Information Sciences 448–449 (2018) 134–148

Contents lists available at ScienceDirect

Information Sciences

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

Causal inference for multivariate stochastic process prediction

Simona Cabuz[a,*], Giuseppe Abreu[a,b]

[a] Focus Area Mobility, Jacobs University Bremen, Campus Ring 1, Bremen 28759, Germany
[b] Department of Electrical and Electronic Engineering, Ritsumeikan University, Kusatu, Shiga 525-8577, Japan

ARTICLE INFO

Article history:
Received 28 October 2016
Revised 10 March 2018
Accepted 14 March 2018
Available online 15 March 2018

ABSTRACT

Numerous real world systems of major interest are modeled as sets of analog continuous stochastic processes with delayed and varying causal relationships. Yet studying their dynamic becomes often difficult, as it involves sensing, understanding and predicting a system of inter-dependent random variables in a given context and over time. In the present work we develop systematic, rigorous and efficient framework to structurally characterize and forecast such systems in a flexible manner. In particular we use a graph method based on a maximum spanning tree approach, to capture the causal dependence structure based on directed information theory. To this end we address the sparsity problem in information causality estimation in general, and we propose a new method that identifies and eliminates redundant calculations. To forecast child nodes based on their inferred causal parents we use a linear model aiming to capture the closest approximation of functional relations. We further account for dependencies using causal conditional information by adding links that improve child nodes estimation. The result is a comprehensive and flexible approach to understanding and predicting large sets of inter-dependent narrowband processes, as we demonstrate on both synthetic and real datasets.

1. Introduction

Some of the most important but least understood areas of science today deal with complex probabilistic systems involving intricate patterns of inter-dependencies and correlations in space and in time, often masked under significant uncertainty and noise. Systems of such nature include power grids, neural networks, financial systems, biological organisms, communication networks, weather, astronomical signals, social and political trends, and several other, currently on the frontier of many scientific fields.[1]

The recent trend of "big data" high-throughput in all these fields leads to an increasing interest in extracting meaning out of ever larger amounts of data. This meaning comes in two general forms corresponding to their dependency structure, and underlying functional behavior among components. Significant efforts have already been made to bring together a complex tool in order to both structurally and predictively characterize complex probabilistic systems. Clearly, several methods have been designed to estimate information measures in order to capture causal relations in discrete cases [9]. We will not examine nor attempt improving on those particular cases. Instead, we take this problem one step further and address a

* Corresponding author.
E-mail addresses: s.poilinca@jacobs-university.de (S. Cabuz), g.abreu@jacobs-university.de, g-abreu@fc.ritsumei.ac.jp (G. Abreu).

[1] By flexible we mean that the method has a relatively low complexity, and a tendency build-in by construction, to extract an accurate prediction model of the process which has a smaller number of nodes and causal relationships than the true structure of the process itself. This give the method the ability to handle processes with both small and large numbers of nodes and inter-connections.

https://doi.org/10.1016/j.ins.2018.03.038
0020-0255/

broader class of signals: continuous, narrowband signals defined on infinite alphabets with a delayed memory. These are particularly interesting in several applications as discussed below.

Before going into further details on our contribution, let us briefly review the literature. If in the past researchers have used correlative models, a new information theoretic approach is being developed and increasingly used in several fields including neuroscience [22,25,34,35], climatology [45], content-based image retrieval [43], pattern classification [42] and stock market [33]. With a root in communication theory for channels with feedback dating back to the 80's [27,29], directed information theory was recently linked to Granger concept of causality [5]. This led to the introduction of directed information graphs, Granger causality graphs and directed networks, which are relatively recent additions to the data scientist's toolbox.

Marko [27] defined directive information in an average sense, in terms of the limit of the relative (difference of) conditional entropy of a variable conditioned on a sequence of itself and on a sequence of itself and another. A while later, Massey [29] introduced the notion of *directed information*, which measures the ability to describe the *current* state of a random variable based on the *past and current* states of *another* related variable. Unlike Marko's, Massey's definition is time-varying. Kramer [19] formalized the notion of directed information, rigorously studying the relationship between the latter and various other classic measures of "relative information" often encountered in information theory (entropy, conditional entropy, mutual information, etc). In [1,2], Abdallah et al. also discusses the relationship of various forms of "relative information" in the context of predictability. He goes on to define "predictive information" and "binding" information, both of which attempt to measure how the observation of a variable up to a point can be used to predict the behaviour of another, beyond that point. Since the notion of mutual information is well disseminated and understood, it is useful to explain directed information by drawing a parallel with mutual information. An excellent description in those lines was offered in [34], i.e: "as one can determine the degree of *correlation* by computing *mutual information*, one can determine the degree of *causation* by computing the *directed information*".

In [34,35] the authors use directed information as a non-parametric measure of "causality" among a network of processes, and develop a parametric estimator to estimate directed information from data. More recently, the same authors in [35] combine the idea with the optimality of maximum spanning trees, to build directed information-based "causal graphs" of multiple variables. However, as put by Liu et al. [25], "the main advantage of using directed information to measure the causal relationship between variables is that it is non-parametric, as it does not assume any underlying model". Nonetheless, these approaches generally lack a rigorous strategy to derive the underlying dependence structures, or are designed for a narrow class of signals.

Following these recent ideas, our work provides a systematic and computationally efficient way to extract fully descriptive structural and predictive meaning. To this end we address the sparsity problem and propose a new method to adaptively estimate causal measures for a broader class of processes. Estimators are initially designed for mutual information between two random variables but provide an extension to N random variables. Although they are robust and have excellent performance when applied on two variables only, when the dimension grows, the data becomes sparse and the accuracy is drastically affected. Ways to overcome this have already been proposed. For instance, a way to overcome the sparsity problem is to assume the processes are first order Markovian stationary processes. While this applies very well to numerous spike recordings and is suitable to model neural activity as used in [17,35], the problem of narrowband processes with possibly non-stationary relationships or lagged effects on each other remains open.

Another way to overcome the sparsity problem is based on the ideas laid out in [13]. The authors in [36,37] explore the structural properties of the time series graphs to reduce dimensionality when estimating transfer entropy. The idea is to build a graph in which nodes are subprocesses of each observed variable and separate the nodes in terms of conditional independencies. In this way one can quantify the couplings between the nodes and use only a subset of components. The procedure used to obtain an estimate of the graphical model is based on the PC algorithm [40]. This way dimension is potentially greatly reduced and the components with a direct influence on the current variable are chosen in a judicious way. However conditioning terms, together with current components may still sum up to a considerable number of variables to be estimated, which we want to avoid.

The idea to restrict directed information estimation to a subset of variables was proposed also in [21–25]. While the solution used in [25] is to compute directed information on two previous successive points only, in [23,24] the same authors improved their estimation by using a time lagged directed information.

Using a time window of length L, when the length of time series N is significantly larger than L they use an average time lagged approximation for directed information. However, even this solution considers its last L *coupled* components, which for $L > 2$ is already computationally heavy.

To overcome all these constraints we propose identifying and using the most causally relevant component for each pairwise interaction, regardless of how far it is in the past. The detailed analysis, together with the exact equations are provided later on in Section 3. By having an accurate estimate of causal relationships enables us to use a maximum spanning tree method to *efficiently* encapsulate the structural information, together with a linear approximation to capture the functional dependence while keeping a low complexity profile. We extend the maximum spanning tree structure to a graph, using prediction of the nodes. Namely, the structure will add new connections while the prediction of nodes can be improved and will converge to perform the best prediction using only linear modeling.

We must emphasize that beyond its attractive computational simplicity, our adaptive delay estimation approach is a very natural way to model real-world systems where delays are common due to *propagation* effects in all kinds of networks: