X_{V \setminus L}$. This applies in particular to the last two cases in Tab. 4.1. For an example, we consider again the process (4.1) with associated path diagram in Fig. 4.1(a). By Tab. 4.1, the subpath $1 \leftarrow 4 \rightarrow 2$ should be replaced by $1 \leftrightarrow 2$, which suggests that X_1 Granger-causes X_2 (as does the path $1 \leftarrow 4 \rightarrow 2$ in the original path diagram), while in fact the structure can be represented by the graph in Fig. 4.1(c).

4.2. Ancestral graphs

For systems with latent variables, the set of Granger-causal relationships and contemporaneous independencies that hold for the observed process does not uniquely determine a graphical representation within the class of general path diagrams. As an example, Fig. 4.2 displays two graphs that are Markov equivalent, that is, they encode the same set of Granger-causal and contemporaneous independence relations among the variables. Therefore, the corresponding graphical models—models that obey the conditional independence constraints imposed by the graph—are statistically indistinguishable. This suggests to choose one unique representative for each Markov equivalence class and to restrict model selection to these.

Following Richardson and Spirtes (2002), one suitable choice are maximal ancestral graphs. For vertices $a, b \in V$, we say that a is an ancestor of b if $a = b$ or there exists a directed path $a \rightarrow \dots \rightarrow b$ in G . The set of ancestors of b is denoted by $\text{an}(b)$. Then $G = (V, E)$ is an ancestral graph if

$$a \in \text{an}(b) \Rightarrow a \dashrightarrow b \notin E. \quad (4.2)$$

for all distinct $a, b \in V$. We note that, in contrast to Richardson and Spirtes (2002), we do not require acyclicity (which is hidden in the time ordering). Furthermore, an ancestral graph G is maximal if addition of further edges changes the independence models; for details, we refer to Richardson and Spirtes (2002).

5. IDENTIFICATION OF CAUSAL EFFECTS

Suppose that we are interested in studying the causal effect of an intervention in X on future instances of Y . As we want to be sure about the effect of the intervention before actually intervening, the effect must be predicted from data obtained under the observational regime.

Following the definition of causal effects in section 2.1, let Z be a set of other relevant variables such that the process $V = (X, Y, Z)$ satisfies the stability assumptions (I1) to (I4). For simplicity, we consider the case of a simple intervention, that is, an intervention in one variable at a single point in time. Then, using the law of iterated expectation, the ACE can be written as

$$\text{ACE}_s = \mathbb{E}_s \mathbb{E}_s [\mathbb{E}_s (Y_{t+h} | X^t, Y^t, Z^t) | X^{t-1}, Y^t, Z^t].$$

Noting that under the simplifying assumptions conditions (2.1) and (2.2) hold, we can use invariance under change of regime to get

$$\text{ACE}_s = \mathbb{E}_\theta \mathbb{E}_s [\mathbb{E}_\theta (Y_{t+h} | X^t, Y^t, Z^t) | X^{t-1}, Y^t, Z^t].$$

If the full process V is observable, the causal effect can be estimated from data. For instance, if V is a Gaussian stationary process, the causal effect of setting X_t to x^* on $Y_{t'}$ is given by the corresponding coefficient in a linear regression of $Y_{t'}$ on V^t .

The problem is that usually not all variables that are deemed as relevant and included in Z are observable. Therefore, suppose that only the subprocess \tilde{Z} of Z is observed. In case that the desired intervention is conditional as in Definition 2.1 (iii) and (iv), we assume that the conditioning set C is a subset of the variable in (X, Y, \tilde{Z}) . The following result states sufficient conditions under which the above derivation of the ACE with Z replaced by \tilde{Z} continues to hold. This so-called back-door criterion has been established first by Eichler and Didelez (2007) and reflects what is known in epidemiology as adjusting for confounding. The name is due to the graphical way of checking this criterion. For a more detailed discussion including proofs of the presented results, we refer to Eichler and Didelez (2010). Chapter 8 of this book by Berzuini, Dawid and Didelez, the back-door criterion

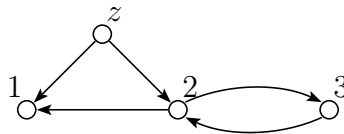
Theorem 5.1 (Back-door criterion). *Suppose that the process $V = (X, Y, Z)$ satisfies assumptions (2.1) and (2.2) hold and that for some $S \subseteq Z$*

$$Y_{t+h} \perp\!\!\!\perp \sigma_t | X^t, Y^t, S^t \quad (5.1)$$

for all $h > 0$. Then $\tilde{V} = (X, Y, S)$ identifies the effect of $\sigma_t = s$ on Y_{t+h} for all $h > 0$, and the average causal effect ACE_s is given by

$$\mathbb{E}_s Y_{t+h} = \mathbb{E}_\theta \mathbb{E}_s [\mathbb{E}_\theta (Y_{t+h} | X^t, Y^t, S^t) | X^{t-1}, Y^t, S^t]. \quad (5.2)$$

In (5.2) we can estimate $\mathbb{E}_\theta (Y_{t+h} | X^t, Y^t, S^t)$ from observational data, while the second expectation is with respect to the interventional distribution, which is fully known. The outer expectation is again observational. Hence, provided that (X, Y, S) has been observed, we can use the above to estimate the causal effect ignoring any variables that are in Z but not in S . Dawid (2002) calls such a set (X, Y, S) “sufficient covariates” or “de-confounder”. Note that under the stability assumptions, $V = (X, Y, Z)$ always identifies the causal effect due to condition (2.2). In this sense we could say that the whole system V contains all “relevant” variables or components to identify an individual causal effect of X_t on Y_{t+h} . Note, however, that if an

FIGURE 5.1. Mixed graph associated with the processes X and Z in Example 5.3.

intervention in a different variable \tilde{X} is considered, a different system \tilde{V} might be required to justify the stability assumptions with respect to this intervention.

In the case of ordinary multivariate distribution, the back-door criterion has an intuitive graphical representation which yields Using these graphs or path diagrams, we devise graphical rules to check if a set $S \subset V$ exists that satisfies the back-door or front-door criteria. This will provide us with further interesting insights into the relation between Granger-causality and intervention causality.

Theorem 5.2 (Back-door criterion). *Consider a multivariate time series X_V that obeys the global Markov properties for a graph G . Furthermore, assume that a is an ancestor of b ($a \in \text{an}(b)$).*

- (i) *Assumption (5.1) of Theorem 5.1 is satisfied if all $\text{an}(b)$ -pointing back-door paths between a and $\text{an}(b)$ are m -blocked given S .*
- (ii) *The minimal set S satisfying (i) is given by $S = \{a, b\} \cup \text{pa}(a) \cup D$, where D is the set of all nodes v such that there is a back-door path from node a to v for which all intermediate nodes are m -colliders and all intermediate nodes as well as v itself are ancestors of b .*

The following example illustrates the application of the graphical back-door criterion. A more complex example with sequential interventions can be found in Eichler and Didelez (2010).

Example 5.3. Consider the following trivariate Gaussian process X with

$$\begin{aligned} X_{1,t} &= \alpha_1 Z_{t-2} + \beta_{12} X_{2,t-1} + \varepsilon_{1,t}, \\ X_{2,t} &= \alpha_2 Z_{t-1} + \beta_{23} X_{3,t-1} + \varepsilon_{2,t}, \\ X_{3,t} &= \beta_{32} X_{2,t-1} + \varepsilon_{3,t}, \end{aligned}$$

where Z and ε_i , $i = 1, 2, 3$, are independent Gaussian white noise processes with mean 0 and variance σ^2 . The corresponding graph is shown in Figure 5.1.

Now suppose that we are interested in the effect of an intervention s setting $X_{3,t}$ to x_3^* on $X_{1,t+2}$. If both X and Z have been observed, the equations for the full model immediately yield

$$\mathbb{E}_s X_{1,t+2} = \beta_{12} \beta_{23} x_3^*.$$

If only X has been observed while the process Z takes the role of an unobserved variable, the ACE can be still computed. First, straightforward calculations show that X has the autoregressive representation

$$\begin{aligned} X_{1,t} &= \left(\frac{\alpha_1 \alpha_2}{1 + \alpha_2^2} + \beta_{12} \right) X_{2,t-1} - \frac{\alpha_1 \alpha_2 \beta_{23}}{1 + \alpha_2^2} X_{3,t-2} + \tilde{\varepsilon}_{1,t}, \\ X_{2,t} &= \beta_{23} X_{3,t-1} + \tilde{\varepsilon}_{2,t}, \\ X_{3,t} &= \beta_{32} X_{2,t-1} + \tilde{\varepsilon}_{3,t}, \end{aligned}$$

where $\tilde{\varepsilon}_i$, $i = 1, 2, 3$, are again independent zero mean Gaussian white noise processes and independent of the other X components. To apply the back-door criterion in Theorem 5.2, we note that in Figure 5.1 every pointing back-door path between 3 and some other node v is bi-pointing starting with the edge $3 \leftarrow 2$ and hence is m -blocked given $S = \{1, 2, 3\}$ because 2 is a noncollider. Thus, S identifies the effect of $X_{3,t}$ on $X_{1,t+2}$ and the average causal effect can be obtained from the above autoregressive representation of X as

$$\mathbb{E}_s X_{1,t+2} = \phi_{13}^{(2)}(1) = \phi_{12}(1)\phi_{23}(1) + \phi_{13}(2) = \beta_{12}\beta_{23}x_3^*.$$

Similarly, the example also shows that the effect of $X_{3,t}$ on $X_{1,t+2}$ is not identified if Z instead of X_2 is observed as there exists a back-door path $3 \leftarrow 2 \rightarrow 1$ that is not blocked by z .

We close our discussion of the identification of causal effects by an example that points out some limitation of the theory due to serial correlation usually present in time series.

Example 5.4. Consider a four-dimensional autoregressive process of the form

$$\begin{aligned} X_{1,t} &= \phi_{11} X_{1,t-1} + \phi_{12} X_{2,t-1} + \varepsilon_{1,t}, \\ X_{2,t} &= \phi_{22} X_{2,t-1} + \varepsilon_{2,t}, \\ X_{3,t} &= \phi_{32} X_{2,t-1} + \phi_{33} X_{3,t-1} + \varepsilon_{3,t}, \\ X_{4,t} &= \phi_{41} X_{1,t-1} + \phi_{43} X_{3,t-1} + \phi_{44} X_{4,t-1} + \varepsilon_{4,t}, \end{aligned}$$

where $(\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4)$ is a Gaussian white noise process with mean zero and covariance matrix $\sigma^2 I$. The path diagram is shown in Figure 5.2(a).

Suppose that we are interested in the effect of an intervention s setting $X_{1,t}$ to x^* on $X_{4,t+2}$. If all variables are observed, the causal effect of $X_{1,t}$ on $X_{4,t+2}$ is identified and given by

$$\mathbb{E}_s X_{4,t+2} = (\phi_{41} \phi_{11} + \phi_{44} \phi_{41}) x^*.$$

Similarly, knowledge of X_2 is sufficient for identification of the causal effect as every back-door path is m -blocked given $S = \{1, 2, 4\}$. In contrast, knowledge of only (X_1, X_3, X_4) is not sufficient as the path $1 \leftarrow 2 \rightarrow 1 \rightarrow 4$ is unblocked.

To understand why this path induces a confounding association between $X_{1,t}$ and $X_{4,t+2}$, we consider explicitly the instances at the relevant points in time along the path. Thus, the above path $1 \leftarrow 2 \rightarrow 1 \rightarrow 4$ corresponds to the path

$$X_{1,t} \leftarrow X_{2,t-1} \rightarrow X_{2,t+1} \rightarrow X_{1,t+1} \rightarrow X_{4,t+2}.$$

The problem is that every variable that directly influences a variable of interest acts as a confounder for all causal links from this variable of interest to other variables due to serial correlation.

The problem can be resolved if additional information about the process constrains the dynamic dependence structure further. For example, if $\phi_{22} = 0$ for the above process X , the causal effect of $X_{1,t}$ on $X_{4,t+1}$ is identified by (X_1, X_3, X_4) . However, a graphical criterion for identifiability must be based on the full time series chain graph depicted in Figure 5.2(b) as the Granger causality graph cannot encode such additional specific constraints.

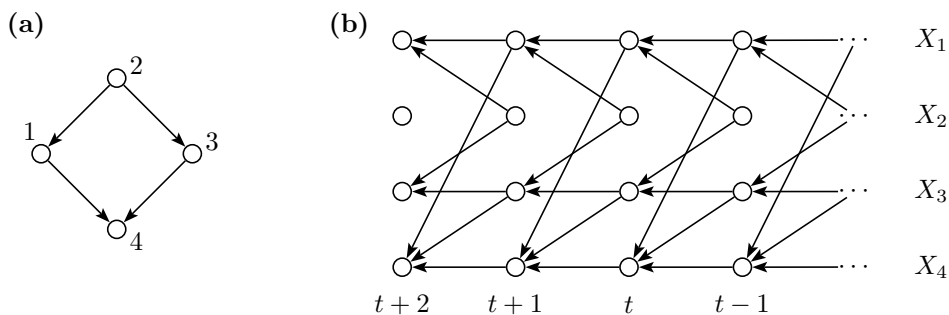


FIGURE 5.2. (a) Granger causality graph and (b) time series chain graph (with additional imposed constraint $\phi_{22} = 0$) for the process X in Example 5.4.

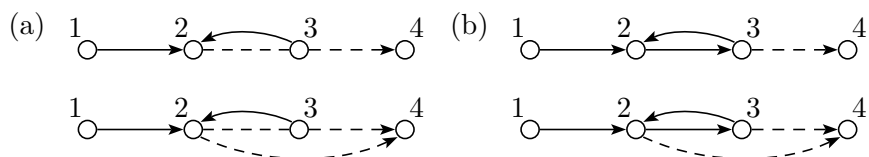


FIGURE 6.1. Inducing paths: (a) Dynamic ancestral graph with an inducing 4-pointing path between 2 and 4 and its corresponding Markov equivalent maximal dynamic ancestral graph. (b) Maximal dynamic ancestral graph with an inducing 4-pointing path between 2 and 4; the graph with additional edge $2 \dashrightarrow 4$ is not Markov equivalent.

6. LEARNING CAUSAL STRUCTURES

There are two major approaches for learning causal structures: One approach utilizes constraint-based search algorithms such as the Fast Causal Inference (FCI) algorithm (Spirtes et al. 2001) while the other consists of score-based model selection. Both approaches have been adapted to the time series case (Eichler 2009, 2010). In this and the next section, we briefly discuss the two approaches.

The first approach requires as input the set of all Granger-noncausal relations and contemporaneous independences that hold for the process. Usually, this will be accomplished by fitting vector autoregressions to all subseries X_S of X_V . Then the constraint-based search tries to find a graph that matches the empirically determined conditional independences. It usually consists of two steps:

1. identification of adjacencies of the graph;
2. identification of the type and the orientation of edges whenever possible.

In the case of ancestral graphs, the first step makes use of the fact that for every ancestral graph there exists a unique Markov-equivalent maximal ancestral graph (MAG), in which every missing edge corresponds to a conditional independence relation among the variables. Here an ancestral graph G is said to be maximal if addition of further edges would change the Markov equivalence class. MAGs are closely related to the concept of inducing paths; in fact, Richardson and Spirtes (2002) used inducing paths to define MAGs and then showed the maximality property.

Definition 6.1. In a (dynamic) ancestral graph, a path π between two vertices a and b is called an inducing path if every intermediate vertex on π is, firstly, a collider on π and, secondly, an ancestor of a or b .

Figures 6.1(a,b) give two examples of dynamic ancestral graphs, in which $2 \dashrightarrow 3 \dashrightarrow 4$ resp. $2 \rightarrow 3 \dashrightarrow 4$ are inducing 4-pointing paths. The graph in (b) shows that—unlike in the case of ordinary ancestral graphs—inducing paths may start with a tail at one of two vertices. As a consequence, insertion of an edge $2 \dashrightarrow 4$ or $2 \rightarrow 4$ changes the encoded Granger-causal relationships: while the upper graph implies that X_1 is Granger-noncausal for X_4 with respect to $X_{\{1,2,3,4\}}$, this is not true for the lower graph. It follows that the method used for identifying adjacencies in ordinary MAGs does not apply to dynamic ancestral graphs.

The problem can be solved by observing that m -connecting pointing paths not only encode Granger-causal relationships but, depending on whether they start with $a \rightarrow c$ or $a \dashrightarrow c$, also a related type of conditional independences. More precisely, we have the following result.

Proposition 6.2. *Suppose that a and b are not connected by an edge $a \rightarrow b$ or $a \dashrightarrow b$ or by a b -pointing inducing path starting with $a \leftarrow c$, $a \dashleftarrow c$, or $a \dashrightarrow c$. Then there exist disjoint subsets S_1, S_2 with $b \in S_1$ and $a \notin S_1 \cup S_2$ such that*

$$X_{a,t-k} \perp\!\!\!\perp X_{b,t+1} \mid X_{S_1}^t, X_{S_2}^{t-k}, X_a^{t-k-1}$$

for all $k \in \mathbb{N}$ and all $t \in \mathbb{Z}$.

The proof is based on the fact that inducing paths starting with an edge $a \rightarrow c$ or $a \dashrightarrow c$ only induce an association between $X_{a,t-k}$ and $X_{b,t+1}$ if one conditions on $X_{c,t-k+1}, \dots, X_{c,t}$. To block any other paths, we set S_2 to be the set of all intermediate vertices on all b -pointing inducing paths connecting a and b , and S_1 to be the set of all ancestors of a and b except a and S_2 .

This leads us to the following algorithm for the identification of the Markov equivalence classes of dynamic ancestral graphs. Here, we use dotted directed edges \dashrightarrow to indicate that the tail of the directed edge is (yet) undetermined.

Identification of adjacencies:

1. insert $a \dashrightarrow b$ whenever X_a and X_b are not contemporaneously independent with respect to X_V ;
2. insert $a \dashrightarrow b$ whenever
 - (i) X_a Granger-causes X_b with respect to X_S for all $S \subseteq V$ with $a, b \in S$ and
 - (ii) $X_{a,t-k}$ and $X_{b,t+1}$ are not conditionally independent given $X_{S_1}^t, X_{S_2}^{t-k}, X_a^{t-k-1}$ for some $k \in \mathbb{N}$, all $t \in \mathbb{Z}$, and all disjoint $S_1, S_2 \subseteq V$ with $b \in S_1$ and $a \notin S_1 \cup S_2$

Identification of tails:

1. *Colliders:*
Suppose that G does not contain $a \dashrightarrow b$, $a \rightarrow b$, or $a \dashrightarrow b$. If $a \dashrightarrow c \dashrightarrow b$ and X_a is Granger-noncausal for X_b with respect to X_S for some set S with $c \notin S$, replace $c \dashrightarrow b$ by $c \dashrightarrow b$.
2. *Non-colliders:*
Suppose that G does not contain $a \dashrightarrow b$, $a \rightarrow b$, or $a \dashrightarrow b$. If $a \dashrightarrow c \dashrightarrow b$ and X_a is Granger-noncausal for X_b with respect to X_S for some set S with $c \in S$, replace $c \dashrightarrow b$ by $c \rightarrow b$.
3. *Ancestors:*
if $a \in \text{an}(b)$ replace $a \dashrightarrow b$ by $a \rightarrow b$;

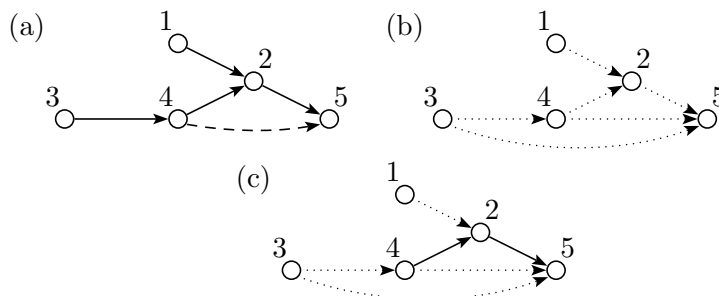


FIGURE 6.2. Identification of dynamic ancestral graphs: (a) underlying structure; (b) adjacencies; (c) identification of tails.

4. *Discriminating paths:*

A fourth rule is based on the concept of discriminating paths. For details, we refer to Ali et al. (2004).

We note that in contrast to the case of ordinary ancestral graphs only the tails of the dotted directed edges need to be identified. The positions of the arrow heads are determined by the time ordering of the Granger-causal relationships. The above algorithm probably can be complemented by further rules, see also Zhang and Spirtes (2005).

To illustrate the identification algorithm, we consider the graphs in Fig. 6.2. The original general path diagram is depicted in (a). Since 4 is an ancestor of 5, this graph is not ancestral. The adjacencies determined by the algorithm are shown in (b). Since X_1 does not Granger-cause X_5 with respect to X_V , we find that 2 and 5 are connected by $2 \rightarrow 5$. Similarly, X_3 is Granger-noncausal for X_2 with respect to $X_{\{2,3,4\}}$, which implies that the graph contains also the edge $4 \rightarrow 2$. No further tails can be identified; the final graph is given in (c).

7. A NEW PARAMETRIC MODEL

The second major approach for identifying causal structures requires fitting suitable constraint models that are Markov with respect to a given path diagram to the data. Searching over all possible path diagrams, the fit is evaluated by a model selection criterion such as AIC or BIC. The path diagram with the lowest score is taken as an estimate of the true causal structure.

In the case of simple path diagrams with no dashed directed edges, Definition 3.3 states the constraints that are required for the model being Markov with respect to the graph. In particular, we note that the undirected edges in the path diagram correspond to dependences in the innovation process ε . Eichler (2010) showed that by allowing correlations between lagged instances of the innovation process ε dashed directed edges can be incorporated into the framework of graphical vector autoregressive models.

More precisely, we consider multivariate stationary Gaussian processes $X = X_V$ that are given by

$$X_t = \sum_{u=1}^p \Phi(u) X_{t-u} + \varepsilon_t, \quad (7.1)$$

where $\varepsilon = \varepsilon_V$ is a stationary Gaussian process with mean zero and covariances

$$\text{cov}(\varepsilon_t, \varepsilon_{t-u}) = \begin{cases} \Omega(u) & \text{if } |u| \leq q \\ 0 & \text{otherwise} \end{cases} \quad (7.2)$$

for some $q \in \mathbb{N}$. The distribution of these processes can be parametrized by the vector $\theta = (\phi, \omega)$ with

$$\phi = \text{vec}(\Phi(1), \dots, \Phi(p))$$

and

$$\omega = \begin{pmatrix} \text{vech } \Omega(0) \\ \text{vec}(\Omega(1), \dots, \Omega(q)) \end{pmatrix},$$

where as usual the vec operator stacks the columns of a matrix and the vech operator stacks only the elements contained in the lower triangular submatrix. The parameter ϕ corresponds to the autoregressive structure of the process X while ω parametrizes the dependence structure of the innovation process ε . The following result states that suitable zero constraints on the parameters ϕ and ω ensure that the model is Markov with respect to a given mixed graph G .

Theorem 7.1. *Let $X = X_V$ be a stationary Gaussian process of the form (7.1) and (7.2), and suppose that Assumption 3.1 holds. Furthermore, let $G = (V, E)$ be a mixed graph such that the model parameters ϕ and ω satisfy the following constraints:*

- (i) if $a \rightarrow b \notin E$ then $\Phi_{ba}(u) = 0$ for all $u > 0$;
- (ii) if $a \dashrightarrow b \notin E$ then $\Omega_{ba}(u) = 0$ for all $u > 0$;
- (iii) if $a \dash\vdash b \notin E$ then $\Omega_{ba}(0) = \Omega_{ab}(0) = 0$.

Then X is Markov for G .

Sufficient conditions for X to satisfy Assumption 3.1 can be formulated in terms of the parameters ϕ and ω . For this, let $\Phi(z) = I - \Phi(1)z - \dots - \Phi(p)z^p$ the characteristic polynomial of the autoregressive part of X , and let $g_\omega(\lambda)$ be the spectral matrix of ε . Then, if $\det \Phi(z) \neq 0$ for all $z \in \mathbb{C}$ such that $|z| \leq 1$ and the eigenvalues of $g_\omega(\lambda)$ are bounded and bounded away from zero uniformly for all $\lambda \in [-\pi, \pi]$, Assumption 3.1 holds for the process X . For details on this and the proof of the above result we refer to Eichler (2010).

Autoregressive models with an innovation process having correlation structure of the form (7.2) are uncommon in time series analysis. However, it has been shown by Eichler (2010) that models of this form can be viewed as graphical multivariate ARMA models that satisfy the conditional independence constraints encoded by general path diagrams. We note that under the assumptions the error process ε has a moving average representation. Since the covariance function $\Omega(u)$ vanishes for $|u| > q$, it follows that the moving average representation is of order q and we have

$$\varepsilon_t = \sum_{u=0}^q \Theta(u) \eta_{t-u},$$

where $\Theta(0) = I$ and η is a Gaussian white noise process with mean zero and covariance matrix Σ . The coefficients of this moving average representation are uniquely determined by the equation system

$$\Omega(v) = \sum_{u=0}^{q-v} \Theta(u)' \Sigma \Theta(u+v), \quad (7.3)$$

which can be iteratively solved for $\Theta(u)$ and Σ (Tunncliffe Wilson 1972). It follows that the process X can be represented as a multivariate ARMA(p, q) process

$$X_t = \sum_{u=1}^p \Phi(u) X_{t-u} + \sum_{u=0}^q \Theta(u) \eta_{t-u}.$$

We note that, because of (7.3), the zero constraints on the matrices $\Omega(u)$ do not translate into equally simple constraints on the parameters $\Theta(1), \dots, \Theta(q)$ and Σ .

8. CONCLUDING REMARKS

In time series analysis, inference about causal relationships is still predominantly based on the concept of Granger causality as it can be simply formulated and tested in many standard time series models. Nevertheless, Granger causality has always been criticised as being not a true notion of causality due to the fact that it can lead to spurious causalities when confounding variables are not included in the analysis.

In this chapter, we have pointed out the usefulness of Granger causality as an empirical concept for causal inference by reviewing recent results that link the traditional concept of Granger causality to the modern graph-based approach to causality. Here it is helpful to distinguish between the theoretical notion of Granger causality as originally defined by Granger (1969, 1980, 1988) and its various empirical versions which are usually referred to as “Granger causality”. As Granger (2003) stated himself, these empirical versions should be seen as implications of Granger causality and not mistaken as the definition itself.

With this in mind, we have compared four definitions of causality, namely intervention, direct structural, Granger, and Sims causality used in the context of time series. We found that all four concepts are closely related and differ only in whether they are concerned with the total or only the direct causal effect. The subtle differences that do exist we think are not of practical relevance.

Although Granger causality does not require a scientific model—that is, no pre-knowledge about the system to analysed—it is limited in its scope by the statistical model used for the empirical analysis. The traditional framework to study Granger causality are vector autoregressive models (cf Eichler 2005, 2006). As a consequence, only linear relationships are covered by the model and any nonlinear cause-effect relationship might stay undetected. Although there exist general tests for nonlinear Granger causality that are based on nonparametric estimation methods (e.g. Su and White 2007, Bouezmarni et al. 2009, Huang 2009, Marinazzo et al. 2011) applying this on large scale systems usually is not feasible. Inference can be greatly facilitated if additional background knowledge about the system can be incorporated in the model. As an example, we mention the recent article by Valdes-Sosa et al. (2011), which lists various complex state-space models used for causal inference from brain-imaging data.

An important topic that we have not raised in our discussion is model uncertainty. The score-based approach yields a single model—specified by a path diagram—that minimizes some model selection criterion, but at the same time there are other models that have only a slightly larger score and therefore should be deemed competitive. Such models usually do not give a significantly worse fit if tested by an appropriate test (e.g. Vuong 1989). One possible solution to this model uncertainty could be model averaging but it is not immediately clear how to “average” path diagrams and how to draw conclusions from such an averaged path diagram. In

the constraint-based approach, model uncertainty enters through the set of Granger noncausality statements that serve as an input to the identification algorithm. As the set of statements is determined by a multitude of statistical tests, the empirical set deviates from the true set of statements that hold for the process. The uncertainty introduced in this way in the identification algorithm has been so far neglected.

Finally, we note that there is no reason to restrict the discussion to time-discrete stochastic processes. Indeed, Comte and Renault (1996) and Florens and Fougère (1996) extended the notion of Granger causality to time-continuous processes. Closely related to these definitions is the notion of local independence introduced earlier by Schweder (1970) and Aalen (1987) and used by Didelez (2008) to define graphical models for marked point processes and by Commenges and Gègout-Petit (2009) for causal inference.

REFERENCES

- Aalen, O. O. (1987). Dynamic modeling and causality. *Scandinavian Actuarial Journal* pp. 177–190.
- Ali, R. A., Richardson, T. S. and Spirtes, P. (2004). Markov equivalence for ancestral graphs. *Technical Report No. 466*, Department of Statistics, University of Washington.
- Berzuini, C. and Dawid, A. P. (2011). Inference about biological mechanism on the basis of epidemiological data. In C. Berzuini, P. Dawid and L. Bernardinelli (eds), *Causality: Statistical Perspectives and Applications*, Wiley, chapter 14.
- Bouezmarni, T., Rombouts, J. V. K. and Taamouti, A. (2009). A nonparametric copula based test for conditional independence with applications to Granger causality. *Economics Working Papers we093419*, Universidad Carlos III, Departamento de Economía.
- Boyan, X., Friedman, N. and Koller, D. (1999). Discovering the hidden structure of complex dynamic systems. *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann, San Francisco, pp. 91–100.
- Commenges, D. and Gègout-Petit, A. (2009). A general dynamical statistical model with causal interpretation. *JRSSB* **71**, 1–18.
- Comte, F. and Renault, E. (1996). Noncausality in continuous time models. *Econometric Theory* **12**, 215–256.
- Dahlhaus, R. (2000). Graphical interaction models for multivariate time series. *Metrika* **51**, 157–172.
- Dahlhaus, R. and Eichler, M. (2003). Causality and graphical models in time series analysis. In P. Green, N. Hjort and S. Richardson (eds), *Highly structured stochastic systems*, University Press, Oxford, pp. 115–137.
- Dawid, A. P. (2002). Influence diagrams for causal modelling and inference. *International Statistical Review* **70**, 161–189.
- Dawid, A. P. (2011). The decision-theoretic approach to causal inference. In C. Berzuini, P. Dawid and L. Bernardinelli (eds), *Causality: Statistical Perspectives and Applications*, Wiley, chapter 3.
- Dawid, A. P. and Didelez, V. (2005). Identifying the consequences of dynamic treatment strategies. *Technical Report 262*, Department of Statistical Science, University College London.
- Demiralp, S. and Hoover, K. D. (2003). Searching for the causal structure of a vector autoregression. *Oxford Bulletin of Economics and Statistics* **65**(Supplement), 745–767.
- Didelez, V. (2008). Graphical models for marked point processes based on local independence. *Journal of the Royal Statistical Society Series B* **70**, 245–264.

- Eichler, M. (2005). A graphical approach for evaluating effective connectivity in neural systems. *Philosophical Transactions of The Royal Society B* **360**, 953–967.
- Eichler, M. (2006). Graphical modelling of dynamic relationships in multivariate time series. In M. Winterhalder, B. Schelter and J. Timmer (eds), *Handbook of Time Series Analysis*, Wiley-VCH, pp. 335–372.
- Eichler, M. (2007). Granger causality and path diagrams for multivariate time series. *Journal of Econometrics* **137**, 334–353.
- Eichler, M. (2009). Causal inference from multivariate time series: What can be learned from Granger causality. In C. Glymour, W. Wang and D. Westerståhl (eds), *Proceedings from the 13th International Congress of Logic, Methodology and Philosophy of Science*, College Publications, London, pp. 481–496.
- Eichler, M. (2010). Graphical Gaussian modelling of multivariate time series with latent variables. *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, Journal of Machine Learning Research W&CP 9.
- Eichler, M. (2011). Graphical modelling of multivariate time series. *Probability Theory and Related Fields* (DOI:10.1007/s00440-011-0345-8).
- Eichler, M. and Didelez, V. (2007). Causal reasoning in graphical time series models. *Proceedings of the 23rd Conference on Uncertainty in Artificial Intelligence*.
- Eichler, M. and Didelez, V. (2010). On Granger-causality and the effect of interventions in time series. *Life Time Data Analysis* **16**, 3–32.
- Florens, J. P. and Fougère, D. (1996). Noncausality in continuous time. *Econometrica* **64**, 1195–1212.
- Florens, J. P. and Mouchart, M. (1982). A note on noncausality. *Econometrica* **50**, 583–591.
- Florens, J. P. and Mouchart, M. (1985). A linear theory for noncausality. *Econometrica* **53**, 157–175.
- Good, I. J. (1961). A causal calculus (I). *British Journal for the Philosophy of Science* **11**, 305–318.
- Good, I. J. (1962). A causal calculus (II). *British Journal of the Philosophy of Science* **12**, 43–51.
- Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* **37**, 424–438.
- Granger, C. W. J. (1980). Testing for causality, a personal viewpoint. *Journal of Economic Dynamics and Control* **2**, 329–352.
- Granger, C. W. J. (1988). Some recent developments in a concept of causality. *Journal of Econometrics* **39**, 199–211.
- Granger, C. W. J. (2003). Some aspects of causal relationships. *Journal of Econometrics* **112**, 69–71.
- Hosoya, Y. (1977). On the Granger condition for non-causality. *Econometrica* **45**, 1735–1736.
- Hsiao, C. (1982). Autoregressive modeling and causal ordering of econometric variables. *Journal of Economic Dynamics and Control* **4**, 243–259.
- Huang, M. (2009). *Essays on testing conditional independence*. Doctoral thesis, University of California, San Diego.
- Hyvärinen, A., Zhang, K., Shimizu, S. and Hoyer, P. (2010). Estimation of a structural vector autoregression model using non-Gaussianity. *Journal of Machine Learning Research* **11**, 1709–1731.
- Kamiński, M., Ding, M., Truccolo, W. A. and Bressler, S. L. (2001). Evaluating causal relations in neural systems: Granger causality, directed transfer function and statistical assessment of significance. *Biological Cybernetics* **85**, 145–157.

- Koster, J. T. A. (1999). On the validity of the Markov interpretation of path diagrams of Gaussian structural equations systems with correlated errors. *Scandinavian Journal of Statistics* **26**, 413–431.
- Lauritzen, S. L. (2001). Causal inference from graphical models. In O. E. Barndorff-Nielsen, D. R. Cox and C. Klüppelberg (eds), *Complex stochastic systems*, CRC Press, London, pp. 63–107.
- Lütkepohl, H. (1993). *Introduction to Multiple Time Series Analysis*. Springer, New York.
- Lynggaard, H. and Walther, K. H. (1993). *Dynamic modelling with mixed graphical association models*. Master's thesis, Aalborg University.
- Marinazzo, D., Liao, W., Chen, H. and Stramaglia, S. (2011). Nonlinear connectivity by Granger causality. *NeuroImage* **58**, 330–338.
- Moneta, A. (2007). Graphical causal models and VARs: an empirical assessment of the real business cycles hypothesis. *Empirical Economics* DOI 10.1007/s00181-007-0159-9.
- Moneta, A. and Spirtes, P. (2006). Graphical models for the identification of causal structures in multivariate time series models. *Joint Conference on Information Sciences Proceedings*, Atlantis Press.
- Pearl, J. (1993). Graphical models, causality and interventions. *Statistical Science* **8**, 266–269.
- Pearl, J. (1995). Causal diagrams for empirical research (with discussion). *Biometrika* **82**, 669–710.
- Pearl, J. (2000). *Causality*. Cambridge University Press, Cambridge, UK.
- Pearl, J. and Verma, T. (1991). A theory of inferred causation. In J. A. Allen, F. Fikes and E. Sandewall (eds), *Principles of Knowledge Representation and Reasoning: Proceedings of the 2nd International Conference*, Morgan Kaufmann, San Mateo, CA, pp. 441–452.
- Ramsahai, R. R. (2011). Supplementary variables for causal estimation. In C. Berzuini, P. Dawid and L. Bernardinelli (eds), *Causality: Statistical Perspectives and Applications*, Wiley, chapter 16.
- Reale, M. and Tunnicliffe Wilson, G. (2001). Identification of vector AR models with recursive structural errors using conditional independence graphs. *Statistical Methods and Applications* **10**, 49–65.
- Richardson, T. (2003). Markov properties for acyclic directed mixed graphs. *Scandinavian Journal of Statistics* **30**, 145–157.
- Richardson, T. and Spirtes, P. (2002). Ancestral graph Markov models. *Annals of Statistics* **30**, 962–1030.
- Schweder, T. (1970). Composable markov processes. *Journal of Applied Probability* **7**, 400–410.
- Shpitser, I. (2011). Structural equations, graphs and interventions. In C. Berzuini, P. Dawid and L. Bernardinelli (eds), *Causality: Statistical Perspectives and Applications*, Wiley, chapter 3.
- Sims, C. A. (1972). Money, income and causality. *American Economic Review* **62**, 540–552.
- Spirtes, P., Glymour, C. and Scheines, R. (2001). *Causation, Prediction, and Search*. 2nd edn, MIT Press, Cambridge, MA. With additional material by David Heckerman, Christopher Meek, Gregory F. Cooper and Thomas Richardson.
- Spirtes, P., Richardson, T. S., Meek, C., Scheines, R. and Glymour, C. (1998). Using path diagrams as a structural equation modelling tool. *Soc. Methods Res.* **27**, 182–225.
- Su, L. and White, H. (2007). A consistent characteristic function based test for conditional independence. *Journal of Econometrics* **141**, 807–834.
- Suppes, P. (1970). *A probabilistic theory of causality*. North-Holland, Amsterdam.
- Swanson, N. R. and Granger, C. W. J. (1997). Impulse response functions based on a causal approach to residual orthogonalization in vector autoregressions. *Journal of the American Statistical Association* **92**, 357–367.

- Tunncliffe Wilson, G. (1972). The factorization of matricial spectral densities. *SIAM J. Appl. Math.* **23**, 420–426.
- Valdes-Sosa, P. A., Roebroeck, A., Daunizeau, J. and Friston, K. (2011). Effective connectivity: influence, causality and biophysical modeling. *NeuroImage* **58**, 339–361.
- Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica* **57**, 307–333.
- White, H. and Lu, X. (2010). Granger causality and dynamic structural systems. *Journal of Financial Econometrics* **8**, 193–243.
- Zhang, J. and Spirtes, P. (2005). A characterization of Markov equivalence classes for ancestral graphical models. *Technical Report 168*, Department of Philosophy, Carnegie Mellon University.