Inferring causality in time series data

A concise review of the major approaches.

Shay Palachy
Nov 12, 2019 · 43 min read ★

The question of what event caused another, or what brought about a certain change in a phenomenon, is a common one. Examples include whether a drug caused an improvement in some medical condition (versus the placebo effect, additional hospital visits, etc.), tracking down the cause for a malfunction in an assembly line or determining what caused an upsurge in a website's traffic.

While a naive interpretation of the problem may suggest simple approaches like equating causality with high correlation, or to infer the degree to which *x* causes *y* from the degree of *x*'s goodness as a predicator of *y*, the problem turns out to be much more complex. As a result, rigorous ways to approach this question were developed in several fields of science.

The task of causal inference divides into two major classes:

1. Causal inference over random variables, representing different events. The most common example are two variables, each representing one alternative of an A/B test, and each with a set of samples/observations associated with it.

2. Causal inference over time series data (and thus over stochastic processes). Examples include determining whether (and to what degree) aggregate daily stock prices drive (and are driven by) daily trading volume, or causal relations between volumes of Pacific sardine catches, northern anchovy catches, and sea surface temperature.

This post deals only with the second class of problems.

Scope

This post is meant to provide a concise technical review of the major approaches found in academic literature and online resources for the purpose of inferring causality in time series data, the methods derived from them and their implementation in code form.

It aims to touch upon both (1) classical *statistical* approaches, created mainly in the *econometrics* field of research, including modern developments (2) and adaptions and original approaches coming from various other research communities, such as those dealing with dynamic systems or information theory.

The general subject of causal inference is both too large and not directly applicable enough to cover in this post. The same is true for the intersection between causal inference in general (which in many cases is done on general probability distributions, or their samples, and not on time series data) and machine learning. Nevertheless, I have included a few notable resources I encountered on these topics in the *Other Notable Literature* section.

Organization & Notation

General remarks are sometimes given in quote format in-body, while optional notes are given in clickable footnotes (each with a link to send you back to its origin). A clickable table-of-contents is also provided to assist in navigation. I've added the 📖 link to the header of each section; click on it to quickly return to the table of contents. Finally, complete references for the literature used accompanies the post.

Required background

The post is written in technical language, and while I obviously cannot go as deep as the academic papers referenced throughout this post, I will not shy away from including equations, notations and text requiring technical background for their understanding.[9]

As such, a background of — or equivalent to — at least a thorough undergraduate theoretical course in probability and statistics, including the required mathematical background is assumed. If you are not familiar with stochastic processes theory, you can find a concise review of it in my post on stationarity in time series data. Familiarity with this theory is required for further reading, as it is the framework upon which statistical notions of causality in time series are built.

Table of Contents

. . .

Background: Notions of causality in time series data

Throughout the years a number of different notions of causality were suggested by scholars of statistics and economics. I give here an overview of the major ones. This part is based, in part, on [Eichler, 2011] and [Runge, 2014]. 📖

Granger causality[8]

The earliest concept of causality for time series data was suggested by Granger [1969, 1980], building on a notion from [Wiener, 1956]. It is based on contrasting the ability to predict a stochastic process *Y* using *all the information in the universe,* denoted with *U*, with doing the same using all information in *U* except for some stochastic process *X*; this is denoted with *U\X*. The core idea is that if discarding X reduces the predictive power regarding Y, then X contains some unique information regarding Y, and we thus say that *X Granger-causes Y*.

More formally:

- Let *X* and *Y* be stationary stochastic processes.
- Denote with $\mathscr{U}_i=(U_{i-1},\dots,U_{i-\infty})$ all the information in the universe until time *i,* and with $\mathscr{X}_i=(X_{i-1},\dots,X_{i-\infty})$ all information in *X* until time *i*.
- Denote with $\sigma^2(Y_i|\mathscr{U}_i)$ the variance of the residual of predicting Y_i using $\mathscr{U}_i$ at time *i*.

- Denote with $\sigma^2(Y_i|\mathscr{U}_i \backslash \mathscr{X}_i)$ the variance of the residual of predicting Y_i using all information in $\mathscr{U}_i$ at time *i* except $\mathscr{X}_i$.

Definition 1: If $\sigma^2(Y_i|\mathscr{U}_i) < \sigma^2(Y_i|\mathscr{U}_i \backslash \mathscr{X}_i)$ then we say that *X Granger-causes Y,* and write $X \Rightarrow Y$.

Definition 2: If $X \Rightarrow Y$ and $Y \Rightarrow X$ we say that *feedback* is occurring, and write $X \Leftrightarrow Y$.

As noted by Granger himself, the requirement of having access to all the information in the universe is extremely unrealistic. In practice *U* is replaced by a limited set of observed time series $\mathbf{X}$, with $X \in \mathbf{X}$, and the above definition reads *X Granger-causes Y with respect to* $\mathbf{X}$.

Finally, this definition does not specify the prediction method used for σ^2, and thus allows for both linear and non-linear models, but the use of the variance to quantify the closeness of prediction restricts this notion of causality to causality in mean.

This notion is usually referred to as *strong Granger causality*; other related notions of causality are *Granger causality in mean* [Granger 1980, 1988] and *linear Granger causality* [Hosoya 1977, Florens and Mouchart 1985].

Instantaneous causality: A related kind of causality, modifying Granger causality slightly, is *instantaneous causality* [Price, 1979]. We say that X and Y has *instantaneous causality* between them if, at time *i*, adding X_i to the information set helps to improve the predicted value of Y_i.

More formally:

- Let *X* and *Y* be stationary stochastic processes.
- Denote with $\mathscr{U}_i = (U_{i-1}, \dots, U_{i-\infty})$ all the information in the universe until time *i,* and with $\mathscr{X}_i = (X_{i-1}, \dots, X_{i-\infty})$ all information in *X* until time *i* (in both cases, not including information from time *i* itself).

Definition 3: If $\sigma^2(Y_i|\mathscr{U}_i \cup \{X_i\}) < \sigma^2(Y_i|\mathscr{U}_i)$ then we say that there is *instantaneous causality* between *X* and *Y*.

Note that this type of causality is not directional, but rather symmetric, and it can be shown that if the above definition holds then the symmetric statement — that $\sigma^2(X_i|\mathscr{U}_i \cup \{Y_i\}) < \sigma^2(X_i|\mathscr{U}_i)$ — also holds(see [Lütkepohl, 2007] for proof). Thus, we do not say that *X* instantaneously causes *Y*, but rather that there is instantaneous causality between *X* and *Y*.

Multi-step causality: In a bivariate system, if the 1-step ahead forecasts of one variable cannot be improved by using the information in the other variable, the same holds for all *h*-step forecasts for any *h=1,2,…*, and so the 1-step ahead criteria is sufficient to define Granger causality. This result does not hold anymore if the information set contains additional variables. [Lütkepohl and Müller, 1994]

Thus, in a multivariate system, we say that variable X_i is *h*-step causal for another variable Y_i if the information in X_i helps improve the *j*-step forecasts of Y_i for some *j=1, 2, …, h*.

Sims causality

In an influential paper, [Sims, 1972] showed — in the context of covariance stationary processes, and restricted to linear predictors — that in the bivariate case the definition of Granger causality is equivalent to parameter restrictions of the moving average or distributed lag representations of the processes x[t], y[t]. When the system is covariance stationary it can be represented as:

$$x_t = \sum_{j=0}^{\infty} a_j u_{t-j} + \sum_{j=0}^{\infty} b_j v_{t-j}$$

$$y_t = \sum_{j=0}^{\infty} c_j u_{t-j} + \sum_{j=0}^{\infty} d_j v_{t-j}$$

Equation 8: The Sims representation for covariant stationary processes

where a_j, b_j, c_j and d_j are constants and *u[t]* and *v[t]* are mutually uncorrelated white noise processes. Sims shows that the condition *x[t] does not Granger cause y[t+1]* is equivalent to c_j or $_j$ being chosen identically zero for all *j*.

In contrast to Granger's definition, which considers temporal precedence in the form of a link from the past to the present, Sims' notion considers temporal precedence in the form of a link from the present to the future. As a consequence, the potential causal relationship considered runs from the dependent variable to future values, or ‘leads’, of the regressor.

While at the time of its introduction in [Sims, 1972] it was presented as an equivalent definition to Granger’s, it was since contrasted with it and was shown to be inequivalent when the measure of uncorrelatedness of time series used is independence [Florens and Mouchart, 1982]; rather, it is shown that Granger causality is a stronger condition, and that while Granger causality implies Sims causality, the inverse is not true.

In spite of this inequivalency, most statistical tests for causal inference in time series data focus on Granger’s definition. However, at least in the case where a vector autoregressive (VAR) model is used, these tests can be modified to test for Sims’ causality (see here for an example highlighting the difference between the tests for the linear case).

Structural Causality

Introduced by White and Lu (2010), structural causality assumes that the data-generating process (DGP) has a recursive dynamic structure in which predecessors structurally determine successors. Specifically, for two processes X — the potential cause — and *Y* — the response, we assume they are generated by

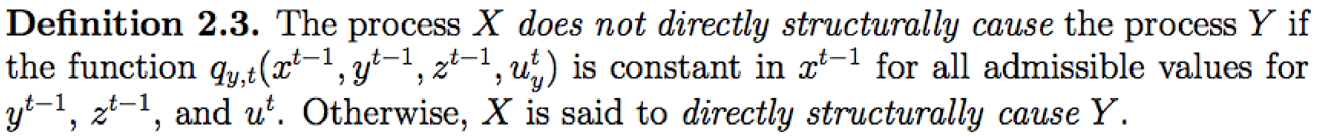
$$
\begin{aligned}
X_t &= q_{x,t}(X^{t-1}, Y^{t-1}, Z^{t-1}, U_x^t) \\
Y_t &= q_{y,t}(X^{t-1}, Y^{t-1}, Z^{t-1}, U_y^t)
\end{aligned}
$$

Equation 9: The structural causality DGP

for all $t \in Z$. Here, the process *Z* includes all relevant observed variables while the realizations of $U=(U_x, U_y)$ are assumed to be unobserved, and the functions *q[x,t]* and *q[y,t]* are assumed to be unknown.

Observe that this dynamic structure is general, in that the structural relations may be nonlinear and non-monotonic in their arguments and non-separable between observables and unobservables. The unobservables may be countably infinite in number. Finally, this system may generate stationary processes, non-stationary processes, or both.

A structural notion of causality can then be defined:



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^t. Otherwise, X is said to *directly structurally cause* Y.

The authors then go on to analyze the relations between Granger causality and this notion of structural causality. Additionally, building on classical notions of Granger causality, they introduce two extensions of it: *weak Granger causality* and *retrospective weak Granger causality*.

Finally, the authors construct practical tests for their two notions of Granger causality. Specifically, *weak Granger causality* is shown to be detectable by testing for the conditional independence of the response *Y* and the potential cause *X* given the history of the response and the near-histories of the observable covariates of the processes.

Intervention causality

The use of intervention as the basis for a statistical theory of causality inference, as championed by Judea Pearl, can be traced back at least to the early 90's [Pearl and Verma, 1991] [Pearl, 1993]. It's application to time series data, however, has begun to receive rigorous treatment only recently [White, 2006] [Eichler and Didelez, 2007]. This approach to causality is closely related to the use of impulse response analysis in economics.

Eichler and Didelez define a set of possible intervention regimes corresponding to different possible types of interventions in a multivariate stationary time series *X* with *d* components. Interventions are denoted by the intervention indicator σ which takes values in $\{\emptyset, s\in\mathcal{S}\}$; I use X_a to denote a component in *X*, and X^U to denote a subset of components in *X*, where $a\in V$ and $U\subset V$ for $V=\{1,\ldots,d\}$. I also use σ_a to denote an intervention in component X_a.

Intervention types are:

1. *Idle regime*: When $\sigma_a(t) = \emptyset$, $X_a(t)$ arises naturally without intervention. A.k.a. *the observational regime*.

2. *Atomic interventions*: Here $\mathcal{S}=X$, the domain of x_a, and $\sigma_a(t) = x^\bullet$ denotes an intervention forcing $X_a(t)$ to assume the value x•.

3. *Conditional intervention*: Here $\mathcal{S}$ consists of functions $g(x^U(t\text{-}1))\in X$, $U\subset V$, such that $\sigma_a(t)=g$ means $X_a(t)$ is forced to take on a value that

depends on past observations of $X^{u}(t=1)$.

4. *Random intervention:* Here $\mathcal{S}$ consists of distributions meaning that $X_{a}(t)$ is forced to arise from such a distribution.

Then, under some assumptions ensuring an intervention is an isolated exogenous change of the system, the *average causal effect* (ACE) of interventions in X according to strategy s on the response variable $Y[t`]$ is defined to be (assuming w.l.o.g. that $\mathbb{E}[Y[t`]]=0$):

$$ACE_s = \mathbb{E}_{\sigma_t=s} Y_{t'} - \mathbb{E} Y_{t'} = \mathbb{E}_{\sigma_t=s} Y_{t'}$$

Equation 10: The average causal effect (ACE) of interventions according to strategy s

Thus, the $ACE[s]$ can be regarded as the average difference between no intervention and intervention strategy s. Additionally, different strategies can be compared by considering $ACE[s_1]-ACE[s_2]$, or other functionals of the post-intervention distribution $\mathbb{P}[s](Y[t`])$.

Now, a priori there is no reason why data that is not collected under the intervention regime of interest should allow estimation of the *ACE*. However, the authors then go on to show the possibility of expressing the *ACE* in terms of quantities that are known or estimable under the observational regime, using what they call the *back-door criterion*.

I find this a very elegant reconciliation of causality in time series data with the the highly influential concept of intervention-based causality in general.

Additionally, a recent paper by Samartsidis et al. provides a thorough review of other methods for assessing the causal effect of interventions from aggregate time-series observational data for the specific case of *binary interventions* [Samartsidis et al. 2018].

> ***Note:*** *The notion of intervention causality is fundamentally different from the other three notions presented here; while Granger causality, Sims causality and structural causality all assume an observational framework, intervention causality makes the much stronger assumption that intervention can be performed in the studies processes. As such, it is significantly less applicable in many real-life scenarios.*

. . .

Classical methods for causality inference in time series data 📖

This section covers the two most basic approaches to causality inference, based on classical statistical approaches.

Non-directional lagged interactions

The perhaps most basic approach to inferring causal relationships between two time series X and Y is to use non-directional measures of correspondence between a lagged (back-shifted) version of the potentially-causing time series X to the (non-lagged) potentially-caused time series Y.

If a high degree of correspondence is found between a k-lag of X and (non-lagged) Y, then a very weak notion of X-causing-Y can be inferred; the direction is thus inferred from the fact that a lag of X has high correspondence with Y. Various measures of correspondence can be used; among them are Pearson correlation (e.g. [Tsonis and Roebber, 2004]), mutual information (e.g. [Donges at al. 2009]) and phase synchronization (e.g. [Pikovsky et al. 2003]).

When the chosen correspondence measure is the Pearson correlation, this is equivalent to looking at the cross-correlation function of the two time series for different positive and negative lags, and taking the maximum value it attains over the chosen range as the strength of the causal link, with the sign of the lag indicating the causal direction. Naively, if the function achieves positive values over both positive and negative lags then bi-directional causality can be inferred. In any case, the autocorrelation of both series must be taken into account in order to arrive at a valid interpretation.

This approach is employed mainly in climate research [Yamasaki et al. 2008] [Malik et al. 2012] [Radebach et al. 2013]. It was shown to have significant problems that might produce misleading conclusions, as discussed in chapter 4.5 of SIFT's online manual and demonstrated in section 5.2.1 of [Runge, 2014].

Parametric VAR-based tests for Granger causality

A concise breakdown of the classical parametric tests for Granger causality is given in [Greene, 2002]. A substantial number of these tests were constructed over the years to test for Granger causality.[4] I thus give a brief overview of the ones I have encountered, focusing on tests for which I could find an implementation in common data-processing programming languages (i.e. Python and R).

In general, the first phase in these tests is to make sure that all examined series are stationary — stationarizing them, usually through trend removal and/or differencing, if they are not.

Then, in pair-wise tests, for each pair of time series and for each specific direction of causality *X⇒Y*, a (usually manually) number of negative (past) lags of the potentially-causing series *X* are generated (including the zero-lag, which is *X* itself). The maximal lag length to take is a model selection consideration, and thus should be chosen based on some information criteria (e.g. Akaike information criterion, Bayesian information criterion, etc.).

> ***Note:*** *If checking a large number of pairs, you will also have to consider how to deal with problems arising from multiple hypothesis testing.*

The model used in all below cases is a vector autoregressive model of the endogenous (potentially caused) time series *Y* as a stochastic process; two such models are stated.

The first model — called the restricted model — assumes that *Y* linearly depends only on past values of itself with linear coefficients γ_i and a time-dependent noise term *e[t]*:

$$Y_t = \gamma_0 + \sum_{i=1}^{p} \gamma_i Y_{t-i} + e_t$$

Equation 11: The restricted model in a VAR-based Granger causality test

Conversely, the second — called the unrestricted model — assumes that Y linearly depends past values of both X and Y, determined by coefficients α_i, β_i and a time-dependent noise term u*[t]*:

$$Y_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i Y_{t-i} + \sum_{i=1}^{p} \beta_i X_{t-i} + u_t$$

Equation 12: The unrestricted model in a VAR-based Granger causality test

The unformalized null hypothesis is that the second model does not add information, or provides a better model of Y, when comparing it to the first model. This need to be formalized into a testable null hypothesis; a common approach is to to state that the null hypothesis H_0 is that $\forall i, \beta_i = 0$.

Finally, one of the below test procedures is applied to all such pairs of a lag of *X* and the unlagged *Y*. To check for causality in both directions, lags of *Y* are added to the set of examined series.

> ***Note:*** *Granger-causality tests are very sensitive to the choice of lag length and to the methods employed in dealing with any non-stationarity of the time series.*

SSR-based F-test for Granger causality: Parameters are estimated for both the restricted and the unrestricted model (usually using ordinary least squares). An F-statistic[1] is then computed using the RSS of the two series, which is given by:

$$\frac{(RSS_R - RSS_{UR})/p}{RSS_{UR}/(T-2p-1)} \sim F_{p,T-2p-1}$$

Equation 13: RSS-based F statistic for Granger causality

where *T* is time series length and *p* is the number of lags.

A nice overview of the bivariate case of this test is given here and here. A bivariate version is implemented in the *statsmodels* Python package [Python], in the MSBVAR package [R], the lmtest package [R][3], the NlinTS package [R] and the vars package [R].

Peasron's Chi-Square test for Granger causality: First, model parameters are estimated using OLS. A Chi-square-statistic[2] is computed using the SSR of the two series and the Peasron's Chi-Square test procedure is performed. A bivariate version is implemented in the *statsmodels* package [Python].

The likelihood ratio Chi-Square test (aka G-test) for Granger causality: A Chi-square-statistic[2] is computed using the likelihood ratio of the two series and the standard test procedure is followed. A bivariate version is implemented in the *statsmodels* package [Python].

Heteroskedasticity-robust F-test for Granger causality: Introduced in [Hafner and Herwartz, 2009], this procedure uses bootstrapping for parameter estimation that is robust to heteroskedasticity (and yields a more efficient estimator than OLS in this case, for example), and a custom Wald test statistic. A bivariate version is implemented in the vars package [R] (see the second test implemented in the *causality* method).

The Toda and Yamamoto procedure: Introduced in [Toda and Yamamoto, 1995], this procedure is meant to deal with testing for Granger causality in cases where the examined series are either integrated or cointegrated of an arbitrary order (or both). In these cases the tests statistics in the aforementioned tests do not follow their usual asymptotic distribution under the null; the procedure was developed to address this problem. The authors give a detailed test procedure that uses a standard Wald test as a component, in such a way that a properly distributed (under the null hypothesis) test statistic is achieved. Dave Giles has an outstanding blog post on the procedure.

> *I did not found a code implementation of this procedure, but I have included it because of its importance. It can be implemented by using existing implementations of all the procedures composing it.*

Other tests for linear Granger causality: Linear Granger causality tests were developed in many directions, e.g., [Hurlin and Venet, 2001] proposed a procedure for causality tests with panel data, while [Ghysels et al. 2016] introduced a test for Granger causality with mixed frequency data.

The above linear methods are appropriate for testing Granger causality in the mean. However they are not able to detect Granger causality in higher moments, e.g., in the variance. To deal with this challenge, and additional deficiencies in the classical model for Granger causality, a plethora of methods were suggested; these include nonlinear parametric approaches, and various non-parametric approaches.

The following sections then aim to concisely cover the numerous alternatives approaches to inferring causality in time series data, inspired by various fields in nature sciences.

. . .

Alternative parametric Granger causality measures for time series data 📖

Most of the following causality measures were overviewed and compared in [Papana et al. 2013].

Conditional Granger Causality Index (CGCI)

Introduced in [Geweke, 1984], it was the first attempt to suggest measures for the *degree* of linear dependence and feedback between multiple time series. The author introduced the decomposition of the linear causal relationship between X and Y as the sum of linear causality from X to Y, linear causality from Y to X, and instantaneous linear feedback between the two series. Furthermore, the introduced measures can (under certain conditions) be additively decomposed by frequency.

Using the same VAR model as the original linear Granger causality measure, the CGCI is similarly defined to be the natural logarithm of the ratio between the residual variance of the restricted model and that of the unrestricted model. The difference is thus just in the inclusion of additional time series besides X_1 and X_2 in both the restricted and unrestricted models; thus, if X_2 is only mediating the influence of some other time series Z on X_1, we can again expect the residual errors of the unrestricted model to be similar to that of the restricted one, in which case the index will be close to zero.

$$\mathrm{CGCI}_{X_2 \to X_1 | Z} = \ln(s_{1R}^2 / s_{1U}^2)$$

Equation 14: The Conditional Granger Causality Index

Restricted variants of the VAR model CGCI uses were proposed to handle higher-dimensional data, a smaller amount of samples or non-linear causal relations. [Siggiridou and Kugiumtzis, 2016] includes an overview of several such variants (and introduces another).

MLP-based F-test for Granger causality

This approach is very similar to the aforementioned VAR-based approach, but a perceptron replaces the VAR as the explaining model. Two multi-layer perceptron (MLP) neural networks models are trained — one just for the endogenous time series and one for both — and an F-test is performed to test the null hypothesis that the exogenous time series does not improve predictability of the endogenous time series. Implemented in the NlinTS package [R].

RBF models for Granger causality

[Ancona et al. 2004] suggested a non-linear parametric model for Granger causality, replacing the VAR model with the more rich family of radial basis functions (RBF), which was shown to able to approximate any real function to a desired degree, Recently, [Ancona and Stramaglia, 2006] showed that not all nonlinear prediction schemes are suitable to evaluate causality between two time series, since they should be invariant if statistically independent variables are added to the set of input variables. Driven by this finding, [Marinazzo et al. 2006] aims to find the largest class of RBF models suitable to evaluate causality.

Partial Granger Causality Index (PGCI)

CGCI (and its extensions) still assume the inclusion of all relevant variables. PGCI was introduced in [Guo et al. 2008] as a causality index that can handle the existence of exogenous inputs and latent (i.e. unobservable) variables in the examined system.

To determine the direct causal link from a variable Y to a variable X, given another variable Z (this can be naturally extended to multiple variables) and both exogenous input to the system and unobserved latent variables, the authors suggest the following restricted VAR model, with a noise covariance matrix S:

$$
\begin{aligned}
X_t &= \sum_{i=1}^{\infty} a_{1i} X_{t-i} + \sum_{i=1}^{\infty} c_{1i} Z_{t-i} + \overrightarrow{\varepsilon_{1t}} + \overrightarrow{\varepsilon^{\mathrm{E}}_{1t}} + \overrightarrow{B_1(L)\varepsilon^{L}_{1t}} \\
Z_t &= \sum_{i=1}^{\infty} b_{1i} Z_{t-i} + \sum_{i=1}^{\infty} d_{1i} X_{t-i} + \overrightarrow{\varepsilon_{2t}} + \overrightarrow{\varepsilon^{\mathrm{E}}_{2t}} + \overrightarrow{B_2(L)\varepsilon^{L}_{2t}}
\end{aligned}
$$

Equation 15: The Restricted VAR model for PGCI

And the following unrestricted VAR model, with a noise covariance matrix Σ:

$$
\begin{aligned}
X_t &= \sum_{i=1}^{\infty} a_{2i} X_{t-i} + \sum_{i=1}^{\infty} b_{2i} Y_{t-i} + \sum_{i=1}^{\infty} c_{2i} Z_{t-i} + \overrightarrow{\varepsilon_{3t}} + \overrightarrow{\varepsilon^{\mathrm{E}}_{3t}} + \overrightarrow{B_3(L)\varepsilon^{L}_{3t}} \\
Y_t &= \sum_{i=1}^{\infty} d_{2i} X_{t-i} + \sum_{i=1}^{\infty} e_{2i} Y_{t-i} + \sum_{i=1}^{\infty} f_{2i} Z_{t-i} + \overrightarrow{\varepsilon_{4t}} + \overrightarrow{\varepsilon^{\mathrm{E}}_{4t}} + \overrightarrow{B_4(L)\varepsilon^{L}_{4t}} \\
Z_t &= \sum_{i=1}^{\infty} g_{2i} X_{t-i} + \sum_{i=1}^{\infty} h_{2i} Y_{t-i} + \sum_{i=1}^{\infty} k_{2i} Z_{t-i} + \overrightarrow{\varepsilon_{5t}} + \overrightarrow{\varepsilon^{\mathrm{E}}_{5t}} + \overrightarrow{B_5(L)\varepsilon^{L}_{5t}}
\end{aligned}
$$

Equation 16: The Unrestricted VAR model for PGCI

Like in previous VAR models, matrices A_1,B_1, A_2, E_2 and K_2 model the autoregressive effect in each series, other matrices model the effect of the different lag of each model on the others, and ε_i are white noise processes. The new components here are ε_i^E, which are independent random vectors representing exogenous inputs, and ε_i^L, which are independent random vectors representing latent variables.

The authors go on to develop two measures: (1) A measure of the accuracy of the autoregressive prediction of X based on its previous values conditioned on Z by eliminating the influence of ε_i^E and ε_i^L. (2) A measure for the accuracy of predicting present value of X based on the previous history of both X and Y conditioned on Z by eliminating the influence of ε_i^E and ε_i^L. They then define PGCI to be the natural logarithm of the ratio between the two. Given in terms of the noise covariance matrices S and Σ of the two models, the index can be written as:

$$F_1 = \ln\left(\frac{S_{11} - S_{12}S_{22}^{-1}S_{21}}{\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}}\right)$$

Equation 17: PGCI in terms of the noise covariance matrices of the VAR models

By comparison, the standard Granger causality index can be expressed as $GCI = ln(|S_{11}|\backslash|\Sigma_{11}|)$.

The authors also extend their measure to the nonlinear case by using nonlinear RBF parametric models, as refined in [Marinazzo et al. 2006]. The index remains as in Eq. 1.

Directed Coherence Measures

The bivariate function of *coherence* is commonly used in signal processing to estimate the power transfer between the input and output of a linear system. [Saito and Harashima, 1981] has expanded on the concept by defining *directed coherence* (DC), which decomposes coherence into two components of direct coherence measures: one representing the feedforward dynamic and the other representing feedback dynamic in the examined system. The original paper used a bivariate autoregressive model, which was later generalized in several variations for the multivariate case.

[Baccalá and Sameshima, 2001] extended the concept of *directed coherence* with the definition of *Partial Directed Coherence (PDC),* which is based on the *partial coherence function*, as a coherence-based measure of Granger causality for the multivariate case.

. . .

Alternative non-parametric causality measures for time series data 📖

Note that most methods presented in the following sections, dealing with chaos and dynamic system theory approaches and information theoretic approaches to causality inference, are also non-parametric.

The Bouezmarni-Taamouti test

The authors of [Bouezmarni and Taamouti, 2010] give a non-parametric test for conditional independence and Granger causality for the bivariate case. Unlike most tests, which focus on causality in mean, the authors base their test on *conditional distributions*.

Other non-parametric approaches to testing causality are those suggested in [Bell et al. 1996] and [Su and White, 2003].

. . .

Chaos and dynamic system theory approaches for causality inference in time series data 📖

This section covers methods for causality inference based on the two closely related fields of chaos theory and dynamic system analysis. Naturally, these approaches are also related to information theory to some degree, which is covered in the next section.

The Hiemstra-Jones test

[Baek and Brock, 1992] developed a nonlinear Granger causality test which was modified by [Hiemstra and Jones, 1994] later on to study the bivariate nonlinear causal relationship between stock returns and stock trading volume. This test has become common in testing for non-linear Granger causality in the years that followed, and has been extended to the multivariate case by [Bai et al. 2010].

Although not commonly presented as such, their non-parametric dependence estimator is based on so-called correlation integral, a probability distribution and entropy estimator, developed by physicists Grassberger and Procaccia in the field of nonlinear dynamics and deterministic chaos as a characterization tool of chaotic attractors. [Hlaváčková-Schindler et al. 2007] It is thus also closely related to the CTIR measure examined in the following section, dealing with information theoretic measures for causality.

[Diks and Panchenko, 2005] have showed that the relationship tested by the process is not implied by the null hypothesis of Granger non-causality, and that the actual rejection rate may tend to one as the sample size increases. As a result, the test was revisited and a new version of it, overcoming some of the aforementioned problems (namely the growth of rejection rate with sample size), was suggested for both the bivariate case (in [Bai et al. 2017]) and the multivariate case (in [Bai et al. 2018]).

The Diks-Panchenko test

Building on their investigation into the problems of the Hiemstra-Jones test in [Diks and Panchenko, 2005], the authors suggest in [Diks and Panchenko, 2006] a new bivariate nonparametric test for Granger causality. They show significantly better behavior of the size[5] and power of their test as the sample size increases, when comparing to the Hiemstra-Jones test, while also testing for a relationship equivalent to the desired null hypothesis. [Diks and Wolski, 2015] extend the test to multivariate settings.

Extended Granger Causality Index (EGCI)

Introduced in [Chen et al. 2004], this method extends the classical Granger causality index to the non-linear case by restricting the application to local linear models in reduced neighborhoods and then averaging the resulting statistical quantity over the entire dataset.

This approach makes use of a technique from the field of dynamical system theory; *delay coordinate embedding* is used to reconstruct a phase space R, and the autoregressive model is then fitted in the reconstructed space R instead of the original space of the samples. The model is fitted for all points in the neighborhood (determined by a distance parameter δ) of a reference point z_0. The residual variance in the EGCI measure are then estimated using averaging over the neighborhood sampling the entire attractor. Finally, the EGCI is computed as a function of the neighborhood size δ. For linear systems the index should stay roughly the same as δ

becomes smaller, while for nonlinear systems it (supposedly) reveals nonlinear causal relation as δ grows smaller.

The authors also suggest a conditional variant of the index, the Conditional Extended Granger Causality Index (CEGCI), to deal with the multivariate case.

Convergent cross mapping (CCM)

Introduced in [Sugihara et al. 2012], CCM is a method for causality inference based on nonlinear state space reconstruction, a mathematical model commonly used in the theory of dynamical systems, and which can be applied to systems where causal variables have synergistic effects (unlike Granger causality tests). The authors demonstrate successful discerning between true coupled variables and the case of external forcing of non-coupled variables.

The method was implemented by some of the authors in the rEDM package [R], the pyEDM package [Python] and cppEDM package [C++], which are accompanied by a comprehensive tutorial.

Information theoretic approaches to causality inference in time series data 📖

Most of the following causality measures were overviewed and compared in [Papana et al. 2013].

Coarse-grained trans-information rate (CTIR)

CTIR is a measure introduced in [Paluš et al. 2001] for the detection of the “direction of information flow” between coupled systems in a bivariate time series scenario, based on *conditional mutual information*.

Defined in information theoretic form, rather than a measure of strength, it measures the average rate of the net amount of information “transferred” from a process Y to the process X, or, in other words, the average rate of the net information flow by which the process Y influences the process X.

[Hlaváčková-Schindler et al. 2007] provide an extremely thorough overview of both CTIR and conditional mutual information, and the various methods used to estimate them. The same paper also includes a proof that the two measures are equivalent, given proper conditioning.

Transfer entropy measures

The information theoretic concept of *transfer entropy* was introduced in [Schreiber, 2000] as a measure quantifying the statistical coherence between systems evolving in time in a way that can distinguish and exclude information that is actually exchanged from shared information due to common history and input signal. Alternatively, it can be said to quantify the the amount of information explained in X_1 at h steps ahead from the state of X_2, accounting for the concurrent state of X_1. *Transfer entropy* is given by:

$$\begin{aligned}\mathrm{TE}_{X_2\to X_1} &= I(x_{1,t+h};\mathbf{x}_{2,t}|\mathbf{x}_{1,t}) = H(x_{1,t+h}|\mathbf{x}_{1,t}) - H(x_{1,t+h}|\mathbf{x}_{2,t},\mathbf{x}_{1,t})\\ &= H(\mathbf{x}_{2,t},\mathbf{x}_{1,t}) - H(x_{1,t+h},\mathbf{x}_{2,t},\mathbf{x}_{1,t}) + H(x_{1,t+h},\mathbf{x}_{1,t}) - H(\mathbf{x}_{1,t})\end{aligned}$$

Equation 18: Transfer Entropy

Where $I(x_1+h;x_2|x_1)$ is the *conditional mutual information*, which gives the expected value of the mutual information of X_1 at h steps and X_2 given the current value of X_1; *H(X)* is *Shannon's entropy;* and *H(X,Y)* is the *joint Shannon's entropy*. The first equivalency was shown in [Paluš and Vejmelka, 2007].

Transfer entropy was since used frequently as a measure of causality in various papers, in fields such as neuroscience (for example[Vicente, 2011]) and extended to use other entropy measures, such as Reyni's, in [Jizba et al. 2012]. [Verdes, 2005] suggested a variant measure better suited for causality detection in homogeneous spatially extended systems.

***Partial transfer entropy* (PTE)**, presented in [Vakorin et al, 2009], is an extension of *transfer entropy* designed to measure the direct causality of X_2 on X_1, conditioning on the remaining variables in *Z:*

$$\mathrm{PTE}_{X_2\to X_1|Z} = H(x_{1,t+h}|\mathbf{x}_{1,t},\mathbf{z}_t) - H(x_{1,t+h}|\mathbf{x}_{2,t},\mathbf{x}_{1,t},\mathbf{z}_t)$$

Equation 19: Partial Transfer Entropy

Symbolic Transfer Entropy (STE): The STE measures amounts to transfer entropy estimated on an embedding space (of dimension *d*) of rank-points (i.e. symbols) formed by the reconstructed vectors of the variables.

$$\mathrm{STE}_{X_2\to X_1} = H(\hat{\mathbf{x}}_{1,t+h}|\hat{\mathbf{x}}_{1,t}) - H(\hat{\mathbf{x}}_{1,t+h}|\hat{\mathbf{x}}_{2,t},\hat{\mathbf{x}}_{1,t})$$

Equation 20: Symbolic Transfer Entropy

Where $\hat{X}_{1}$*,t* is the *ordinal pattern* of order *d* at time *t* of the vector *X_{1},t*, (see [Keller and Sinn, 2005]) which, given time delay *τ*, is defined to be the permutation (r$_0$, r$_1$,⋯, rd) of (0, 1,⋯, d) satisfying

$$x_{t-r_0\tau} \geq x_{t-r_1\tau} \geq \ldots \geq x_{t-r_{d-1}\tau} \geq x_{t-r_d\tau}$$

Partial symbolic transfer entropy (PSTE): STE was extended to multivariate settings in an identical way to PTE:

$$\mathrm{PSTE}_{X_2\to X_1|Z} = H(\hat{\mathbf{x}}_{1,t+h}|\hat{\mathbf{x}}_{1,t},\hat{\mathbf{z}}_t) - H(\hat{\mathbf{x}}_{1,t+h}|\hat{\mathbf{x}}_{2,t},\hat{\mathbf{x}}_{1,t},\hat{\mathbf{z}}_t)$$

Equation 21: Partial Symbolic Transfer Entropy

Additional transfer entropy based measures for causality include transfer entropy on rank vectors (TERV), introduced in [Kugiumtzis, 2012], and its multivariate extension partial transfer entropy on ranks (PTERV), introduced in [Kugiumtzis, 2013A].

Mutual Information from Mixed Embedding (MIME)

Introduced in [Vlachos and Kugiumtzis, 2010], MIME is mutual information driven state space reconstruction technique for time series analysis, including causality (and thus could also be placed under the dynamic system theory approaches section).

In the bivariate case the scheme gives a mixed embedding of varying delays from the variables, *X_1* and *X_2*, that best explains the future of *X_1*. The mixed embedding vector, *W[t]*, may contain lagged components of both *X_1* and *X_2*, defining two complementary subsets *W[t]=[W^{x1}t, W^{x2}t]*. The MIME is then estimated as:

$$\mathrm{MIME}_{X_2\to X_1} = \frac{I(\mathbf{x}^h_{1,t};\mathbf{w}^{X_2}_t|\mathbf{w}^{X_1}_t)}{I(\mathbf{x}^h_{1,t};\mathbf{w}_t)}$$

Equation 22: Mutual Information from Mixed Embedding

The numerator in Eq. 22 is the conditional as for the TE in Eq. 18, but for non-uniform embedding vectors of X_1 and X_2. MIME can thus be considered as a normalized version of the TE for optimized non-uniform embedding of X_1 and X_2.

Partial Mutual Information from Mixed Embedding (PMIME) is an extension of MIME for multivariate settings, described in [Kugiumtzis, 2013B], done by additionally conditioning on all environment variables Z, much like in PTE and PSTE. The mixed embedding vector that best describes the future of X_1 is now formed potentially by all K lagged variables, i.e., X_1, X_2 and the other K-2 variables in Z, and it can be decomposed to the three respective subsets as $W[t]=[W^{x1}t, W^{x2}t, W^{z}t]$. The PMIME is then estimated as:

$$\mathrm{PMIME}_{X_2 \to X_1 | Z} = \frac{I\left(\mathbf{x}_{1,t}^{h}; \mathbf{w}_t^{X_2} | \mathbf{w}_t^{X_1}, \mathbf{w}_t^{Z}\right)}{I\left(\mathbf{x}_{1,t}^{h}; \mathbf{w}_t\right)}$$

Equation 22: Partial Mutual Information from Mixed Embedding

The method was implemented by the authors in a Matlab package.

. . .

Graphical approaches for causality inference in time series data 📖

A graphical approach is often used to model Granger causality in multivariate setting: Each variable (in our case, corresponding to a time series) is considered to be a node in a Granger network, with directed edges denoting a causal link, possibly with a delay (see Figure 2).

Causal graph search algorithms (SGS, PC and FCI)

A family of causal search algorithms that use principles of conditional dependence and application of the causal Markov condition[7] to reconstruct the causal graph of the data-generating process, made up of three related algorithms: SGS, PC and FCI. See [Spirtes et al. 2000] for a thorough overview.

The main structure of these algorithms is similar:

1. **Initialization:** The full undirected graph over all variables V is initialized (i.e. assuming all causal connections).
2. **Skeleton Construction:** Then, edges are eliminated by testing for conditional independence with increasing degrees of dependence (here the algorithms differ; SGS tests for every possible conditioning set, while PC only includes connected variables).
3. **Edge elimination:** Finally, a set of statistical and logical rules are applied to determine the direction of edges (i.e. causality) in the graph.

Between the first two, SGS is considered as possibly more robust to nonlinearities, while the complexity of PC — the more commonly used of the two — does not grow exponentially with the number of variables (as a result of the difference in the edge elimination phase). Finally, the PC algorithm cannot handle unobserved confounders, a problem which its extension, FCI, aims to remedy.

> *[Runge et al, 2017] deem PC to be inappropriate to use with time series data, claiming the use of autocorrelation can lead to high false positive rates based on numerical experiments.*

PCMCI

PCMCI is a causal discovery method described in [Runge et al, 2017], and implemented in the *Tigramite* Python package. The authors claim it is suitable for large datasets (~*O(100k)*) of variables featuring linear and nonlinear, time-delayed dependencies, given sample sizes of a few hundreds or more, and that is shows consistency and higher detecting power with reliable false positive control, when compared with methods such as Lasso Granger and the CI family of algorithms.

The method consists of two stages:

1. PC₁ — A Markov set discovery algorithm based on the PC algorithm that removes irrelevant conditions for each variable by iterative independence testing.
2. MCI — The *momentary conditional independence* test, meant to addresses false positive control for the highly-interdependent time series case, conditions on the parents of both variables in the potential causal link. To test whether X^i affects X^j with lag τ, the following is tested (where $\mathscr{P}(X^i)$ is the set of parent nodes of X^i):

$$MCI: X^i \perp\!\!\!\perp X^j | \mathscr{P}(X^j) \setminus \{X^i\}, \mathscr{P}(X^i)$$

$$MCI: \quad X_{t-\tau} \perp\!\!\!\perp X_t \,|\, \mathcal{P}(X_t)\setminus\{X_{t-\tau}\}, \mathcal{P}(X_{t-\tau})$$

The MCI test

Like in the skeleton construction phase of the PC-family of algorithms, both steps of PCMCI can be combined with any conditional independence test. The authors examine linear partial correlation tests for the linear case, and the GPDC and CMI tests for the non-linear case.

Lasso-Granger

This method was introduced in [Arnold et al. 2007] as way to apply Granger causality models in high-dimensional multivariate settings, by utilizing the variable selection nature of the Lasso method.

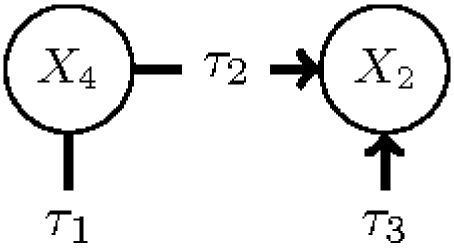
The method was also adapted to deal with *irregular* time series (series with samples missing at blocks of sampling points or collected at non-uniformly spaced time points) in [Bahadori and Liu, 2012A] and [Bahadori and Liu, 2012B] with the Generalized Lasso Granger (GLG) and the Weighted Generalized Lasso Granger (W-GLG) variants.

A thorough review of Lasso-Granger methods, although in the specific context of gene expression regulatory networks, is given in [Hlaváčková-Schindler and Pereverzyev, 2015].

Copula-Granger

Copula-Granger is a semi-parametric Granger causality inference algorithm developed and introduced in [Bahadori and Liu, 2012B] and [Bahadori and Liu, 2013]. The copula approach was first suggested for time series analysis in [Embrechts et al. 2002], and later used in [Liu et al, 2009] to learn the dependency graph among time series.

The authors examine in [Bahadori and Liu, 2013] both two existing approaches and their algorithm in terms of two main properties: (1) The ability to handle *spurious effects* of confounders, and (2) *consistency*[6]. The entire analysis is done under the strong assumption of *Causal Sufficiency* — i.e. that no common cause of any two observed variables in the system is left out.



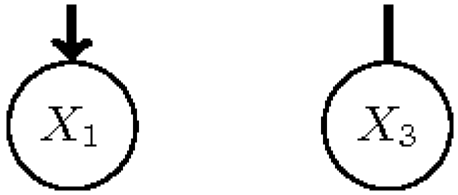


Figure 2: A toy Granger graphical model, with delays τ_i. When X_4 is unobserved, a spurious edge $X_1 \leftarrow X_3$ is detected by some algorithms.

The authors highlight deficiencies of two major approaches for VAR models — *significance test* and *Lasso-Granger* — in the above terms, and show that their approach is both consistent in high dimensions, and can capture non-linearity in the data (for a simple polynomial case). The two main novelties in the proposed approach is the explicit treatment of delays in causality paths in the graphs, used to prevent identification of spurious effects (see Figure 2), and the projection of observations into *copula space* alongside the incorporation of *non-paranormal* (nonparametric normal) distributions into the DGP.

It was generalized, in the same way that elastic net generalizes lasso, in [Furqan et al. 2016], to use the elastic net regularization method, to overcome the natural limitations of lasso: instability when used for high-dimensional data and limited variable selection before saturation when the number of variables is greater than the number of observation points.

Forward-Backward Lasso Granger (FBLG)

Both Lasso Granger and Copula-Granger were extended in [Cheng et al. 2014] with a bagging-like meta-algorithm called Forward-Backward, which enriches the dataset with a reversal of the input time series.

. . .

Choosing which approach to use 📖

In general, the decision on which approach to infer or detect causality in your data will be driven mainly by the data itself and it characteristics, and by what assumptions you feel confident you can make about it and the real processes generating it.

Granger causality vs other approaches

The key requirement of Granger causality is separability, meaning that information about causing effects is not contained in time series for the caused effects, and can be eliminated by removing that variable from the model.

Generally, separability is characteristic of purely stochastic and linear systems, and Granger causality can be useful for detecting interactions between strongly coupled (synchronized) variables in nonlinear systems. Separability reflects the view that systems can be understood a piece at a time rather than as a whole. [Sugihara et al. 2012]

As such, the first criteria to using classical Granger-causality-based approaches is the ability to separate your data into several mutually-exclusive (information-wise) time series, for which the ability to determine that several specific time series cause some other specific time series is valuable.

In complex dynamic systems where these conditions cannot be met, modern approaches meant to infer causality in such systems, like CCM or PCMCI, might be more appropriate.

Parametric vs non-parametric approaches

Model misspecification is always a challenge in causal inference, whether the selected system model is linear or non-linear. If you believe non of the available methods can model the system under question and the flow of causality in it to a good degree— a determination usually made using domain knowledge and intuition — then a non-parametric approach might be more appropriate, such as most of those presented in the sections dealing with dynamic system theoretic and information theoretic causality measures.

A notable caveat, that is often not obvious at first sight, is that the requirements or assumptions some approaches make can render a non-parametric approach parametric in practice. A notable example is PCMCI, which assumes input time series are generated by stationary processes. As non-parametric tests for stationarity (unlike unit root) are few and far between, not to mention that no method or transformation can guarantee the transformation of non-stationary data to stationary one, this assumption will force the user PCMCI to use parametric approaches for the detection and transformation of non-stationarity in input data. This situation is made worse by the lack of accepted and well-defined notions of near-stationarity (a few do exist) and of ways to quantify it and determine when it is sufficient for an inference method to function properly.

Causal graph extraction

If the extraction of a causal graph is a goal, then the PCMCI and Copula-Granger (and its extension FBCLG) stand out among the graphical algorithms. Both methods can handle confounders successfully, with PCMCI also claiming resiliency to high autocorrelation in the data and boasting a convenient Python implementation.

System observability

The observability of the system is also a parameter to be taken in to account. If the strong assumption of *causal sufficiency* cannot be met, then the many methods that presuppose it — including PCMCI and Copula-Granger — cannot be relied upon for correct inference of causal relations. In that case, alternative causality measures aimed at dealing with latent variables in the system, such as PGCI, should be considered.

Deciding between different tests for Granger causality

The takeaway here is pretty simple: Unless you can justify the very strong assumption of a linear relationship between the exogenous and the endogenous variables, a non-parametric approach is the proper one, as it makes much weaker assumptions on your stochastic system and on the flow of causality.

In that case, the Diks-Panchenko test stands out among the non-parametric tests for Granger causality, in terms of power and size[5] of the test. It also solves the discrepancy between the definition of Granger causality and the actual relationship tested by the Hiemstra-Jones test, which is not solved even by the Bai et al variants of the test.

If a linear model of the system is sufficient, then the Toda and Yamamoto procedure is the most rigorous method for linear Granger causality inference, dealing with important phenomena such as integrated or cointegrated time series.

Researchers to follow 📖

Prof. Cees Diks consistently publishes papers on non-linear Granger causality and non-linear dynamics in general. This include, among other topics, building Financial networks based on Granger causality, examining the effect of different resampling methods for causality testing and causality measures for multivariate analysis.

Prof. Dimitris Kugiumtzis does incredible work on time series analysis generally, and causality inference in time series data specifically, driven by information theoretic approaches, notably the MIME and PMIME measures.

Prof. George Sugihara is a theoretical biologist who has worked across a variety of fields, introducing inductive theoretical approaches to understanding complex dynamic systems in nature from observational data. Chief among those is *Empirical Dynamic Modeling*, a non-parametric approach for the analysis and forecast of complex dynamical systems rooted in chaos theory, represented in this post by the CCM method. His work involves inductive theoretical approaches to understanding nature from observational data.

Dr. Jakob Runge did substantial work on causality in time series data, mainly in the context of climate research; he is also the creator of tigarmite, a Python library for causal inference in time series data using the PCMCI method.

Youssef Hmamouche is one of the authors and maintainers of the NlinTS R package for neural network-based time series forecasting and causality detection in time series data, and recently wrote about a Causality-Based Feature Selection Approach For Multivariate Time Series Forecasting.

. . .

Other notable literature 📖

Learning and causal inference

Judea Pearl, who is a prominent researcher in the field and who has developed the structural approach to casual inference, recently wrote a very interesting piece on causal inference tools and reflections on machine learning [Pearl, 2018]. He also wrote a thorough overview of the topic of causal inference [Pearl, 2009].

David Lopez-Paz, a research scientist at Facebook AI Research, leads very interesting research on casual inference in general, and in the context of learning frameworks and deep learning specifically. Highlights include posing causal inference as a learning problem (specifically of classifying probability distributions), causal generative neural networks, incorporation of an adversarial framework for causal discovery and discovering causal signals in images.

Krzysztof Chalupka has done some fascinating research in the intersection of deep learning and causal inference. Highlights include a deep-learning-based conditional independence test, causal feature learning, visual causal feature learning and causal regularization.

Finally, [Dong et al. 2012] have used Multi-Step Granger Causality Method (MSGCM), a method developed to identify feedback loops embedded in biological networks using time-series experimental measurements, for the identification of feedback loops in neural networks.

. . .

References 📖

Academic Literature: Causality and causality inference

- [Eberhardt, 2007] Eberhardt, F. (2007), *Causation and Intervention*, (Ph.D. Thesis), Carnegie Mellon University.

- [Geweke et al. 1983] Geweke, J., Meese, R., and Dent, W. (1983), *Comparing Alternative Tests of Causality in Temporal Systems: Analytic Results and Experimental Evidence*, Journal of Econometrics, 21, 161–194.

- [Pearl, 1993], Pearl, J. *[Bayesian Analysis in Expert Systems] Comment: Graphical Models.* Causality and Intervention. Statist. Sci., Volume 8, Number 3 (1993), 258–261.

- [Pearl, 2000] Pearl, J. *Causality: Models, Reasoning, and Inference*. Oxford University Press, 2000.

- [Pearl, 2009] Pearl, J. *Causal Inference in Statistics : An Overview.* Statistics Surveys, Vol. 0, 2009.

- [Pearl, 2018] Pearl, J. The Seven Tools of Causal Inference with Reflections on Machine Learning. Communications of Association for Computing Machinery, 2018.

- [Pearl and Verma, 1991] Pearl, J. and Verma, T.S. *A Theory Of Inferred Causation*. Second International Conference conference on the Principles of Knowledge Representation and Reasoning, Cambridge, Massachusetts, April 1991.

Academic Literature: Causality inference in time series data

- [Ancona et al. 2004] Ancona N, Marinazzo D, Stramaglia S. Radial basis function approach to nonlinear granger causality of time series. Phys Rev E 2004;70:056221.

- [Ancona and Stramaglia, 2006] Ancona, N., & Stramaglia, S. (2006). *An invariance property of predictors in kernel-induced hypothesis spaces*. *Neural Computation, 18*, 749–759.

- [Arnold et al. 2007] A. Arnold, Y. Liu, and N. Abe. Temporal causal modeling with graphical Granger methods. In Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 66–75, USA, 2007. ACM New York

- [Bahadori and Liu, 2012A] Bahadori, M. T., and Liu, Y. 2012. *Granger causality analysis in irregular time series*. In SIAM International Conference on Data Mining (SDM 2012). SIAM. [PDF]

- [Bahadori and Liu, 2012B] Bahadori, M.T. and Liu Y. *On Causality Inference in Time Series*. Discovery Informatics: The Role of AI Research in Innovating Scientific Processes, Papers from the 2012 AAAI Fall Symposium, 2012.

- [Bahadori and Liu, 2013] M. T. Bahadori and Y. Liu. An examination of practical granger causality inference. In SDM, 2013. [PDF]

- [Bai et al. 2010] Bai ZD, Wong WK, Zhang BZ. Multivariate linear and nonlinear causality tests. Mathematics and Computers in simulation. 2010; 81: 5–17. doi: 10.1016/j.matcom.2010.06.008

- [Bai et al. 2016] Bai ZD, Hui YC, Lv ZH, Wong WK, Zhu ZZ. *The Hiemstra-Jones Test Revisited*. 2016; arXiv:1701.03992.

- [Bai et al. 2018] Bai, Z., Hui, Y., Jiang, D., Lv, Z., Wong, W.K. and Zheng, S., 2018. *A new test of multivariate nonlinear causality*. PloS one, 13(1), p.e0185155.

- [Baek, 1992] Baek, E.G. and Brock, A.W. (1992). *A General Test for Non-Linear Granger Causality: Bivariate Model*.

- [Bell et al. 1996] Bell, D., Kay, J. and Malley, J. (1996). A non-parametric approach to non-linear causality testing. Economics Letters, 51, 7–18.

- [Bouezmarni and Taamouti, 2010] Bouezmarni, T. and Taamouti, A. *Nonparametric Tests for Conditional Independence Using Conditional Distributions*. Journal of Nonparametric Statistics, Volume 26, 2014 — Issue 4.

- [Chen et al. 2004] Chen Y, Rangarajan G, Feng J, Ding M. Analyzing multiple nonlinear time series with extended Granger causality. Phys Lett A 2004;324:26–35.

- [Cheng et al. 2014] Cheng D, Bahadori MT, Liu Y. FBLG: a simple and effective approach for temporal dependence discovery from time series data. Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining; New York, New York, USA. 2623709: ACM; 2014. p. 382–91.

- [Diks and Panchenko, 2005] Diks, C. and Panchenko, V. *A Note on the Hiemstra-Jones Test for Granger Non-causality*. Studies in Nonlinear Dynamics & Econometrics, Volume 9, Issue 2.

- [Diks and Panchenko, 2006] Diks C, Panchenko V. 2006. *A new statistic and practical guidelines for nonparametric Granger causality testing.* Journal of Economic Dynamics and Control 30(9–10): 1647–1669. (PDF)

- [Diks and Wolski, 2015] Diks, C., & Wolski, M. (2015). *Nonlinear granger causality: Guidelines for multivariate analysis*. Journal of Applied Econometrics. (PDF)

- [Eichler, 2011] Eichler, M. *Causal Inference in Time Series Analysis.* Chapter 22 in *Causality: Statistical Perspectives and Applications,* 2011. (PDF)

- [Eichler and Didelez, 2007] Eichler, M. and Didelez V. *Causal Reasoning in Graphical Time Series Models*. Proceedings of the 23rd Conference on Uncertainty in Artificial Intelligence.

- [Embrechts et al. 2002] Embrechts, P.; Mcneil, A.; and Straumann, D. 2002. Correlation and dependence in risk management: properties and pitfalls. In Dempster, M. H. A., ed., Risk Management: Value at Risk and Beyond. Cambridge: Cambridge University Press.

- [Florens and Mouchart, 1982] Florens, J. P. and Mouchart, M. (1982). *A note on noncausality*. Econometrica 50, 583– 591.

- [Florens and Mouchart, 1985] Florens, J. P. and Mouchart, M. (1985). *A linear theory for noncausality*. Econometrica 53, 157–175. Good, I. J. (1961)

- [Furqan et al. 2016] Furqan MS, Siyal MY. Elastic-Net copula granger causality for inference of biological networks. PLoS One. 2016;11(10):e0165612.

- [Geweke, 1984] Geweke J. Measures of conditional linear dependence and feedback between time series. J Am Stat Assoc 1984;79:907–15.

- [Ghysels et al. 2016] Ghysels E, Hill JB and Motegi K. *Testing for Granger causality with mixed frequency data*. Journal of Econometrics. 2016; 192(1): 207–230. doi: 10.1016/j.jeconom.2015.07.007

- [Granger, 1969] Granger, C. W. J. *Investigating causal relations by econometric models and crossspectral methods*. Econometrica 37, 424–438.

- [Granger, 1980] Granger, C. W. J. *Testing for causality, a personal viewpoint*. Journal of Economic Dynamics and Control 2, 329–352.

- [Granger, 1988] Granger, C. W. J. (1988). *Some recent developments in a concept of causality*. Journal of Econometrics 39, 199–211.

- [Greene, 2002] Greene, W. (2002) *Econometric analysis*, 5th ed. Prentice Hall, Upper Saddle River.

- [Guo et al. 2008] Guo, S., Seth, A.K., Kendrick, K.M., Zhou, C., Feng, J., 2008. Partial Granger causality — eliminating exogenous inputs and latent variables. J. Neurosci. Meth. 172 (1), 79–93

- [Hafner and Herwartz, 2009] Christian M. Hafner & Helmut Herwartz, 2009. *Testing for linear vector autoregressive dynamics under multivariate generalized autoregressive heteroskedasticity*. Statistica Neerlandica, Netherlands Society for Statistics and Operations Research, vol. 63(3), pages 294–323.

- [Hiemstra and Jones, 1994] Hiemstra, C. and Jones, J. D. *Testing for linear and nonlinear Granger causality in the stock price-volume relation*. Journal of Finance. 1994; 49(5): 1639–1664. doi: 10.2307/2329266 [PDF]

- [Hlaváčková-Schindler et al. 2007] Hlaváčková-Schindler, K., Palus, M., Vejmelka, M., & Bhattacharya, J. (2007). Causality detection based on information-theoretic approaches in time series analysis. *Physics Reports, 441*, 1–46.

- [Hlaváčková-Schindler and Pereverzyev, 2015] Hlaváčková-Schindler, K., and Pereverzyev, S. (2015). "*Lasso granger causal models: some strategies and their efficiency for gene expression regulatory networks*," in *Decision Making: Uncertainty,* *Imperfection, Deliberation and Scalability Studies in Computational Intelligence*, eds T. Guy, M. Kárný, and D. Wolpert (Cham: Springer), 91–117. doi: 10.1007/978–3–319–15144–1_4

- [Hosoya, 1977] Hosoya, Y. (1977). *On the Granger condition for non-causality*. Econometrica 45, 1735– 1736.
- [Hurlin and Venet, 2001] Hurlin C. and Venet B. (2001). *Granger causality tests in panel data models with fixed coefficients*. EURIsCO Université Paris Dauphine.
- [Jizba et al. 2012] Jizba, P.; Kleinert, H.; Shefaat, M. *Rényi's information transfer between financial time series*. *Physica A: Stat. Mech. Appl.* **2012**, *391*, 2971–2989.
- [Kugiumtzis, 2012] Kugiumtzis, D. Transfer entropy on rank vectors. J. Nonlinear Syst. Appl. 2012, 3, 73–81.
- [Kugiumtzis, 2013A] Kugiumtzis, D. Partial transfer entropy on rank vectors. Eur. Phys. J. Spec. Top. 2013, 222, 401–420
- [Kugiumtzis, 2013B] Kugiumtzis, D. *Direct-coupling information measure from nonuniform embedding*. *Phys. Rev. E* **2013**, *87*, 062918.
- [Kugiumtzis and Kimiskidis, 2015] Kugiumtzis D, Kimiskidis VK (2015) Direct causal networks for the study of transcranial magnetic stimulation effects on focal epileptiform discharges. Int J Neural Syst 25(5):1550006
- [Liu et al, 2009] Liu, H.; Lafferty, J. D.; and Wasserman, L. A. 2009. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. JMLR
- [Lütkepohl and Müller, 1994] Lütkepohl, H. and Müller, M. (1994). *Testing for Multi-Step Causality in Time Series*.
- [Lütkepohl, 2007] Lütkepohl, H. *New Introduction to Multiple Time Series Analysis*. 2007.
- [Marinazzo et al. 2006] Marinazzo D, Pellicoro M, Stramaglia S (2006) *Nonlinear parametric model for Granger causality of time series*. Phys Rev E Stat Nonlin Soft Matter Phys 73: 066216.
- [Paluš et al. 2001] M. Paluš, V. Komárek, Z. Hrnˇcíˇr, K. Štˇerbová, Synchronization as adjustment of information rates: detection from bivariate time

series, Phys. Rev. E 63 (2001) 046211.[PDF]

- [Paluš and Vejmelka, 2007] Paluš M, Vejmelka M. Directionality from coupling between bivariate time series: how to avoid false causalities and missed connections. Phys Rev E 2007;75:056211.
- [Papana et al. 2013] Papana, A., C. Kyrtsou, D. Kugiumtzis, and C. Diks. 2013. " *Simulation Study of Direct Causality Measures in Multivariate Time Series*. " Entropy 15: 2635–2661.
- [Papana et al. 2017] Papana, A., Kyrtsou, C., Kugiumtzis, D., & Diks, C. (2017). *Financial networks based on Granger causality: A case study*. Physica A: Statistical Mechanics and Its Applications, *482, 65–73.* doi:10.1016/j.physa.2017.04.046
- [Popescu and Guyon, 2009] Popescu, F. and Guyon, I. *Causality in Time Series*. Challenges in Machine Learning, Volume 5; based on the MiniSymposium on Time Series Causality at NIPS 2009.
- [Price, 1979] Michael Price, J. (1979). *The characterization of instantaneous causality*. Journal of Econometrics, 10(2), 253–256. doi:10.1016/0304–4076(79)90009–5
- [Runge, 2014] Runge, J. *Detecting and Quantifying Causality from Time Series of Complex Systems*. Ph.D. dissertation. (PDF link)
- [Runge et al, 2017] Runge, J. Nowack, P. Kretschmer, M. Flaxman, S. and Sejdinovic, D. *Detecting causal associations in large nonlinear time series datasets*. arXiv.org.
- [Runge, 2018] Runge, J. *Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information*. Proceedings of the 21st International Conference on Artificial Intelligence and Statistics (AISTATS) 2018. PMLR: Volume 84.
- [Samartsidis et al., 2018] Samartsidis, P., Seaman, S. R., Presanis, A. M., Hickman, M. and De Angelis, D. *Review of methods for assessing the causal effect of binary interventions from aggregate time-series observational data*. arXiv.org.
- [Schreiber, 2000] T. Schreiber, Measuring information transfer, Phys. Rev. Lett. 85 (2000) 461–464.

- [Siggiridou and Kugiumtzis, 2016] Siggiridou, E. & Kugiumtzis, D. *Granger causality in multivariate time series using a time-ordered restricted vector autoregressive model*. IEEE Transactions on Signal Processing **64**, 1759–1773, doi:10.1109/TSP.2015.2500893 (2016).

- [Sims, 1972] Sims, C. A. (1972). *Money, income and causality*. American Economic Review 62, 540–552.

- [Spirtes et al. 2000] P. Spirtes, C. Glymour, and R. Scheines. Causation, Prediction, and Search. The MIT Press, Cambridge, Massachusetts, London, England, 2000.

- [Su and White, 2003] Su, L. and White, H. (2003). A nonparametric Hellinger metric test for conditional independence. Technical Report. Department of Economics, UCSD.

- [Sugihara et al. 2012] Sugihara, G., May, R., Ye, H., Hsieh, C. -h., Deyle, E., Fogarty, M., & Munch, S. (2012). Detecting Causality in Complex Ecosystems. Science, 338(6106), 496–500. doi:10.1126/science.1227079 [PDF]

- [Toda and Yamamoto, 1995] Toda, H. Y., & Yamamoto, T. (1995). *Statistical inference in vector autoregressions with possibly integrated processes*. Journal of Econometrics, 66(1–2), 225–250. doi:10.1016/0304–4076(94)01616–8

- [Vakorin et al, 2009] Vakorin, V.A.; Krakovska, O.A.; McIntosh, A.R. Confounding effects of indirect connections on causality estimation. *J. Neurosci. Methods* **2009**, *184*, 152–160.

- [Verdes, 2005] P.F. Verdes, Assessing causality from multivariate time series, Phys. Rev. E 72 (2005) 026222.

- [Vicente, 2011] Vicente, R, Wibral, M, Lindner, M, Pipa, G Transfer entropy — A model-free measure of effective connectivity for the neurosciences.. *J Comput Neurosci*. (2011). *30* 45–67

- [Vlachos and Kugiumtzis, 2010] Vlachos, I.; Kugiumtzis, D. Non-uniform state space reconstruction and coupling detection. *Phys. Rev. E* **2010**, *82*, 016207.

- [White, 2006] White, H. *Time-series estimation of the effect of natural experiments*. Journal of Econometrics, 135, pp. 527–566.

- [White and Lu, 2010] White, H. and Lu, X. *Granger Causality and Dynamic Structural Systems*. Journal of Financial Econometrics 8, 193–243.

Academic Literature: Other

- [Baccalá and Sameshima, 2001] Baccalá L , Sameshima K. *Partial directed coherence: a new concept in neural structure determination*. Biol Cybern 84: 463–474, 2001. [PDF]

- [Dong et al. 2012] Chao-Yi Dong, Dongkwan Shin, Sunghoon Joo, Yoonkey Nam, Kwang-Hyun Cho. Identification of feedback loops in neural networks based on multi-step Granger causality. Bioinformatics. 2012 Aug 15; 28(16): 2146–2153. Published online 2012 Jun 23. doi: 10.1093/bioinformatics/bts354

- [Donges at al. 2009] Donges, J. F., Y. Zou, N. Marwan, and J. Kurths (2009). *The backbone of the climate network*. Europhysics Letters 87.4 (cit. on p. 12) (in arXiv version cit. on p. 2).

- [Keller and Sinn, 2005] K. Keller and M. Sinn, Ordinal analysis of time series, Physica A 356 (2005) 114–120.

- [Malik et al. 2012] Malik, N., B. Bookhagen, N. Marwan, and J. Kurths (2012). *Analysis of spatial and temporal extreme monsoonal rainfall over South Asia using complex networks*. Climate Dynamics 39.3–4, pp. 971–987 (cit. on p. 13).

- [Pikovsky et al. 2003] Pikovsky, A., M. Rosenblum, and J. Kurths (2003). *Synchronization: a universal concept in nonlinear sciences*. Vol. 12. Cambridge: Cambridge Univ Press (cit. on pp. 12, 18).

- [Radebach et al. 2013] Radebach, A., R. V. Donner, J. Runge, J. F. Donges, and J. Kurths (2013). *Disentangling different types of El Niño episodes by evolving climate network analysis*. Physical Review E 88.5, p. 052807 (cit. on p. 13).

- [Saito and Harashima, 1981] Saito Y, Harashima H (1981) Tracking of information within multichannel EEG record — Causal analysis in EEG. In: Yamaguchi N, Fujisawa K, editors. Recent Advances in EEG and EMG data processing. Amsterdam: Elsevier. pp. 133–146.

- [Tsonis and Roebber, 2004] Tsonis, A. A. and P. J. Roebber. *The architecture of the climate network*. Physica A Volume 333, 15 February 2004, Pages 497–504.

- [Wiener, 1956] Wiener N. The theory of prediction. In: Beckenbach EF, editor. Modern mathematics for engineers. New York: McGraw-Hill; 1956 [chapter 8].

- [Yamasaki et al. 2008] Yamasaki, K., A. Gozolchiani, and S. Havlin (2008). *Climate Networks around the Globe are Significantly Affected by El Nino*. Physical Review Letters 100.22, p. 228501 (cit. on p. 13).

Other Online Sources

- The homepage of Prof. Cees Diks.
- The homepage of David Lopez-Paz
- Granger Causality on Scholarpedia
- F-test on Barbara Webb's website
- Wikipedia: G-test, Mutual Information, Likelihood-ratio test, F-test, Granger causality and other articles.
- Likelihood-ratio test on the Engineering Statistics Handbook
- The documentation and source code of the *grangercausalitytests* method of the *statsmodels* Python package.
- The source code of the *lmtest* R package.
- FAQ: How are the likelihood ratio, Wald, and Lagrange multiplier (score) tests different and/or similar?
- Testing for Granger Causality on Econometrics Beat: Dave Giles' Blog
- Chapter 8.1 of Forecasting: Principles and Practice by Rob J Hyndman and George Athanasopoulos, a great online textbook on forecasting.

Footnotes 📖

1. For a brief overview of the F-test see here and here. ↺
2. For an overview of the Chi-squared test see the Wikipedia article on the topic. ↺
3. In *lmtest*, the *grangertest* method calls the *waldtest* method without assigning a value to its *test* parameter (which determines whether an F test or a chi-square test is applied), so an F test is used by default. ↺
4. Geweke at al. performed a comparison of 8 methods for inferring causality in time series data [Geweke et al. 1983] and found that Wald variants of a test attributed

to Granger, and a lagged dependent variable version of Sims’ test introduced in that paper, are equivalent in all relevant respects and are preferred to the other tests discussed. ↺

5. The size of a statistical test is the probability of it making a Type I error; i.e. falsely rejecting the null hypothesis. ↺
6. A method is consistent if its probability of errors goes to zero as the number of observations increase. ↺
7. The Causal Markov Condition: A variable in a graph is, conditional on its parents, probabilistically independent of all other variables that are neither its parents nor its descendants. ↺
8. ***Ψ-causality* vs *Ψ-non-causality*:** The same definition of causality can sometimes be referred to in two seemingly contradictory names; e.g. Granger causality and Granger non-causality refer to the same definition of causality. This is the case because in many cases the definitions are given for the inverse condition, and *X* is said to be *Ψ*-causing *Y* if the given condition does not hold. I use only the first form for consistency. ↺
9. **A note on notation:** I try to keep notation as close to the source material as possible, but as Medium does not support inline math expressions, I do the best I can with Unicode characters. Specifically, square brackets are used repeatedly where subscript would have been used but is not available in Unicode; for example, the *i*-th element of a vector *v* will be denoted by v_i, but the *t*-th element will be denoted by *v[t]*. ↺

Thanks to Michal Eisenberg.

Statistics Causality Causal Inference Data Science Time Series Analysis