Detecting Causality With Convergent Cross Mapping

August Birk, Clotilde Armand Prætroius, Frederikke Rasmussen & Sune Halkjær

I. Introduction

Correlation and causality are often considered closely related terms, but concluding on causality only based on observations of correlation between variables can lead to incorrect and contradictory results. Correlation is not necessary for causality and neither does it directly imply causality. At the same time, causality is an important property to understand in complex systems and crucial for making political decisions about e.g. climate or epidemiology. Methods are thus needed to detect causality, and [1] presents a previous method and a new alternative method with examples of the latter used on simple models and real data, which will be summarised here.

One commonly used method for determining causality between time-series variables is called ”Granger causality” (GC) and it relies on the predictability of a time-series variable, Y, given a causally related time-series variable, X. X is said to ”Granger cause” Y if lagged values of X can significantly predict values of Y, which also means that the predictability of Y has to decline if X is removed from the system. Using GC requires that the system is separable, meaning that if the system consists of variables X and Y, where X causes Y, information about the effects of X is only contained in time-series for X. This is known to be the case for purely stochastic and linear systems. Consequently, GC is best suited to study such systems. GC is not a useful method when the system is non-separable, the variables are only weakly connected or if the system consists of several variables that are all influenced by a common external variable, which is usually the case for e.g. complex ecosystems. In this case alternative methods are needed, and this is where convergent cross mapping (CCM) comes into the picture.

II. Convergent Cross Mapping

Convergent cross mapping is developed as a method for detecting causality in deterministic dynamical systems in which the dynamics are not purely random but governed by an underlying ”attractor manifold”. This manifold, M, is the state space of the system with the variables e.g. $X(t)$ and $Y(t)$ on the axes and a projection of the time series variables on the manifold surface. When a variable X causes another variable Y, CCM looks for causality between those by studying the extent to which the time series of X can be reconstructed from the time series of Y. Takens' theorem says that lagged time-series coordinates of X and Y generates the so-called shadow manifolds of M, M_X and M_Y with e.g. $X(t)$ and $X(t-\tau)$ on the axes, and each point on the shadow manifolds maps to a point on the attractor manifold. This means that if the variables are causally connected, it is also possible to map directly between the two shadow manifolds (this concept of cross mapping is shown in Figure 1). CCM thus attempts to estimate X from M_Y and Y from M_X. More specifically, the algorithm takes a subset of M_Y with size L, estimates $\hat{X}$ from this subset and computes the correlation ρ between X and $\hat{X}$. Letting L increase from a small number to the total size of M_Y, ρ should increase and converge in the case that X is a cause for Y. In CCM, it is thus possible to estimate X from Y only if X drives Y which runs counter to GC.

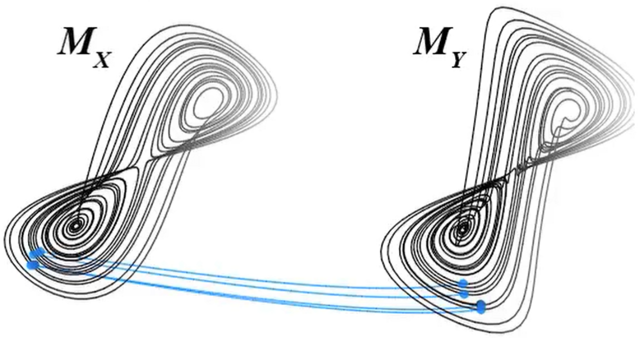

Fig. 1. Cross mapping between shadow manifolds of variables X and Y in the canonical Lorenz system (figure from video $S2$ in supplementary material of [**1**]).

III. Results

A. Simple model

In [1] an example of the use of CCM is presented where the underlying equations are known making it possible to test the performance of the method. The simple non-linear system

$$\begin{aligned} X(t+1) &= X(t)[r_x - r_x X(t) - \beta_{x,y} Y(t)] \\ Y(t+1) &= Y(t)[r_y - r_y Y(t) - \beta_{y,x} X(t)], \end{aligned} \tag{1}$$

is studied for $r_x = 3.8$, $r_y = 3.5$, $\beta_{x,y} = 0.02$ and $\beta_{y,x} = 0.1$ meaning that there is a bidirectional coupling between X and Y. This coupling is detected from convergence of the correlation coefficient, ρ as a function of L. ρ is calculated for the estimate of X from M_Y, $\hat{X}$, and for the estimate of Y from M_X, $\hat{Y}$. The former converges faster because of the larger value of $\beta_{y,x}$ meaning that X drives Y to a larger degree than the reverse, making the estimation of X from M_Y better when using CCM. This can be seen in Figure 2.

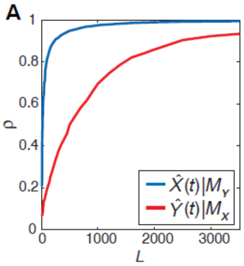

Fig. 2. Convergence of the correlation coefficients, ρ, in the simple system from Eq. 1. The blue graph showing the cross mapping of X using M_Y converges faster because of the stronger effect of X on Y[1].

An important thing to note from this example is that Eq. 1 can be algebraically rearranged describing $X(t+1)$ in terms of $X(t)$ and $X(t-1)$ meaning that we can remove $Y(t)$ as an explicit variable without affecting the prediction of X. In this case, GC would incorrectly deduce that Y does not cause X, whereas CCM, as we saw, correctly concludes the opposite.

The unidirectional case of Eq. 1 is also tested where $\beta_{x,y} = 0$. As expected, this shows that Y can no longer be estimated from X, but $\hat{X}$ remains well estimated. For unidirectional coupling we encounter a special case when X very strongly forces Y. Here, the dynamics of Y become subordinate and we will see a convergence of ρ wrongly suggesting bidirectional coupling. This is phenomenon is called "synchrony" and has to be ruled out before concluding bidirectional coupling with CCM.

B. Complex models

Another use of the method arises when considering two non-interacting species, X and Y, driven by a common external forcing variable Z. Here, cross-correlation (between X and Y) might initially suggest that the two species are coupled, however CCM shows that ρ does not converge, proving no coupling between X and Y. CCM is thereby capable of distinguishing true interaction from simple correlation created by the common driving variable, which is one of the things GC is not able to.

This can be expanded to a more complex system. Species 1, 2 and 3 all interact mutually, and act as external forcing variables with respect to species 4 and 5, which do not interact. 1, 2 and 3 are extension of Z, with 4 and 5 corresponding to the non-coupled pair X and Y. CCM is able to correctly identify the network of all bi- and unidirectional links, as well as the strength of each interaction link.

C. Real-world examples

Ecosystems often differ from systems commonly studied with GC. One of the key differences is that ecosystems are often subject to forcing by external factors such as temperature. Non-interacting species that are exposed to the same driving factor(s) in the same environment may have apparent correlations in their populations. To accommodate spurious correlations of this sort, it is import to address non-separable systems, identify weakly coupled variables - and not least distinguish direct causing interactions of species from effects of their mutual environment.

A specific case of this mechanism is seen in the sardine-anchovy-temperature problem. Historically two main theories for the changes in the population sizes of sardines and anchovies have been proposed: either the species compete for resources or an underlying environmental mechanism drives the changes. Regarding the latter sea surface temperature (SST) is examined in this paper. Both theories serve as explanations for the species reciprocal abundance levels. CCM shows no significant causative relation between the species. Neither do the fishes affect the SST (as one would most likely expect) according to the CCM method. However, SST seems to asymmetrically affect both of the species. See Fig. 3. Thereby CCM strongly suggests the theory of a common environmental driving force being the cause of the fishes apparent correlation in abundance. The coupling between SST and the fishes is weak, implying other driving factors may be at play.

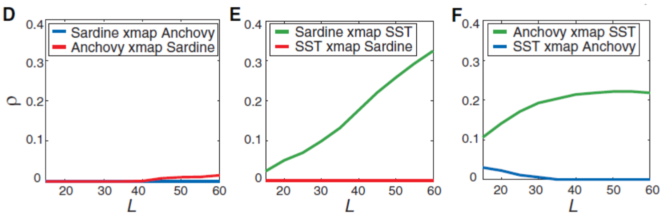

Fig. 3. Sardine-Anchovy-Temperature problem analyzed with CCM. **D**, **E**, **F** respectively: CCM of sardines vs. anchovies, sardines vs. SST, anchovies vs. SST[1]. **D** shows no coupling between sardines and anchovy populations. **E** and **F** suggest a forcing of the SST on both populations.

IV. Conclusion

Understanding the connection but also distinction between correlation and causality is important for understanding complex systems such as climate, epidemics or ecological systems. GC and CCM are two methods that aim to detect causality in complex systems. [1] shows that CCM is a good alternative to GC in cases where GC does not apply, including non-separable systems with weak couplings or systems with a shared driving variable. The method performs well and as expected for constructed models and real data. It is able to predict all uni- and bidirectional couplings expected for the different models studied. The model is furthermore able to distinguish interactions between variables from shared driving variables of these which one of the weaknesses of GC.

References

[1] George et al Sugihara. "Detecting Causality in Complex Ecosystems". In: *Science* 338 (2012). DOI: 10.1126/science.1227079. URL: https://www.science.org/doi/10.1126/science.1227079.