Understanding causation via linear response theory

Angelo VULPIANI

Dept. of Physics, University of Rome "Sapienza", Italy

Rome, January 2022

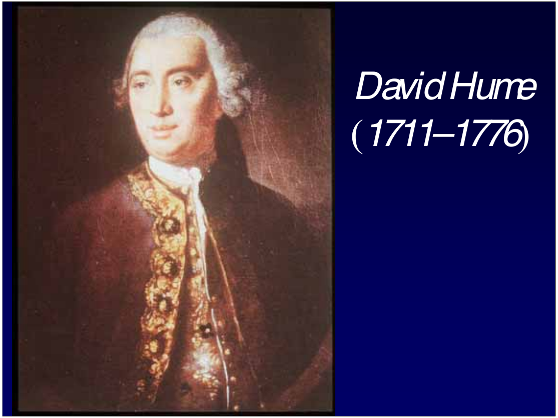
MANY THANKS TO

• **Marco BALDOVIN** and **Fabio CECCONI**
for the friendly, intense and fruitful collaboration.
• **Antonello PROVENZALE**
for the collaboration on paleoclimate.
* M.Baldovin, F.Cecconi and A.Vulpiani:
Understanding causation via correlations and linear response theory,
Physical Review Reserch **2**, 043436 (2020)
* M.Baldovin, F.Cecconi, Provenzale and A.Vulpiani:
Extracting causation from paleoclimate data, Work in progress (2021)

• **Erik AURELL** and **Gino DEL FERRARO**
from whom we had the idea to use response theory for the treatment of the causation.

• **MIUR-PRIN2017** *Coarse-grained description for non-equilibrium systems and transport phenomena (CO-NEST)* for the financial support.

The modern study of causation begins with David Hume.



HUME'S LEGACY

I Analytical vs. empirical claims
He made a sharp distinction between analytical and empirical claims - the former are product of thoughts, the latter matter of fact.

II Causal claims are empirical
He classified causal claims as empirical, rather than analytical.

III All empirical claims originate from experience
He identified the source of all empirical claims with human experience, namely sensory input.

Putting II and III together have left philosophers baffled, for over two centuries, over two major riddles:

- What empirical evidence legitimizes a cause-effect connection?

- What inferences can be drawn from causal information? and how?

The general problem

Detection of causality, i.e. to understand wether, and how much, a variable $\{x_t\}$ influences $\{y_t\}$, is an old important topic.
We can mention, in the context of the climate change, the debated relationship between temperature and CO_2.

In the Western Philosophical tradition **Aristotle** postulated four kinds of causes, **Francis Bacon** later retained only two (the material and the efficient one).
The modern study of causation begins with the **David Hume** who *supposes the law to state that there are propositions "A causes B"' where A and B are classes of events; the fact that such laws do not appear in any well-developed science appears unknown to philosophers.* (B. Russell)

On the other hand the Third Law of Newton states that in Nature there is no separation between cause and effect: both classical and quantum **physics knows only interactions, and these are always mutual......**

Technical problems

How to understand if A $\to$ B;
How to quantify the "degree of causation"

• The first natural idea, according to the latin saying **cum hoc ergo propter hoc** (with this, therefore because of this), it seems rather natural to **look at the correlation** between x and y.

On the other hand it is well known that **correlation does not imply causation**, among the many hilarious examples we mention the strong correlation between
a) a country' s chocolate consumption and Nobel Prize victories;
b) shark attacks and ice cream sales.

BEYOND CORRELATION

• One of the first, and most popular attempt , is due to **Granger**, and it is based on the idea that $x^{(j)}$ influences $x^{(k)}$ if the knowledge of the past history of $x^{(j)}$ enhances the ability to predict future values of $x^{(k)}$.

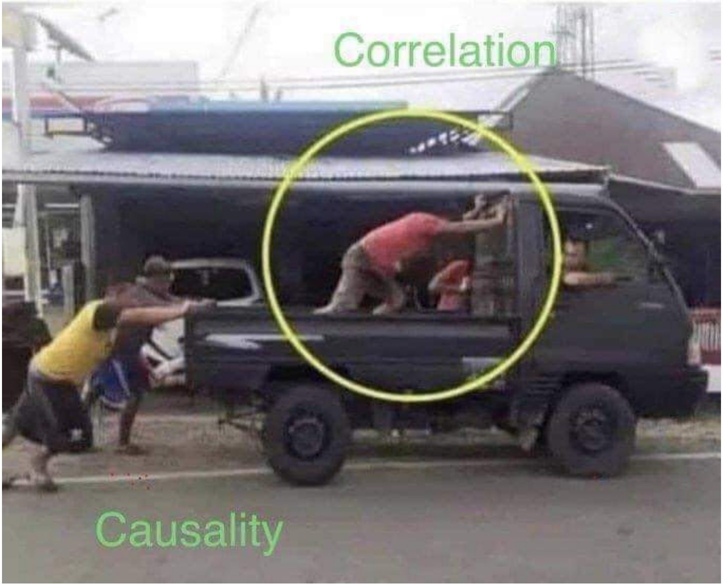
• A different approach consists in defining a degree of information exchange from $x^{(j)}$ to $x^{(k)}$ which quantifies the loss of information about $x^{(k)}$ that one measures if $x^{(j)}$ is ignored, namely the **transfer entropy**, or related quantities in the context of information theory.

• Remarkably, the Granger causality and the transfer entropy have been shown to be equivalent in linear autoregressive systems.

• In more recent studies, the protocol of "convergent cross-mapping" has been able to face the problem of causation characterization by using methods from the **theory of information and dynamical systems**.

A very heavy cat?

The trouble



Interventional and observational causation

We say that two variables can be assumed to be in a cause- effect relationship if an external action on one of them results in a change of the observed value of the second. We call **interventional** this physics- inspired definition of cause- effect relation and **observational** the approach based on tests to determine whether, and to what extent, the knowledge of a certain variable is useful to the actual determination of future values of another.

Let us discuss the **measurement of the electrical current passing trough a resistor**, whose extremities are connected to an external electric potential $v(t)$, if the amperometer is affected by some noise $\eta(t)$ independent of $v(t)$, the measured current is given by

$$J_{meas}(t) = J_{true}(t) + \eta(t) = Gv(t) + \eta(t)$$

J_{true} is the actual (unknown) value of the current and G is the electrical conductance.

A good estimator of the **interventional causality** between $v(t)$ and J_{meas} will only depend on the conductance G since this parameter establishes to which extent an external action on $v(t)$ will influence the observed value of the current, J_{meas}, and G which does not depend on the intensity of the noise.

Conversely, from an observational perspective, the amplitude of the noise also plays a role, since our ability to predict future values of J_{meas} given $v(t)$, crucially depends on it.

If the noise is small, the knowledge of $v(t)$ will suffice to give a good esteem of J_{meas} whereas if it is large, the information about $v(t)$ is almost useless.

• We'll show that linear response theory allows us to understand causal links (in the interventional sense) from time series of data, if the considered process is of Markov type.

A physical definition of causation

Let us consider the system $\mathbf{x}_t = (x_t^{(1)}, x_t^{(2)}, ..., x_t^{(n)})$, where t is a (discrete) time index. We say that $x^{(j)}$ influences $x^{(k)}$ if a perturbation on the variable $x^{(j)}$ at time 0, $x_0^{(j)} \to x_0^{(j)} + \delta x_0^{(j)}$, induces, on average, a change on $x_t^{(k)}$, with $t > 0$.

In formulae, we will say that $x^{(j)}$ has an influence on $x^{(k)}$ if a smooth function $\mathcal{F}(x)$ exists such that

$$\overline{\frac{\delta \mathcal{F}(x_t^{(k)})}{\delta x_0^{(j)}}} \neq 0 \qquad \text{for some } t > 0\,, \tag{1}$$

i.e. if perturbing $x_0^{(j)}$ results in a non-zero average variation of $\mathcal{F}(x_t^{(k)})$ with respect to its unperturbed evolution.
The over-line represents an average over many realizations of the experiment.

Since we will mainly deal with linear Markov systems, considering the identity function $\mathcal{A}(x) = x$ will be sufficient to detect the presence of causal links.

Of course, the response matrix $R_t^{(i,j)}$ can be derived in the numerical computations where the initial conditions can be changed.
Such possibility, at least in principle, can be used in numerical computations, where some details are hidden, which allows to compute $\{\mathbf{x}_{t+1}\}$ from $\{\mathbf{x}_t\}$, where details are hidden, e.g. computations based on neural networks.

This idea to use the response is reminiscent of the framework developed by **Judea Pearl**, in which causation is detected by observing the effects of an action on the system:
Interventions and counterfactuals are defined through a mathematical operator called do(x)*, which simulates physical interventions by deleting certain functions from the model, replacing them with a constant* $X = x$ *while keeping the rest of the model unchanged.*

Fluctuation/response relation: for the computation of the response it is not necessary to perturb the system

If the system admits a (sufficiently smooth) invariant distribution, $P_s(\mathbf{x})$, under rather general conditions, one has

$$R_t^{kj} \equiv \lim_{\delta x_0^{(j)} \to 0} \frac{\overline{\delta x_t^{(k)}}}{\delta x_0^{(j)}} = -\left\langle x_t^{(k)} \frac{\partial \ln P_s(\mathbf{x})}{\partial x^{(j)}} \Big|_{\mathbf{x}_0} \right\rangle , \qquad (2)$$

where the average $\langle \cdot \rangle$ is computed on the unperturbed system, and R_t is the matrix of the linear response functions.

Eq. (2) expresses the link among responses and correlations.
Provided that the functional form of $P_s(\mathbf{x})$ is known, or can be inferred from data (a completely non-trivial task, at least in high-dimensional systems), it is possible to compute the response functions from suitable correlation functions.

A class of systems where the situation is clear

As first we will limit ourselves to the study of linear stochastic processes of the form

$$\mathbf{x}_{t+1} = A\mathbf{x}_t + B\eta_t \tag{3}$$

where A and B are constant $n \times n$ matrices.

In such a class of systems the the following relation between the response matrix and the correlation matrix with entries $C_t^{kj} = \langle x_t^{(k)} x_0^{(j)} \rangle$ holds:

$$R_t = C_t C_0^{-1} \tag{4}$$

where C_0^{-1} is the inverse of C_0.
Therefore **if we know the proper variables which describe the system it is possible to understand causation from correlations**,
The result can be shown to hold also in cases with continuous time.

In order to compare the effect of different "causes", it may be useful to rescale correlations and responses with the standard deviations of the corresponding variables:

$$\tilde{C}_t^{kj} = \frac{1}{\sigma_k \sigma_j} C_t^{kj} \quad \tilde{R}_t^{kj} = \frac{\sigma_j}{\sigma_k} R_t^{kj} \tag{5}$$

Following the idea of **Green-Kubo formula**, which allows to understand the average effect of an electric field on the current in terms of correlations a cumulative "degree of causation" $x^{(j)} \to x^{(k)}$ can be introduced:

$$\mathcal{D}_{j \to k} = \sum_{t=1}^{\infty} \tilde{R}_t^{kj} \, . \tag{6}$$

This quantity characterizes the cumulative effect of the perturbation δx_0^k on the variable $x^{(k)}$.

In linear systems with discrete time, from the relation $R_t = A^t$ it follows that

$$\mathcal{D}_{j\to k} = \frac{\sigma_j}{\sigma_k}\left[A(I_n - A)^{-1}\right]^{kj}, \tag{7}$$

I_n being the $n \times n$ identity matrix.

Let us stress that a vanishing value of $\mathcal{D}_{j\to k}$ does not exclude causation between $x^{(j)}$ and $x^{(k)}$; indeed, since $\tilde{R}_t^{kj}$ can assume both positive and negative values, contributions with opposite signs in sum (6) might eventually compensate and give a null result even in presence of a causal link.

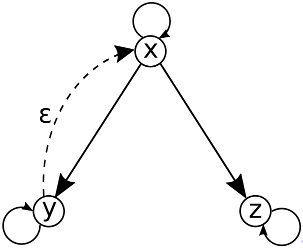
A toy model

$$x_{t+1} = ax_t + \varepsilon y_t + b\eta_t^{(x)} \tag{8a}$$

$$y_{t+1} = ax_t + ay_t + b\eta_t^{(y)} \tag{8b}$$

$$z_{t+1} = ax_t + az_t + b\eta_t^{(z)} \tag{8c}$$

where $\eta^{(x)}, \eta^{(y)}, \eta^{(z)}$ are independent Gaussian processes with zero mean and unitary variance, while a, ε and b are constant parameters.

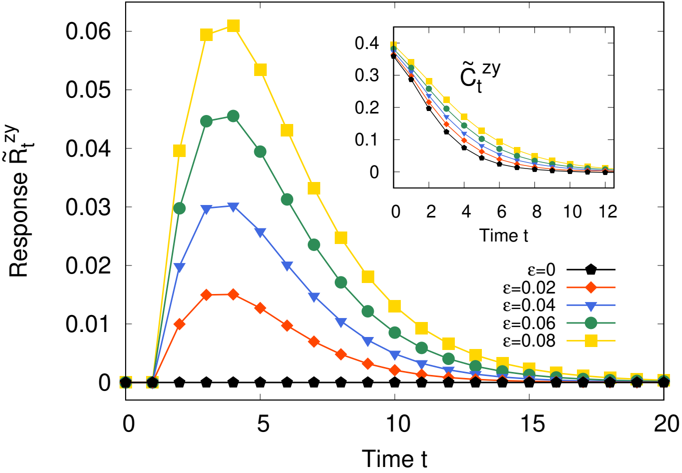


The case $\varepsilon = 0$ is a minimal example in which the behavior of two quantities, y and z, is influenced by a common-causal variable x; as a consequence, y and z are correlated even if there is no causal relationship between them.

The same mechanism may be identified in many cases in which surprising functional dependences arise, as that between the number of Nobel laureates of a country and its chocolate consumption per year in this specific case, both quantities may be expected to be roughly influenced by the gross domestic product of the nation.

According to our definition, in order to decide whether there is a causal relation between y and z, one has to perturb y at time 0 and measure the average variation δz_t for $t > 0$. For $\varepsilon = 0$, not surprisingly, $\tilde{R}_t^{zy} = 0$ for all $t > 0$.

Correlations and Responses ($a = 0.5,\ b = 1$)



The situation completely changes if we allow for a small feedback $\varepsilon \neq 0$ from y to x, which will eventually result in a causal link between y and z.

The corresponding response function correctly reveals that the behavior of z starts to be influenced by a perturbation of y after $t = 2$ time steps, and that the intensity of such causal influence roughly scales with ε.

Remarks:

• None of this conclusions could have been drawn from the simple analysis of the correlation functions.

• Both Granger and transfer entropy approaches would lead to a null causal dependence of z from y for every choice of ε; indeed, once the trajectory of x is known, the knowledge of y does not add any useful information to predict future values of z.

Continuous time and non gaussian effects

The approach can be easily extended to stochastic processes with continuous time of the form

$$\frac{d\mathbf{x}}{dt} = -F\mathbf{x} + B\xi \,, \tag{9}$$

where F and B are $n \times n$ matrices, the eigenvalues of F have positive real part and ξ is a n-dimensional, delta correlated Gaussian noise.

In this case it can be shown that $R_t = \exp(-Ft)$, so that it is again possible to infer F from the study of the response functions, e.g. by considering the matrix $I_n - R_t$ for $t \to 0$, where I_n is the $n \times n$ identity matrix, or by the continuous-time version of Eq. (7),

$$\mathcal{D}_{j\to k} = \frac{\sigma_j}{\sigma_k}[F^{-1}]^{kj} \,. \tag{10}$$

Non-linear effects

Let us now consider a case with non-linear dynamics, a system: $\mathbf{x} = (x, y, z)$ evolves as

$$\dot{x} = -U'(x) - k(x - y) + b\xi^{(x)} \tag{11a}$$

$$\dot{y} = -U'(y) - k(y - x) - k(y - z) + b\xi^{(y)} \tag{11b}$$

$$\dot{z} = -U'(z) - k(z - y) + b\xi^{(z)} \tag{11c}$$

with

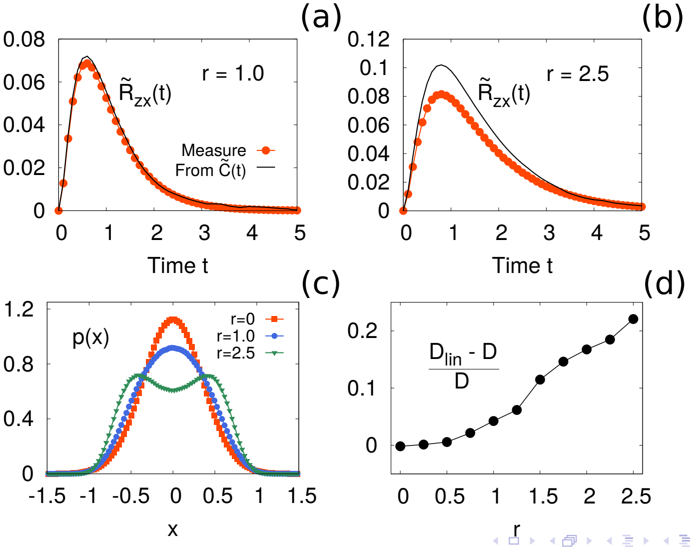
$$U(x) = (1 - r)x^2 + rx^4\,, \tag{12}$$

where k and b are constants, ξ is a delta-correlated Gaussian noise and r is a parameter which determines the degree of nonlinearity of the dynamics: when $r = 0$ the external potential U is harmonic, while for $r > 1$, it takes a double-well shape.

Eq. (2) implies that the general treatment of cases in which the invariant p.d.f. is not Gaussian would require
(i) a careful estimation of the functional form of the joint p.d.f. of *all* variables of the system and
(ii) the knowledge of all correlation functions resulting from the r.h.s. of Eq. (2).
However even in presence of nonlinear contribution to the dynamics, the "linearized" response (4) still gives a meaningful information about the causal relations between the variables of the system.
In particular, the relative error on the indicator $\mathcal{D}_{x \to y}$ defined by Eq. (6) which comes from the responses through Eq. (4), is not large.
We have that the agreement is rather fair even for $r > 1$, when the joint p.d.f. is quite far from a multivariate Gaussian.

Non gaussian effects

The nonlinear terms do not change too much the results



Some general remarks

• **Differences and advantages with other approaches**

1) At variance with the Granger causality and transfer entropy, which are based on the idea of prediction and information, our method has an objective nature, i.e. it is based, as in the Pearl's approach, on the understanding about how the system behaves after a certain operation.

2) It is able to find the "causation link" even in non trivial cases.

3) It allows for a global, as well time dependent, characterization of causation.

4) It is easy to use.

Link of our approach with linear regression

Consider a system described by the time-dependent vector , ruled by some unknown stochastic dynamics. The system is initially in the state $\mathbf{x}_0$, and the dynamics will evolve $\mathbf{x}_t$, where $\mathbf{x}_t$ will, in general, depend both on the initial condition and on the particular realization of the stochastic noise. The best linear approximation to predict $\mathbf{x}_t$ from $\mathbf{x}_0$ will be of the form

$$\mathbf{x}_t \simeq L_t \mathbf{x}_0 + \zeta_t \tag{13}$$

where ζ_t is a vector of random variables. The structure of Eq. (13) is linear, therefore one has the linear regression formula

$$L_t \simeq C_t C_0^{-1} \,. \tag{14}$$

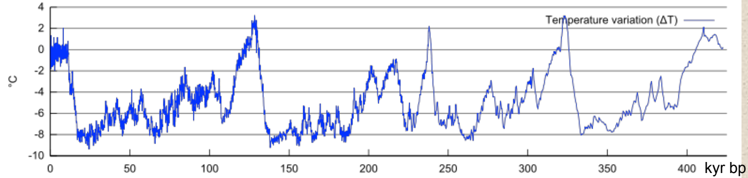
As a consequence, the R_t matrix that one might compute in nonlinear systems by using the "wrong" relation $R_t = C_t C_0^{-1}$ is actually the response associated to the process (13), which is the best linear approximation of the considered transformation $\mathbf{x}_0 \to \mathbf{x}_t$.

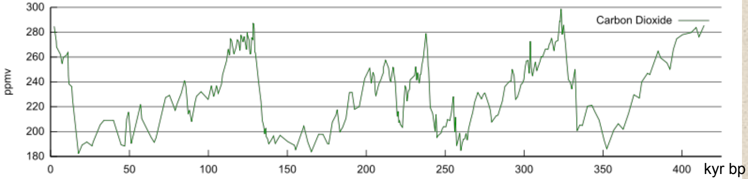
How to face the causation relation in paleoclimate?

Temperature and CO_2 concentration from Vostok ice core data

Sawtooth temperature pattern: increasingly colder glacial period, then rapid return to interglacial

Strong correlation between CO_2 and T





This is a very difficult problem ....

We have to face a very general and difficult topic

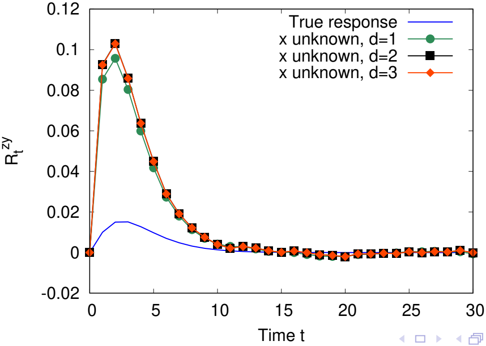
Let us consider the rather common situation in which we do not know the whole state vector $\mathbf{x}$ of the system, but we only have access to the time series of two variables, $\{x_t^{(j)}\}$ and $\{x_t^{(k)}\}$.
In order to show the basic problems in the case the unique information we have comes from the times series we can consider y and z of the toy model (8).
Looking at the vector $\Gamma_t^{(2)} = (y_t, z_t)$ leads to wrong results: for instance, the proxy response computed by applying Eq. (4) to $\Gamma_t^{(2)}$ turns out to be nonzero even in the case $\varepsilon = 0$.
The reason of this failure is that this vector does not completely describe the state of system, and Eq. (2) cannot be derived.

It is tempting, following the approach of Takens, to build a vector of dimension $2d$ by looking at the previous times, e.g.

$$\Gamma_t^{(2d)} = (y_t, y_{t-1}, ..., y_{t-d+1}, z_t, z_{t-1}, ..., z_{t-d+1}),$$

and then to use the protocol based on the FDR, as if $\Gamma_t^{(2d)}$ would describe the whole system we are interested in.

The results unpleasant in the toy model $(a = 0.5, b = 1, \epsilon = 0.02)$



A very bad result

It is possible to show that *at variance with deterministic dynamics, the "embedding" procedure for stochastic processes is not able to reconstruct a finite-dimensional vector whose dynamics is equivalent to that of the original system*.

This implies that to infer causation from time correlations in a stochastic dynamics, we actually need to know all the variables which are relevant to the dynamics of y and z.

At the end we have to face with a difficult problem, which is quite common in all the attempts dealing with data: **the selection of the proper variables**.

Toward an understanding of paleoclimate dynamics

Assume that the paleoclimate is described by the $2D$ vector $\mathbf{x} = (Temperature,\ CO_2)$ whose dynamics has the shape

$$\frac{d\mathbf{x}}{dt} = \mathcal{A}\mathbf{x} + \mathbf{f}(t) + noise$$

where $\mathbf{f}(t)$ indicate the astronomical and tectonic forcing.

We want understand the causation relations $x_1 \to x_2$ and $x_2 \to x_1$ only from the time series of *Temperature* and CO_2), in the general case such a problem is not well posed, see the negative result on the embedding technique.
However there is at least a case where it is possible to use the approach to causation in terms of linear response:
if the characteristic time of the "internal dynamics" (given by the eigenvalues of $\mathcal{A}$) τ_I is much smaller than the typical time $\mathcal{T}$ of the forcing.

The protocol to remove the role of the forcing

Of course, since x_1 and x_2 are both driven by the same forcing there is a strong correlation which is not significative, therefore in order to understand the "internal dynamics", we have to "remove" the effects of the forcing on x_1 and x_2.

The filtering procedure

$$\mathbf{x}_S(t) = (G_\tau * \mathbf{x})(t) = \int G_\tau(t - t')\mathbf{x}(t')\, dt'$$

where the filter $G_\tau(t - t')$ has a maximum at $t = t'$, is very small at large values of $|t - t'|/\tau$ and $\int_{-\infty}^{\infty} G_\tau(t - t')\, dt' = 1$, e.g.

$$G_\tau(t - t') = \frac{1}{\sqrt{2\pi\tau^2}} e^{-\frac{(t-t')^2}{2\tau^2}}$$

In the case of time scale separation i.e.

$$\tau_i \ll \tau \ll \mathcal{T}$$

the protocol allow to detect the "interna" dynamics, namely for the fast variables

$$\mathbf{x}_F(t) = \mathbf{x}(t) - \mathbf{x}_S(t)$$

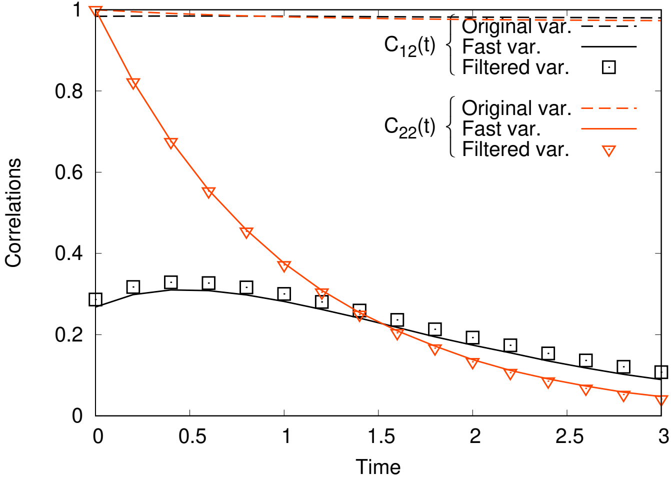
one has

$$\frac{d\mathbf{x}_F}{dt} = \mathcal{A}\mathbf{x}_F + noise$$

therefore one can use the approach in terms of fluctuation/response relation to understand he causation relations $x_1 \to x_2$ and $x_2 \to x_1$ only from the time series.

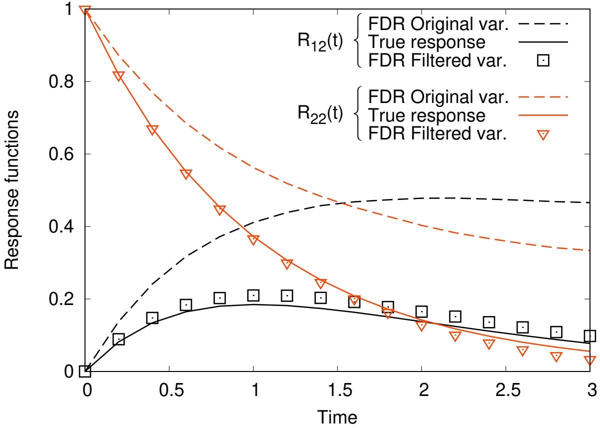
NOTE: the method can work only In the case of time scale separation ($\tau_i \ll \tau \ll \mathcal{T}$).

A numerical test: the correlations



Note the huge effect of the filter procedure!

A numerical test: the responses



Note the huge effect of the filter procedure!

TECHNICAL REMARK: the presence of nonlinear terms does not change too much the results.

VERY NEW results for the data

THE PdF

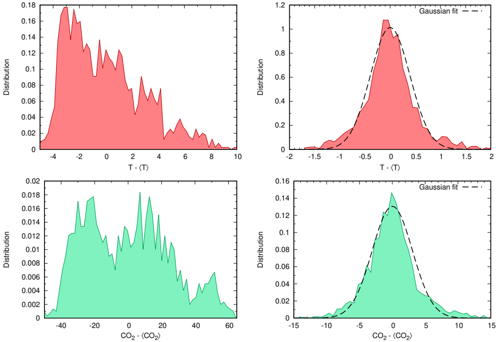


Figura 3: Methods: checking gaussianity *after* filtering

CAUSATION

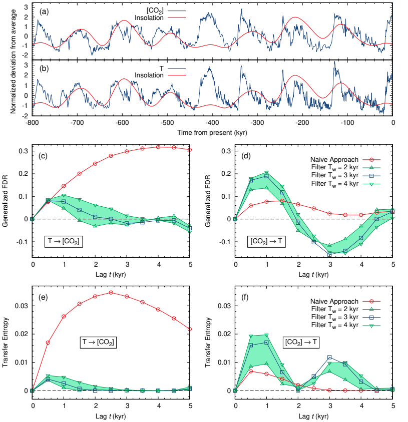


Figure 1. Analysis of the mutual influence between temperature T and CO_2 concentration in paleoclimate data. Panels (a) and (b) show the deviation from average of the two signals as a function of time (blue curves), normalized by the standard deviation. The integrated annual insolation at the 65th north parallel, presenting the typical time-scales of the external driving, is also reported (red curves; see Methods sections for details on the data sources). In panels (c) and (d) the red circles represent the results of a direct application of Eq. (3) on raw data, erroneously suggesting that $T \rightarrow [CO_2]$ is stronger than $[CO_2] \rightarrow T$. The response on data filtered over $T_w = 3$ kyr window (blue squares) instead indicates that the impact of $[CO_2]$ on T is larger. The result is robust with respect of T_w variations by one kyr (green up/down triangles). A similar analysis, where TE are computed instead of generalised FDR, is shown in Panels (e) and (f).

Again an old trouble...

At the end we have to face with a difficult problem, which is quite common in all the attempts dealing with data: **the selection of the proper variables**.

This is an old puzzled question...

How do you know you have taken enough variables, for it to be Markovian?
[L. Onsager, S. Machlup, Phys. Rev. **91**, 1505 (1953)]
The hidden worry of thermodynamics is: we do not know how many coordinates or forces are necessary to completely specify an equilibrium state.
[S.K. Ma *Statistical Mechanics* (World Scientific, 1985), page 29]

Surprisingly many people do not consider seriously such a difficulty.

We can say that the filtering procedure is way to "disentangle" the proper variables from the forcing.

Conclusions and Remarks

• In spite of the (correct) common-wisdom statement *correlation does not imply causation*, a proper employ of time correlations and of fluctuation-response theory allows to understand the causal relations between the variables of a multi-dimensional linear Markov process.
• The fluctuation-response formalism can be used both to find the direct causal links between the variables of a system and to introduce a degree of causation, cumulative in time, whose physical interpretation is straightforward.
• In absence of the knowledge of $\mathbf{x}$, only from the time series $\{x_t^{(i)}\}$ and $\{x_t^{(i)}\}$, in general it is not possible to use of the embedding methodology.
• In presence of time scale separation it is possible to treat time series whose underlying dynamics is generated by

$$\frac{d\mathbf{x}}{dt} = \mathcal{A}\mathbf{x} + \mathbf{f}(t) + noise$$

Some References

* A. Auconi, A. Giansanti, and E. Klipp, Entropy **21**, 177 (2019).
* E. Aurell and G. Del Ferraro, J. Phys. Conf. Ser. **699**, 012002 (2016).
* M. Baldovin, F. Cecconi and A. Vulpiani, Physical Review Reserch **2**, 043436 (2020).
* T. Bossomaier, L. Barnett, M. Harré, and J. T. Lizier, *An Introduction to Transfer Entropy* (Springer, 2016).
* C. W. Granger, Econometrica **37**, 424 (1969).
* D. Hume, *A Treatise of Human Nature*, Vol.1 Edited by D. Norton and M. Norton (Oxford University Press, 2001).
* H. Kantz and T. Schreiber, *Nonlinear Time Series Analysis*, (Cambridge University Press, 1997).
* J. Pearl, *Causality* (Cambridge University Press, 2009).
* J. Runge, J. Heitzig, N. Marwan, and J. Kurths, Phys. Rev. E **86**, 061121 (2012).
* T. Schreiber, Phys. Rev. Lett. **85**, 461 (2000).
* J. Sun, D. Taylor, and E. M. Bollt, SIAM Journal of Applied Dynamical Systems **14**, 73 (2015).