ARTICLE IN PRESS

Future Generation Computer Systems ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

Causal inference from noisy time-series data — Testing the Convergent Cross-Mapping algorithm in the presence of noise and external influence

Dan Mønster [a,b,*], Riccardo Fusaroli [c,b], Kristian Tylén [b,c], Andreas Roepstorff [b], Jacob F. Sherson [d,e]

[a] Department of Economics and Business Economics, Aarhus University, Fuglesangs Allé 4, 8210 Aarhus V, Denmark

[b] Interacting Minds Centre, Aarhus University, Jens Chr. Skous Vej 4, 8000 Aarhus C, Denmark

[c] Center for Semiotics, Aarhus University, Jens Chr. Skous Vej 2, 8000 Aarhus C, Denmark

[d] Department of Physics and Astronomy, Aarhus University, Ny Munkegade 120, 8000 Aarhus C, Denmark

[e] AU Ideas Center for Community Driven Research, Aarhus University, Ny Munkegade 120, 8000 Aarhus C, Denmark

HIGHLIGHTS

- The CCM algorithm is tested on a model system of two coupled logistic maps.
- Noise and an external driving signal are added to test CCM robustness.
- CCM can fail even for low and intermediate coupling.
- We propose low R^2 of fit to exponential convergence as indicator of CCM failure.
- Controlled injection of noise can be used to improve accurate causal inference.

ARTICLE INFO

Article history:
Received 31 August 2016
Received in revised form
20 November 2016
Accepted 6 December 2016
Available online xxxx

Keywords:
Convergent Cross-Mapping
Causality
Logistic map
Noise
Time-series analysis
Nonlinear dynamics
Complex systems

ABSTRACT

Convergent Cross-Mapping (CCM) has shown high potential to perform causal inference in the absence of detailed models. This has implications for the understanding of complex information systems, as well as complex systems more generally. This article assesses the strengths and weaknesses of the CCM algorithm by varying coupling strength and noise levels in a model system consisting of two coupled logistic maps. As expected, it is found that CCM fails to accurately infer coupling strength and even causality direction in strongly coupled synchronized time-series, but surprisingly also in the presence of intermediate coupling. It is further found that the presence of noise reduces the level of cross-mapping fidelity, where the converged value of the CCM correlation decreases roughly linearly as a function of the noise, while the convergence rate of the CCM correlation shows little sensitivity to noise. The article proposes controlled noise injections in intermediate-to-strongly coupled systems could enable more accurate causal inferences. Initial investigation of an external driving signal indicates robustness of CCM toward this potentially confounding influence. Given the inherent noisy nature of real-world systems, the findings enable a more accurate evaluation of CCM applicability and the article advances suggestions on how to overcome the method's weaknesses.

1. Introduction

Complex information systems play a crucial role in infrastructures vital for modern societies. Prime examples are the global Internet, cellular and PSTN networks, power grids, supply networks and the global financial system. The optimal design and operation of complex information systems is an important research topic, and new types of critical infrastructures appear, such

* Corresponding author at: Department of Economics and Business Economics, Aarhus University, Fuglesangs Allé 4, 8210 Aarhus V, Denmark.
E-mail address: danm@econ.au.dk (D. Mønster).

http://dx.doi.org/10.1016/j.future.2016.12.009
0167-739X/

Please cite this article in press as: D. Mønster, et al., Causal inference from noisy time-series data — Testing the Convergent Cross-Mapping algorithm in the presence of noise and external influence, Future Generation Computer Systems (2016), http://dx.doi.org/10.1016/j.future.2016.12.009

as cloud computing and the internet of things. The interaction between many coupled heterogeneous components in such systems makes it difficult to predict the actual behavior at a system-wide level. This is true in particular for predicting events that may be detrimental or even catastrophic to system performance, e.g., network failure [1], rolling blackouts [2], financial crises [3] or even multi-infrastructure failures [4]. Methods that allow us to understand or predict complex systems based on observations are therefore an important alternative approach to trying to predict their behavior simply based on a model or knowledge of their design. Complex information systems usually contain a large number (size) of components that are not identical (heterogeneity); that interact (coupling); and that change over time (dynamics). The combination of size, heterogeneity, coupling, and dynamics means that to characterize complex information systems, time-series of many variables are needed. One of the simplest and most difficult questions one may ask in order to characterize such a data set is: what is the causal relationship between a pair of variables? In other words: does X cause Y, does Y cause X, or does causality go in both directions?

The ability to infer causality from the relation between two variables is an intensely researched area. Common approaches involve having both a model of the system being studied and a series of measurements of that same system [5]. In many cases, however, an adequate model of the system is not available, or there are several conflicting models. Complex natural, social, and technical systems are prime examples, e.g., ecosystems, brains, the climate and the complex information systems mentioned above. Data from such systems are often non-stationary and with nonlinear relationships between variables. This presents a challenge not only to traditional linear statistics based methods, but also to machine learning and big data analytics methods, although for specific applications progress has been made in detecting relevant patterns in the data (e.g., [6]). In most such cases, however, inferring whether one part has a causal influence on another part has to rely on model-free methods. Convergent Cross-Mapping (CCM) [7] is a relatively new method that promises to 'distinguish causality from correlation' in time series data (ibid., p. 496).

While nonlinear methods promise better results they often require more input from the user, e.g. in terms of setting parameters, making unsupervised use more difficult. Further, nonlinear methods often require more computational resources [8]. CCM, however, only requires the specification of two parameters: the embedding dimension and time delay, which can both be estimated with existing algorithms [9–12].

For a system with several variables, for which time series data are available, the CCM method produces a causal network structure describing which variables are causally connected, including the direction of causality. Like mutual information [13], transfer entropy [14], and cross-recurrence quantification [15,16] CCM is a state space method relying on time-delayed embedding of the time series data in a higher dimensional space. While CCM is not a prediction method, using CCM to infer the causal connections between variables can be used to constrain viable predictive models.

CCM has already been used in a wide range of different fields for different kinds of data [17–19], and given the interest and relevance of CCM, it is important to better understand its strengths and limitations. Accordingly, in this article the authors will present results of an in-depth study of a simple model system, the coupled logistic map, with particular emphasis on how the CCM results reflect the model input, varying strength of coupling, levels of noise, and the influence of an external variable. This article will briefly describe how CCM is applied to the model. Then it is examined how coupling strength affects the CCM results, the result of adding noise to the system, and finally the effect of an external driving variable will be investigated. Although noise in real-world data is ubiquitous, the inclusion of noise in model investigations has been largely neglected. It is found that noise can dramatically change the strength of causal inferences. Crucially, it is also observed that the appropriate injection of noise into the dynamics can be used as means of inferring the relative strengths of the coupling (to the extent that the system can be controlled).

2. Related work

CCM was introduced as an alternative to other methods that detect causality between two time series, principally Granger causality [20].

Granger causality [20] was introduced to assess causality between variables in both linear and non-linear systems. The method has gained wide application from contexts of neuroscience to economics, and today several graphical Granger algorithms are available. In a recent study Arnold, Liu and Abe [21] compared a range of the algorithms on both synthetic and real-world data sets, and characterized their performance in terms of ability to recover the correct causal graph, and computational complexity. While Granger causality performs well for certain types of coupled systems, it rests on the assumption of easily separable variables with little or no feedback. As a consequence, it fails to correctly detect the direction of causality in a range of naturally occurring biological, ecological, and social systems that are rather characterized by weak to moderately coupled dynamics [7].

In contrast, CCM is primarily suited for weakly coupled components of non-linear dynamic systems, while it works less well for cases of strong coupling. Accordingly, the two methods have slightly diverging notions of causality, and the statement 'X causes Y' would more accurately be phrased as 'X Granger causes Y' or 'X CCM causes Y'. Sugihara et al. [7] refer to the type of causality captured by CCM as *dynamic causation*, reminiscent of what Lakoff [22] terms systemic causation.

Since its first appearance in 2012, a number of modifications and improvements of the CCM algorithm has been suggested. Cummins, Gedeon, and Spendlove [23] have elaborated the mathematical foundation for CCM to explain in more detail why CCM fails in the limit of strong coupling. They also show that an alternative approach building on the results by Pecora, Carroll, and Heagy [24] can be used, and is more consistent with the mathematical framework developed. Along similar lines, McCracken and Weigel [19] noted that CCM results are not always consistent with theoretical intuitions, and they extended CCM to pairwise asymmetric inference (PAI), which they demonstrated to give results that are in better accordance with the intuitively expected outcomes for several physical systems.

While the original CCM model needs quite extensive time series, Ma, Aihara, and Chen [25] developed cross-map smoothness (CMS) – a method related to CCM – which has the advantage of requiring fewer points in the time series. They also compared CMS to CCM using data from mathematical models, and from a gene regulatory networks, and found that CMS can be used for time-series of length $n \approx O(10)$, whereas CCM requires time-series of length $n \approx O(10^3)$ to yield reliable results.

More recently, Clark et al. [26] developed the CCM technique into multispatial CCM (MCCM) to allow for the analysis of even very short time series ($n < 10$) in the presence of multiple time series, sampled from different spatial locations.

The fact that CCM and related techniques lend themselves to the investigation of also short time-series is important for many applications, however the focus in this article will mainly be on applications with large data sets, so the low n limit, is not a concern here.

Please cite this article in press as: D. Mønster, et al., Causal inference from noisy time-series data — Testing the Convergent Cross-Mapping algorithm in the presence of noise and external influence, Future Generation Computer Systems (2016), http://dx.doi.org/10.1016/j.future.2016.12.009