Deceivingly simple causal effect estimation from time series data

Søren Wengel Mogensen

Center for Statistics
Department of Finance
Copenhagen Business School

November 13, 2024

Deceivingly simple causal effect estimation from time series data

... or *A tale of partially observed time series and causality.*

Søren Wengel Mogensen

Center for Statistics
Department of Finance
Copenhagen Business School

November 13, 2024

Some influential causal models

ADAM and MAKRO are examples of *macroeconomic models* of the Danish economy. They combine historic data and economic theory to produce predictions about economic indicators.

Some influential causal models

ADAM and MAKRO are examples of *macroeconomic models* of the Danish economy. They combine historic data and economic theory to produce predictions about economic indicators.

They are *causal* in the sense that they can produce predictions under unseen interventions, e.g., changes to taxation. This makes them tools for policy evaluation.

The beginning

We assume that $X_t = (X_t^1, X_t^2, \dots, X_t^n)^\top$, $t \in \mathbb{Z}$, is a stochastic process in discrete time. For $D \subseteq V = \{1, 2, \dots, n\}$, we let X_t^D denote $\{X_t^i : i \in D\}$.

$$X_t^j = f_t^j(\dots, X_{t-2}^{pa_{t,2}^j}, X_{t-1}^{pa_{t,1}^j}, \varepsilon_t^j), \tag{1}$$

where ε_t are a sequence of iid vectors with independent entries, and $pa_{t,s}^j \subseteq V$. The collection of *structural assignments* in (1) along with a distribution of ε_t^j constitute a dynamic *structural causal model (SCM)* [Peters et al., 2017].

If we choose a particularly simple SCM, X_t is a *vector-autoregressive process (VAR)*,

$$X_t = \sum_{k=1}^{p} A_k X_{t-i} + \varepsilon_t$$

where A_k are $n \times n$ matrices.

The beginning

We should define what we mean by ‘causal’.[1] The equations are actually assignments in the following sense. We define an *intervention* as an action that exogenously fixes one of the variables at a certain value (this is known as a *do-intervention* [Pearl, 2009]),

$$X_t^j = c. \tag{2}$$

The *interventional distribution* (for the intervention $do(X_t^j = c)$) is the distribution entailed by the SCM where the original X_t^j-equation has been replaced by (2), and all other equations remain unchanged (this is a *hard* intervention, one can define more general interventions).

[1] There are different ways to do this. We follow the approach in Pearl [2009].

The beginning

We should define what we mean by 'causal'.[1] The equations are actually assignments in the following sense. We define an *intervention* as an action that exogenously fixes one of the variables at a certain value (this is known as a *do-intervention* [Pearl, 2009]),

$$X_t^j = c. \tag{2}$$

The *interventional distribution* (for the intervention $do(X_t^j = c)$) is the distribution entailed by the SCM where the original X_t^j-equation has been replaced by (2), and all other equations remain unchanged (this is a *hard* intervention, one can define more general interventions).

This leads to an entire collection of interventional distributions and one observational (no intervention). We will only assume access to (observations from) the observational distribution.

We call this *interventional causality* for ease of reference.

[1] There are different ways to do this. We follow the approach in Pearl [2009].

Causal graphs

A classical statistical model is a collection of distributions. If the model is parametrized, a choice of parameters leads to a single distribution. A *causal* model entails more than a classical model: It specifies not only an *observational distribution*, but also *interventional distributions*.

From a structural causal model, we can define a *full-time causal graph* with nodes $\{X_t^j : t \in \mathbb{Z}, j \in V\}$ such that $X_{t-s}^i \to X_t^j$ if $X_{t-s}^i \in pa_{t,s}^j$. That is, an edge from X_s^i to X_t^j means that X_s^i is an argument in f_t^j.

Causal graphs

As an example, we let $n = 3$, $V = \{1, 2, 3\}$, and for all t

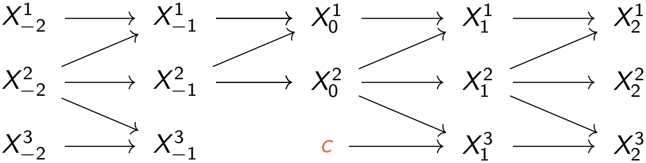
$$\begin{aligned} X_t^1 &= f_t^1(X_{t-1}^1, X_{t-1}^2, \varepsilon_t^1), \\ X_t^2 &= f_t^2(X_{t-1}^2, \varepsilon_t^2), \\ X_t^3 &= f_t^3(X_{t-1}^2, X_{t-1}^3, \varepsilon_t^3). \end{aligned}$$

$$\begin{tikzcd}
X^1_{-2} \arrow[r] & X^1_{-1} \arrow[r] & X^1_0 \arrow[r] & X^1_1 \arrow[r] & X^1_2 \\
X^2_{-2} \arrow[r] \arrow[ur] \arrow[dr] & X^2_{-1} \arrow[r] \arrow[ur] \arrow[dr] & X^2_0 \arrow[r] \arrow[ur] \arrow[dr] & X^2_1 \arrow[r] \arrow[ur] \arrow[dr] & X^2_2 \\
X^3_{-2} \arrow[r] & X^3_{-1} \arrow[r] & X^3_0 \arrow[r] & X^3_1 \arrow[r] & X^3_2
\end{tikzcd}$$

Causal graphs

We now consider the intervention $do(X_0^3 = c)$,

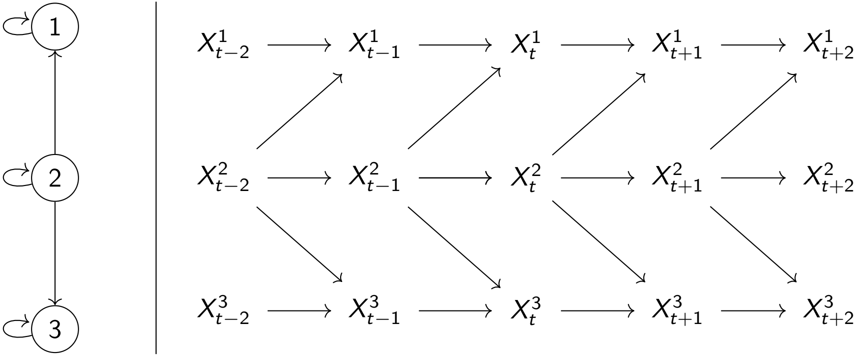
$$
\begin{aligned}
X_t^1 &= f_t^1(X_{t-1}^1, X_{t-1}^2, \varepsilon_t^1), \\
X_t^2 &= f_t^2(X_{t-1}^2, \varepsilon_t^2), \\
X_t^3 &= \begin{cases} f_t^3(X_{t-1}^2, X_{t-1}^3, \varepsilon_t^3), & t \neq 0 \\ {\color{red}c}, & t = 0 \end{cases}.
\end{aligned}
$$



One can also define more general interventions, including *soft* interventions.

Causal graphs

From the full-time causal graph, we define a *(summary) causal graph* on nodes V such that $i \to j$ if there exists an edge $X_s^i \to X_t^j$ for some s, t in the full-time graph.



The story

Today, I will talk about two simple concepts in time series analysis that are ‘causal’, either in name or in interpretation. When these concepts are introduced, they are often given a cautiously causal interpretation.

The story

Today, I will talk about two simple concepts in time series analysis that are ‘causal’, either in name or in interpretation. When these concepts are introduced, they are often given a cautiously causal interpretation.

We will see that, under the right assumptions, both concepts are closely related to our notion of causality from the previous slides.

Granger causality

If $D \subseteq V$, let $X_t^D = \{X_t^i : i \in D\}$ and $X_{<t}^D = \{X_s^i : i \in D, s < t\}$. If $D = \{d\}$, then $X_t^D = X_t^d$ and $X_{<t}^D = \{\dots, X_{t-3}^d, X_{t-2}^d, X_{t-1}^d\}$.

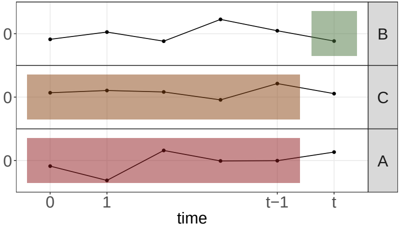
Definition (Granger (non)causality, Granger [1969], Eichler and Didelez [2010])

We say that A is Granger noncausal for B given C, and write $A \not\to B \mid C$, if for all $t \in \mathbb{Z}$,

$$X_{<t}^A \perp\!\!\!\perp X_t^B \mid X_{<t}^C.$$

Granger causality is analogous to local independence [Schweder, 1970, Aalen, 1987] in continuous-time processes.

Note that Granger (non)causality is not symmetric, i.e., $A \not\to B \mid C$ does not imply $B \not\to A \mid C$.



Granger causality

The original paper has 36000+ citations on Google Scholar. Papers in econometrics, neuroscience, environmental sciences, etc., have used Granger causality to study influence between coordinate processes in a multivariate stochastic process.

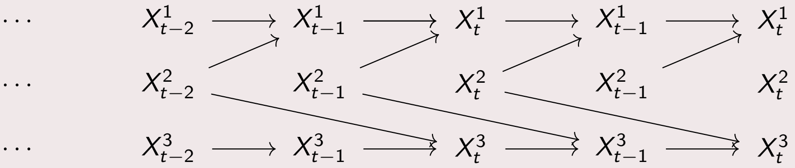
There is a large number of generalizations and other related theoretical work. Shojaie and Fox [2022] give a recent overview.

Granger causality – why it is not causal

It is quite clear that Granger causality does not correspond to interventional causality.[2] Granger (in)dependence would be a better name.

Example (Peters et al. [2017])

Let $V = \{1, 2, 3\}$. We see that an intervention on process 1 will not change the distribution of process 3. However, process 1 may be Granger causal for process 3 given process 3, i.e., $X^1_{<t} \perp\!\!\!\perp X^3_t \mid X^3_{<t}$ need not hold.



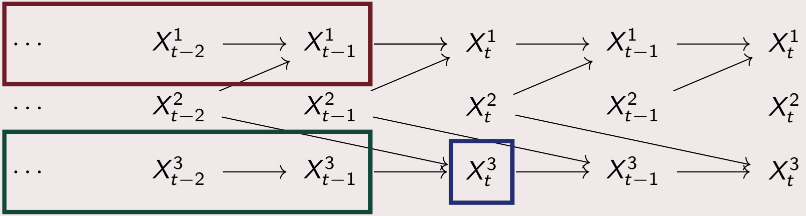
[2] Clive Granger acknowledged a similar limitation in the original paper [Granger, 1969].

Granger causality – why it is not causal

It is quite clear that Granger causality does not correspond to interventional causality.[2] Granger (in)dependence would be a better name.

Example (Peters et al. [2017])

Let $V = \{1, 2, 3\}$. We see that an intervention on process 1 will not change the distribution of process 3. However, process 1 may be Granger causal for process 3 given process 3, i.e., $X^1_{<t} \perp\!\!\!\perp X^3_t \mid X^3_{<t}$ need not hold.



[2] Clive Granger acknowledged a similar limitation in the original paper [Granger, 1969].

Granger causality – learning the causal graph

Applied papers that use Granger causality often state that Granger causality is not 'real causality' – but omit the fact that under the right set of assumptions there is a strong link between Granger causality and interventional causality.

Theorem

Under regularity conditions, i is Granger causal for j given $V \setminus \{i\}$ if and only if the edge $i \rightarrow j$ is in the (summary) causal graph.

Granger causality – learning the causal graph

Applied papers that use Granger causality often state that Granger causality is not 'real causality' – but omit the fact that under the right set of assumptions there is a strong link between Granger causality and interventional causality.

Theorem

Under regularity conditions, i is Granger causal for j given $V \setminus \{i\}$ if and only if the edge $i \to j$ is in the (summary) causal graph.

This means that we can actually learn the causal graph from the observational distribution of X_t using tests of Granger causality!

Granger causality – the partially observed case

As before, we assume that there exists a process, X_t, with an interventionally causal interpretation. Now we observe only a subset of the coordinate processes, $O \subseteq V$. If $O \neq V$, we are not able to test if i is Granger causal for j given $V \setminus \{i\}$.

We can still test if i is Granger causal for j given C when $i, j \in O$ and $C \subseteq O$! Each (summary) causal graph (a *directed graph*) can be *marginalized* to find a *directed mixed graph (DMG)* (with bidirected edges, $\leftrightarrow$, and directed edges $\rightarrow$) that represents the same set of Granger causalities (μ-separations) as the causal graph when restricting to O.

The global Markov property

μ-separation is a graphical algorithm which is analogous to *d*-separation in *directed acyclic graphs (DAGs)* [Pearl, 2009] and a generalization of δ-separation Didelez [2008].

Theorem

We let $\mathcal{D} = (V, E)$ be the causal graph of X_t, and we let $i, j \in V$, $C \subseteq V$. Under regularity conditions,

$$i \perp_\mu j \mid C \; [\mathcal{D}] \Rightarrow i \textit{ is Granger noncausal for } j \textit{ given } C.$$

Markov equivalence

Let $\mathcal{G} = (V, E)$ and $\bar{\mathcal{G}} = (V, \bar{E})$ be DMGs. We say that $\mathcal{G}$ and $\bar{\mathcal{G}}$ are *Markov equivalent* if for all $A, B, C \subseteq V$,

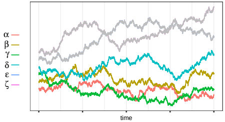
$$A \perp_\mu B \mid C \; [\mathcal{G}] \Leftrightarrow A \perp_\mu B \mid C \; [\bar{\mathcal{G}}].$$

We use $[\mathcal{G}]$ to denote the Markov equivalence class of $\mathcal{G}$. It is possible to characterize Markov equivalence classes of DMGs, to find a graphical representation of each Markov equivalence class, and to learn an equivalence class from data [Mogensen and Hansen, 2020, Mogensen, 2024a].

This means that even when some coordinate processes are unobserved, tests of Granger causality can output the collection of (marginalized) graphs that are equivalent with the causal graph in the sense that they represent the same set of Granger causalities.

This is analogous to causal discovery for DAG-based models using *d*-separation [Spirtes and Zhang, 2018].

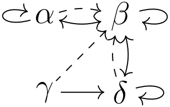
Structure learning

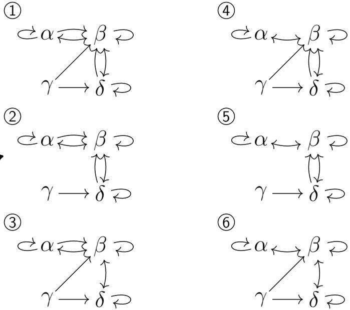


Granger causality test

$$\begin{aligned}\{&\alpha \not\to \gamma \mid \emptyset,\\ &\alpha \not\to \gamma \mid \gamma,\\ &\alpha \not\to \gamma \mid \delta,\\ &\quad\dots\ \}\end{aligned}$$

equivalence theory, learning algorithm





Granger causality

In summary, Granger causality is a ternary independence relation (just like conditional independence of random variables). This independence relation identifies the causal graph when we have full observation. Under partial observation, it can still identify an equivalence class of ‘marginalized graphs’ which contains the marginalized version of the true causal graph.

Causal effects

We saw that Granger causality can tell us something about causal structure. We now look at *causal effects* instead. We assume throughout that X_t is a stationary VAR-process.

Definition (Total causal effect)

Let t, s be integers, and let $k, l \in V = \{1, 2, \ldots, n\}$. The *total causal effect of* X_t^k *on* X_{t+s}^l is defined as

$$\tau_{kl}^s = \tau_{kl}^{ts} = \frac{\partial}{\partial x} E(X_{t+s}^l \mid do(X_t^k = x)).$$

We see that for $s < 0$, $\tau_{kl}^s = 0$.

The *total causal effect of* X_k *on* X_l is the sequence of lag-specific total causal effects, $(\tau_{kl}^s)_s$.

Total causal effects

Using the infinite graph with nodes $\{X_t^k : k \in V, t \in \mathbb{Z}\}$ such that $X_t^k \to X_{t+s}^l$ if $(A_s)_{lk} \neq 0$, τ_{kl}^s can be computed as the sum of products of edge coefficients along directed paths from X_t^k to X_{t+s}^l:

Proposition

It holds that

$$\tau_{kl}^s = \left(\sum_{i_1+\ldots+i_k=s} A_{i_1} \cdot \ldots \cdot A_{i_k} \right)_{lk}$$

where the summation is over all ordered partitions of s.

Total causal effects in VAR-processes are closely related to *impulse response functions*. One can find slightly different versions of these — Lütkepohl [2005] defines *forecast error impulse responses* and these equal total causal effects in VAR-processes.

Impulse responses

Impulse response analysis is commonly used in time series analysis and (implicitly) given a causal interpretation. Lütkepohl [2005] defines the (forecast error) impulse response to process i as the response of the system to a unit shock at time t in process i (i.e., $\varepsilon_t^i = 1$) when all other noise variables are zero.

The impulse responses/total causal effects are also the coefficients, Φ_k, of a *moving-average representation* of X_t,

$$X_t = \sum_{k=0}^{\infty} \Phi_k \varepsilon_{t-k},$$

and this is sometimes used as a definition for the impulse responses.

Impulse responses

Often a word of caution is added when introducing impulse responses. Lütkepohl [2005] writes

“All effects of omitted variables are assumed to be in the innovations. If important variables are omitted from the system, this may lead to major distortions in the impulse responses and makes them worthless for structural interpretations.

[3] All of this can also be defined for the slightly more general VARMA-processes that are closed under marginalization, i.e., impulse responses will also be well-defined for the marginal process.

Impulse responses

Often a word of caution is added when introducing impulse responses. Lütkepohl [2005] writes

"All effects of omitted variables are assumed to be in the innovations. If important variables are omitted from the system, this may lead to major distortions in the impulse responses and makes them worthless for structural interpretations.

In essence, if some processes are unobserved the above cautions us that the impulse responses to a shock in i on j in a marginalized process, X_t^O, $i, j \in O$, are not necessarily the same is in the original process, X_t^V (e.g., when computing them as MA-coefficients).[3]

[3] All of this can also be defined for the slightly more general VARMA-processes that are closed under marginalization, i.e., impulse responses will also be well-defined for the marginal process.

Impulse responses

Often a word of caution is added when introducing impulse responses. Lütkepohl [2005] writes

"All effects of omitted variables are assumed to be in the innovations. If important variables are omitted from the system, this may lead to major distortions in the impulse responses and makes them worthless for structural interpretations.

In essence, if some processes are unobserved the above cautions us that the impulse responses to a shock in i on j in a marginalized process, X_t^O, $i, j \in O$, are not necessarily the same is in the original process, X_t^V (e.g., when computing them as MA-coefficients).[3]

If we have a partially observed time series (we observe coordinate processes O instead of V), it is in general not possible to identify the total causal effects. However, one can find assumptions on the causal graph that imply identifiability of total causal effects.

[3] All of this can also be defined for the slightly more general VARMA-processes that are closed under marginalization, i.e., impulse responses will also be well-defined for the marginal process.

Identification of causal effects under partial observation

We let $k, l \in C$, $C \subseteq V$, and we let $Z_t = X_t^C$ be the subprocess corresponding to the coordinates in C. Under some regularity conditions, we get the following when $\tilde{\tau}_{kl}^s$ is a particular estimator based on the Yule-Walker equations.

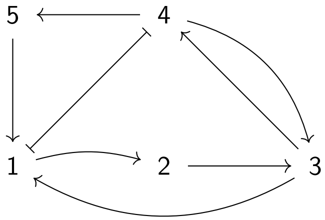
Theorem (Mogensen [2024b])

Let $k, l \in C \subseteq V$. If all μ-connecting walks in the (summary) causal graph from k to l given $C \setminus \{k\}$ have a tail at the endpoint k, $k \rightarrow \ldots l$, then $\tilde{\tau}_{kl}^s$ is a consistent estimator of τ_{kl}^s.

This is analogous to causal adjustment in DAG-based models, see, e.g., Henckel et al. [2022].

(asymptotic normality also holds such that one can analyze the efficiency of different sets C that meet the requirement of the theorem similar to Henckel et al. [2022] in the DAG setting.)

Example



The set $C = \{1, 3\}$ does not meet the requirement in the theorem when estimating the total causal effect of 1 on 3. On the other hand, $C = \{1, 3, 4\}$ does meet the requirement.

This also allows a slight generalization where the noise variables, ε_t^j, may be correlated within lag ($E(\varepsilon_t \varepsilon_t^\top)$ is not diagonal), represented by *blunt edges*, $\vdash\!\dashv$.

The end

Time series literature has much work on causality, often including disclaimers that one should be cautious when there is a risk of unobserved confounding processes.

The end

Time series literature has much work on causality, often including disclaimers that one should be cautious when there is a risk of unobserved confounding processes.

For Granger causality, we saw that there is a very clear link to the causal graph under full observation. However, using tests of Granger causality, we can still learn about the causal graph even in the case of partial observation.

The end

Time series literature has much work on causality, often including disclaimers that one should be cautious when there is a risk of unobserved confounding processes.

For Granger causality, we saw that there is a very clear link to the causal graph under full observation. However, using tests of Granger causality, we can still learn about the causal graph even in the case of partial observation.

Impulse responses can be thought of as total causal effects. Some texts caution that impulse responses are not not necessarily meaningful under partial observation. More precisely, we are interested in impulse responses of the underlying causal process, and those are not identified from the observational distribution without further assumptions.

The end

Time series literature has much work on causality, often including disclaimers that one should be cautious when there is a risk of unobserved confounding processes.

For Granger causality, we saw that there is a very clear link to the causal graph under full observation. However, using tests of Granger causality, we can still learn about the causal graph even in the case of partial observation.

Impulse responses can be thought of as total causal effects. Some texts caution that impulse responses are not not necessarily meaningful under partial observation. More precisely, we are interested in impulse responses of the underlying causal process, and those are not identified from the observational distribution without further assumptions.

Both are useful in the partially observed case, however, only if we delineate what we are assuming about the process.

Thank you for listening!

References I

Odd O. Aalen. Dynamic modelling and causality. *Scandinavian Actuarial Journal*, 1987(3-4):177–190, 1987.

Vanessa Didelez. Graphical models for marked point processes based on local independence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):245–264, 2008.

Michael Eichler and Vanessa Didelez. On Granger causality and the effect of interventions in time series. *Lifetime Data Analysis*, 16(1):3–32, 2010.

Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438, 1969.

Leonard Henckel, Emilija Perković, and Marloes H Maathuis. Graphical criteria for efficient total effect estimation via adjustment in causal linear models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(2): 579–599, 2022.

Helmut Lütkepohl. *New Introduction to Multiple Time Series Analysis*. Springer, Berlin, 2005.

References II

Søren Wengel Mogensen. Weak equivalence of local independence graphs. *Bernoulli*, 2024a. (to appear).

Søren Wengel Mogensen. Causal adjustment in time series. 2024b.

Søren Wengel Mogensen and Niels Richard Hansen. Markov equivalence of marginalized local independence graphs. *The Annals of Statistics*, 48(1): 539–559, 2020.

Judea Pearl. *Causality*. Cambridge University Press, 2009.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. Cambridge, MA: MIT Press, 2017.

Tore Schweder. Composable Markov processes. *Journal of Applied Probability*, 7 (2):400–410, 1970.

Ali Shojaie and Emily B Fox. Granger causality: A review and recent advances. *Annual Review of Statistics and Its Application*, 9(1):289–319, 2022.

Peter Spirtes and Kun Zhang. Search for causal models. In *Handbook of Graphical Models*, pages 439–470. CRC Press, 2018.