Learning causal dependencies in large-variate time series

Gianluca Bontempi
Machine Learning Group, Department of Computer Science
ULB, Université Libre de Bruxelles, Belgium

***Abstract*—A major challenge in causal inference from observational data is to discriminate between associative dependencies and effective causal relationships. This is particularly challenging in large-variate and temporal settings (e.g. in spatio-temporal time series) where the multivariate nature of interactions induces a significant correlation between most of the variables. In recent years, a number of data-driven approaches have been proposed to learn the mapping between some features of the data distribution and the probability of a causal connection between a pair of variables. Most state-of-the-art approaches, however, deal with bivariate cases neglecting the role of the context determined by the other variables. This is a strong limitation in large-variate and temporal settings which are the object of this study. In order to address the context issue, this paper introduces a new set of descriptors based on interaction information to *featurize the context* and justifies its introduction by using a graphical modeling formalism. The resulting causal inference method is assessed on a number of large-variate synthetic stationary time series. The assessment shows that the proposed method outperforms several state-of-the-art causal inference techniques.**

I. Introduction

"We are drowning in data and starving for knowledge" is an old adage of data scientists that nowadays should be rephrased into "we are drowning in associations and starving for causality". The omnipresence of big data and the success of machine learning expose our society to a number of (real or presumed) associations that could have impact on lifestyle, health choices, economic and political decisions. Most recent AI success stories boil down to companies (or scientists) discovering some interesting associations in data and using them for some smart decision making (e.g. control, risk management, or customer interaction). The democratization of machine learning software and platforms grows then the risk of ascribing causal meaning to simple and sometimes brittle associations. The risk of spurious attributions increases in settings characterized by high dimension, multivariate interactions, dynamic behavior where direct manipulation is not only unethical but also impractical. Think, for instance, to genetics, social sciences, economics or climate modeling where the large dimension of the interactions and the spatio-temporal nature of the relevant features [1] makes impossible the direct manipulation of more than few variables (e.g. knocking out few genes). Nevertheless the high societal impact of those domains requires accurate causal inference methodologies to enable transparency, interpretability and effective decision making.

The most successful notion of causality is related to the notion of intervention (e.g. by controlled experimentation) and has been largely discussed in the seminal works of Pearl [2]. According to this interpretation, the best formalism to represent the causal structure underlying observational data is a Directed Acyclic Graph where nodes denote the random variables and edges the causal relationships. The conventional ways to recover a causal structure from observational data are score-based algorithms, which search for a DAG optimising a certain score (e.g. likelihood) and constraint-based algorithms (e.g. PC [3]) [4] which seek a DAG that is compatible with the conditional independencies seen in the dataset. Score-based search methods rely on NP-hard optimization problems and scale badly to high-dimensional data. Constraint-based algorithms suffer as well of many limitations like the presence of indistinguishable configurations (e.g. a closed triplet or a pair of variables) due to equivalence classes, the high complexity (polynomial in the number of variables but exponential on the size of the neighborhood tested for conditional independence) in large variate settings and statistical inconsistency due to multiple hypothesis testing [4]. This opened the way to alternative learning algorithms which pose the problem of causal inference as the classification of probability distributions [5], [6]. The rationale of those algorithms is that the existence of a causal relationship induces a constraint on the observational multivariate distribution. In other words, causality leaves footprints in the data distribution that can be hopefully used to reduce the uncertainty about the causal structure [5]. Typically those approaches *featurize* the data distribution and use those features to train a classifier able to predict the causal patterns between a pair $\mathbf{z}_i, \mathbf{z}_j$ of interest (e.g. $\mathbf{z}_i \rightarrow \mathbf{z}_j$, $\mathbf{z}_i \leftarrow \mathbf{z}_j$ or the absence of a link).

Most algorithms in literature focus on bivariate distributions. Examples are ANM (Additive Noise Model) [7], IGCI (Information Geometry Causality Inference) [8], [9], LiNGAM (Linear Non Gaussian Acyclic Model) [10] and the algorithms described in [11] and [12]. The interest of those strategies became evident thanks to the ChaLearn cause-effect pair challenge [13] organized by I. Guyon, whose results made clear that indistinguishable settings (e.g. in terms of conditional probability) can be in practice set apart once large enough datasets are made available.

An extension of this principle to large variate case has been proposed by the D2C (Dependency to Causality) approach [6]. This approach stands out from the mainstream litterature since it

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

does not use kernel techniques (e.g. kernel mean embedding [5]) to featurize observed data but it relies on asymmetric features (also called descriptors) based on information theory to extract meaningful hints about the causal structure. The D2C algorithm performs three steps to predict the existence of a directed causal link between two variables in a multivariate setting: (i) it estimates the Markov Blankets of the two variables of interest and ranks its components in terms of their causal nature, (ii) it computes a number of asymmetric descriptors and (iii) it learns a classifier (e.g. a Random Forest) returning the probability of a causal link given the descriptors value.

Note that the first step is necessary because of the context issue in multivariate causal inference: the relevance of a statistical descriptor in predicting a dependency between two variables $\mathbf{z}_i$ and $\mathbf{z}_j$ necessarily depends on the *context*, i.e. the set of remaining variables (see Section 4 of [5]). This issue is made explicit by the notion of strong relevance in variable selection [14] which is a necessary condition for causal relevance [15]: given a set $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_n]$ of n random variables, a variable $\mathbf{z}_i$ is strongly relevant for $\mathbf{z}_j$ if the information of $\mathbf{z}_i$ about $\mathbf{z}_j$ given all the remaining ones is larger than zero. Since the relevance of a variable is dependent on the context (e.g. in the XOR case), in a multivariate setting its causal role cannot be fully accounted for by remaining in a bivariate perspective. In [6] the context issue is addressed by first inferring from data the causal components in the Markov blankets of $\mathbf{z}_i$ and $\mathbf{z}_j$. However this bootstrapping procedure induces a vicious circle, since to learn the causal relation between $\mathbf{z}_i$ and $\mathbf{z}_j$, we should already be able to infer the causes of $\mathbf{z}_i$ and $\mathbf{z}_j$.

This paper addresses this dilemma by proposing a way to *featurize the context*, i.e. by including in the set of descriptors a new one, based on the notion of information interaction [16], which is able to account for the context information without relying on an explicit computation of the neighbourhood. This paper extends the D2C algorithm into three main directions: (i) it introduces a new context-aware asymmetric descriptor based on interaction information (ii) it justifies the relevance of such descriptor in a probabilistic way by means of a graphical modelling formalism (Section III), and (iii) it assesses the accuracy of the modified algorithm in a number of large scale temporal tasks by benchmarking it against state-of-the-art approaches (Section V).

In time series analysis, causal inference is typically based on the concept of Granger causality which relies on the concept of temporal precedence and the assumption that a cause contains information not available elsewhere. Notwithstanding its historical role, this measure is an associative measure and not a causal one, since it is more adequate to detect relevant variables than causal ones (see also chapter 10 of the book [17]). In fact, strong relevance is necessary but not sufficient for causal relevance [15]. For this reason, especially in large variate temporal settings, Granger tests are of limited utility in reconstructing the causal dependencies.

The experimental session shows instead that a context-aware learning approach may be accurate in inferring causal relationships in large-variate (up to one hundred dimensions) stationary time series whose generative process is given by a graphical model (e.g. an auto-regressive process). To the best of our knowledge, this is the first work addressing causal inference in such a large scale temporal setting. The most related work is [18] which adopted as well an approach based on supervised learning inspired to D2C but whose multivariate experimental setting is limited to three dimensional time series.

II. Causality and Information Theory

Let us consider a dataset D sampled from a multivariate distribution $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_n]$ and suppose that we want to infer from D the causal relationships underlying the distribution. Let us suppose that the set of causal relationships existing between the variables[1] of interest can be described by a Markov and faithful Directed Acyclic Graphs (DAG) [4], [2]. In this case a DAG is an accurate representation of (in)dependencies between the components $\mathbf{z}_i$ and by *d-separation* we may read from the graph if two sets of nodes are (in)dependent conditioned on a third[2].

A structural notion which can be described in terms of conditional mutual information is the notion of Markov Blanket (MB). The Markov Blanket of the variable $\mathbf{z}_i$ is the smallest subset of variables belonging to $\mathbf{Z} \setminus \mathbf{z}_i$ (where $\setminus$ denotes the set difference operator) which makes $\mathbf{z}_i$ conditionally independent of all the remaining ones, i.e. $I(\mathbf{z}_i; (\mathbf{Z} \setminus (\mathbf{M}_i \cup \mathbf{z}_i))|\mathbf{M}_i) = 0$

Let us suppose that we are interested in predicting the existence of a directed causal link $\mathbf{z}_i \rightarrow \mathbf{z}_j$. In [6] a *dependency descriptor* of the ordered pair $\langle i, j \rangle$ is a function $d(i, j)$ of the distribution of $\mathbf{Z}$ which depends on i and j. Example of dependency descriptors are the correlation $\rho(i, j)$ between $\mathbf{z}_i$ and $\mathbf{z}_j$, the mutual information $I(\mathbf{z}_i; \mathbf{z}_j)$ or the partial correlation between $\mathbf{z}_i$ and $\mathbf{z}_j$ given another variable $\mathbf{z}_k, i \neq j, j \neq k, i \neq k$. A dependency descriptor is *symmetric* if $d(i, j) = d(j, i)$ (e.g. correlation and mutual information) otherwise it is *asymmetric*. The rationale of the D2C approach is that, because of the asymmetric property of causality, asymmetric descriptors $d(i, j)$ are informative about the causal relationship between $\mathbf{z}_i$ and $\mathbf{z}_j$. Useful asymmetric descriptors can be derived once we known the Markov Blankets $\mathbf{M}_i$ and $\mathbf{M}_j$.

Let $\mathbf{m}_i^{(k)}$ and $\mathbf{m}_j^{(k)}$ denote the generic components of the Markov Blankets $\mathbf{M}_i$ and $\mathbf{M}_j$, respectively, with no distinction between cause, effect or spouse. Let us consider for instance the portion of a DAG represented in Figure 1 (excerpt from [6]) where the variable $\mathbf{z}_i$ is a direct cause of $\mathbf{z}_j$. The figure shows also the Markov Blankets $\mathbf{M}_i$, $\mathbf{M}_j$ and their components, i.e. the direct causes (denoted by $\mathbf{c}$), the direct effects ($\mathbf{e}$) and the spouses ($\mathbf{s}$) [20].

[1] For the sake of clarity, we will use the term "variable" to refer to a component of $\mathbf{Z}$ and the term "feature" or "descriptor" to denote the statistic used to featurize the observed data distribution.

[2] Given three continuous random variables $\mathbf{z}_1$, $\mathbf{z}_2$ and $\mathbf{z}_3$ having a joint Lebesgue density, the *conditional mutual information* [19] $I(\mathbf{z}_1; \mathbf{z}_2|\mathbf{z}_3)$ between $\mathbf{z}_1$ and $\mathbf{z}_2$ once $\mathbf{z}_3$ is null if and only if $\mathbf{z}_1$ and $\mathbf{z}_2$ are conditionally independent given $\mathbf{z}_3$

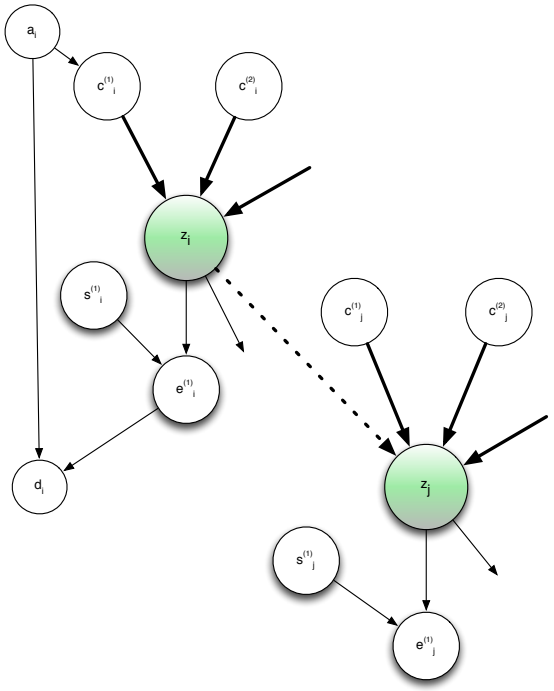

Fig. 1. Two causally connected variables and their Markov Blankets (excerpt from [6]).

Under the following assumptions [6]: (i) the only path between the sets $\mathbf{z}_i \cup \mathbf{M}_i$ and $\mathbf{z}_j \cup \mathbf{M}_j$ is the edge $\mathbf{z}_i \rightarrow \mathbf{z}_j$ and (ii) there is no common ancestor of $\mathbf{z}_i$ ($\mathbf{z}_j$) and its spouses $\mathbf{s}_i$ ($\mathbf{s}_j$), a number of asymmetric conditional (in)dependence relations follow (see Table 1 in [6]) like

$$\mathbf{z}_i \rightarrow \mathbf{z}_j \Rightarrow \mathbf{z}_i \not\perp\!\!\!\perp \mathbf{c}_j^{(k)} | \mathbf{z}_j \text{ and } \mathbf{z}_j \perp\!\!\!\perp \mathbf{c}_i^{(k)} | \mathbf{z}_i \quad \forall k$$

In plain words, by conditioning on the effect $\mathbf{z}_j$ there is a dependence between $\mathbf{z}_i$ and the direct causes of $\mathbf{z}_j$ while by conditioning on the $\mathbf{z}_i$ there is a d-separation between $\mathbf{z}_j$ and the direct causes of $\mathbf{z}_i$. Such reasoning leads to the definition of a number of asymmetric descriptors like

$$d_1^{(k)}(i,j) = I(\mathbf{z}_i; \mathbf{c}_j^{(k)} | \mathbf{z}_j), \quad d_2^{(k)}(i,j) = I(\mathbf{e}_i^{(k)}; \mathbf{c}_j^{(k)} | \mathbf{z}_j),$$
$$d_3^{(k)}(i,j) = I(\mathbf{c}_i^{(k)}; \mathbf{c}_j^{(k)} | \mathbf{z}_j),$$
$$d_4^{(k)}(i,j) = I(\mathbf{z}_j; \mathbf{c}_i^{(k)}) \quad (1)$$

which are all bigger than zero while their asymmetric counterparts are all null.

Though the asymmetric nature of descriptors (1) is encouraging, the computation of those quantities takes for granted that it is possible (i) to infer the Markov Blankets $\mathbf{M}_i$ and $\mathbf{M}_j$ and (ii) to label each of component $\mathbf{m}_i^{(k)} \in \mathbf{M}_i$ and $\mathbf{m}_j^{(k)} \in \mathbf{M}_j$ as causes, effects or spouses. Now if this was the case, we would have already solved the causal inference problem. In practice instead of (1) the algorithm may only compute

$$d_1^{(k)}(i,j) = I(\mathbf{z}_i; \mathbf{m}_j^{(k)} | \mathbf{z}_j), \quad d_2^{(k)}(i,j) = I(\mathbf{m}_i^{(k)}; \mathbf{m}_j^{(k)} | \mathbf{z}_j),$$
$$d_3^{(k)}(i,j) = I(\mathbf{m}_i^{(k)}; \mathbf{m}_j^{(k)} | \mathbf{z}_j),$$
$$d_4^{(k)}(i,j) = I(\mathbf{z}_j; \mathbf{m}_i^{(k)}) \quad (2)$$

where $\mathbf{m}_i^{(k)} \in \mathbf{M}_i$ and $\mathbf{m}_j^{(k)} \in \mathbf{M}_j$. We may define this issue as the *context-dependency* issue according to the discussion in Section 4 of [5].

To address this aspect, in [6] a preliminary causal variable ranking, based on a causal filter [21], [22] is proposed to disambiguate the role of $\mathbf{m}_i^{(k)}$ and $\mathbf{m}_i^{(k)}$ in (2). However this step is time consuming and prone to propagate errors in the subsequent computation of the descriptors (2). In the following section we discuss how it is possible to avoid this bootstrapping step by properly featurizing the context.

III. Context-aware featurization based on information theory

Structure learning approaches to causal inference aim to find necessary and sufficient conditions for associating a DAG structure to an observed dataset. This is often unfeasible because of indeterminate configurations or sequential procedures (e.g. constraint-based algorithms) which reduce the significance of the final output. The rationale of this paper is to address the problem of causal inference in a probabilistic setting and define a set of variables which bring probabilistic information about the existence of a causal link $\mathbf{z}_i \rightarrow \mathbf{z}_j$. The considerations made in the previous section show that if $\mathbf{z}_i \rightarrow \mathbf{z}_j$ holds, then with a certain probability (i.e. the probability that our assumption is true) a number of relationships (1) hold as well (e.g. $I(\mathbf{c}_i^{(k)}; \mathbf{c}_j^{(k)} | \mathbf{z}_j) > 0$). Unfortunately those quantities cannot be computed directly from data. What we can estimate are indeed quantities like $I(\mathbf{m}_i^{(k)}; \mathbf{m}_j^{(k)} | \mathbf{z}_j)$ for which we have no certainty about the role (i.e. cause or effect) of $\mathbf{m}_i^{(k)}$ and $\mathbf{m}_j^{(k)}$ (e.g. is $\mathbf{m}_i^{(k)}$ a cause $\mathbf{c}_i^{(k)}$ or an effect $\mathbf{e}_i^{(k)}$ of $\mathbf{z}_i$?). In other terms the role of the elements used to compute the descriptors are latent since they are not directly observable. However, though their specific role is not observable, we can observe some related statistics, like the information interaction [16], which may provide some insight about the nature of the elements belonging to the Markov blankets. Given three random variables $\mathbf{m}_i^{(1)}, \mathbf{m}_i^{(2)} \in \mathbf{M}_i$ and $\mathbf{z}_i$ the *interaction information* is

$$I(\mathbf{m}_i^{(1)}; \mathbf{m}_i^{(2)}; \mathbf{z}_i) = I(\mathbf{m}_i^{(1)}; \mathbf{m}_i^{(2)}) - I(\mathbf{m}_i^{(1)}; \mathbf{m}_i^{(2)} | \mathbf{z}_i). \quad (3)$$

This quantity sheds a light on the possible causal patterns (e.g. v-structures) existing between them. For instance in [21] it is shown that negative interaction between $\mathbf{m}_i^{(1)}$, $\mathbf{m}_i^{(2)}$ and $\mathbf{z}_i$ are typically associated to the *common effect* configuration (also known as the *explaining-away* effect where $\mathbf{m}_i^{(1)}$, $\mathbf{m}_i^{(2)}$ are both causes of $\mathbf{z}_i$) while positive interaction is associated to the common cause configuration (i.e. $\mathbf{m}_i^{(1)}$, $\mathbf{m}_i^{(2)}$ are both effects of $\mathbf{z}_i$). While in [21] interaction information is used to design

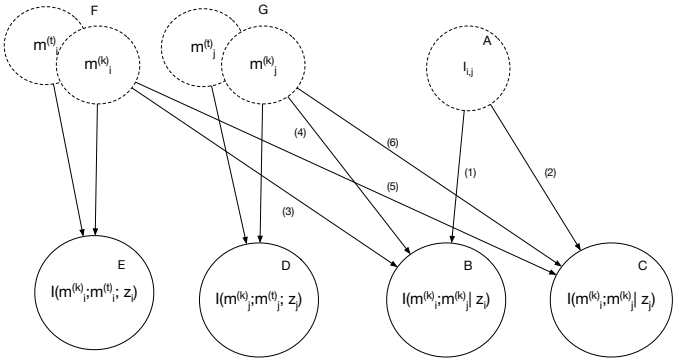

Fig. 2. Graphical modelling of the dependencies between context-aware descriptors based on interactions (nodes D and E), original D2C descriptors (nodes B and C) and the causal relationship $\mathbf{z}_i \rightarrow \mathbf{z}_j$ (node A).

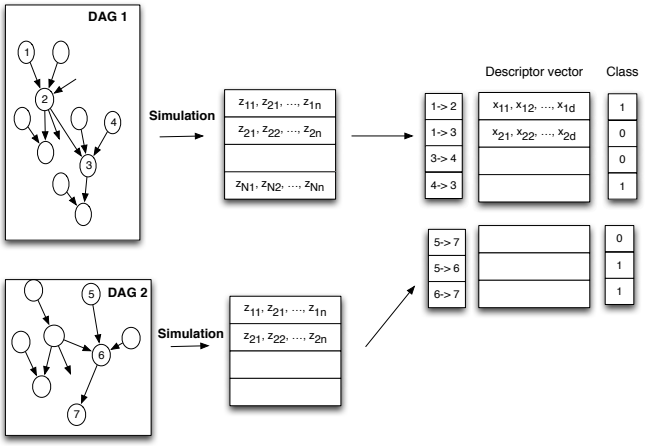

Fig. 3. Training data generation in caD2C.

a causal ranking filter, here we use it as a context-aware and observable proxy of the Markov Blankets $\mathbf{M}_i$ and $\mathbf{M}_j$.

Now, it is possible to show that context-aware descriptors based on interaction (3) and the original D2C descriptors (2) are jointly informative about the causal dependency $\mathbf{z}_i \rightarrow \mathbf{z}_j$, i.e. they reduce the uncertainty about the existence of this link. In order to illustrate this concept, we summarize all the relevant elements in the graphical model of Figure 2 where latent variables are denoted by dotted line nodes and observable statistics by the continuous line nodes. The latent node A represents a binary variable I_{ij} associated to the causal relation $\mathbf{z}_i \rightarrow \mathbf{z}_j$ taking value 1 in the case of an existing link, 0 otherwise. This node is not directly observable but from Section II we know that it conditions (via edges (1) and (2)) the probability distribution of D2C descriptors (nodes B and C). The distribution of those descriptors however are not only dependent on the actual causal relationship but also on the nature of terms $\mathbf{m}_i^{(k)} \in \mathbf{M}_i$ and $\mathbf{m}_j^{(k)} \in \mathbf{M}_j$ (edges (3)-(6)). Since the causal nature of the components of $\mathbf{M}_i$ and $\mathbf{M}_j$ is latent we can denote it by random variables $\mathbf{m}_i$ and $\mathbf{m}_j$ taking values in the ternary set $\{$cause, effect, spouse$\}$. As discussed above, the interaction between members of $\mathbf{M}_i$ and $\mathbf{M}_j$ is informative about the causal nature of those variables. In other terms the latent nature of $\mathbf{m}_i$ and $\mathbf{m}_j$ conditions the distribution of the observable interaction terms (nodes D and E).

From the graphical model it is easy to realize that, given the latent nature of the nodes F and G, no d-separation (i.e. no independence) occurs between the nodes B,C,D,E and the node A. This means that the features associated to the nodes B,C,D,E are informative about the state of the node A. In other words, by measuring the quantities represented by nodes B,C,D,E we may reduce the uncertainty about the binary state of the node A.

IV. The context-aware algorithm

Let us consider a pair of measured variables $\mathbf{z}_i$ and $\mathbf{z}_j$ and suppose we want to estimate the probability of a direct causal link between them. The original D2C algorithm learns from synthetic data the mapping between a number of descriptors (1) and the label of the causal relationships. The training data generation process is sketched in Figure 3. First, several synthetic datasets (in this paper we deal with multivariate time series) are generated on the basis of the associated DAG representation. Then, for each dataset and for (a subset of) all pairs of variables , descriptors are computed. Finally, for each considered pair (e.g nodes 1 and 2 in DAG 1 of Figure 3), a data sample is created from the set of descriptors made of the input vector $x_1, \ldots x_d$ and the related target label (e.g. 1 to denote the existence of the link) of the associated causal edge $(1 \rightarrow 2)$.

Here we briefly summarise the proposed algorithm (caD2C) by insisting on the the changes of the context aware version with respect to the original (D2C). We refer the reader to the D2C paper [6] for more details. Suppose we want to predict the existence of a causal link $\mathbf{z}_i \rightarrow \mathbf{z}_j$. The caD2C steps are

1) it ranks the most relevant variables (e.g. in terms of correlation or mutual information) for $\mathbf{z}_i$ and $\mathbf{z}_j$ and stores them into the sets $\mathbf{M}_i$ and $\mathbf{M}_j$.
2) for each pairs $(\mathbf{m}_i^{(k)}, \mathbf{m}_j^{(k)})$, where $\mathbf{m}_i^{(k)} \in \mathbf{M}_i$ and $\mathbf{m}_j^{(k)} \in \mathbf{M}_j$, it computes the set of (conditional) mutual information descriptors (2) to featurize the dependency between $\mathbf{z}_i$ and $\mathbf{z}_j$
3) for each pairs $(\mathbf{m}_i^{(k)}, \mathbf{m}_i^{(t)})$ and $(\mathbf{m}_j^{(k)}, \mathbf{m}_j^{(t)})$, where $\mathbf{m}_i^{(k)}, \mathbf{m}_i^{(t)} \in \mathbf{M}_i$ and $\mathbf{m}_j^{(k)}, \mathbf{m}_j^{(t)} \in \mathbf{M}_j$, it computes the context aware interaction information descriptor (3)
4) it computes a set of quantiles of the empirical distributions of the terms computed in the two steps before and use them as input vector of the classifier.

Note that in the step 1, since an interaction information descriptor is added, there is no more need to bootstrap the algorithm by identifying a set of putative causes of $\mathbf{z}_i$ and $\mathbf{z}_j$. It is is sufficient to compute a superset $\mathbf{M}_i$ ($\mathbf{M}_j$) of the Markov Blanket of $\mathbf{z}_i$ ($\mathbf{z}_i$) by using a simple ranking algorithm, e.g. by ranking variables on the basis of their correlation to the

target. This step has now a linear complexity in the number n of variables and contains with high probability some causes and effects.

Once the dataset is built, a conventional binary classifier (e.g. Random forest) is used to train the classifier. Note that a problem of unbalancedness may occur (given the edge distribution in DAGs) since the positive class (associated to the existence of a causal relationship) is a minority class. For this reason we have recourse to Easy Ensemble [23] strategies to address the unbalancedness issue. The algorithm is implemented in R in the package D2C[3]. As far as complexity is concerned, given that the proposed algorithm is a modification of D2C, the considerations in Section 3.1 of [6] are still valid, with the exception of the term related to the computation of the Markov blankets of the two nodes whose complexity decreases from $O(n^2)$ to $O(n)$. The complexity for testing the existence of a causal link between a pair of variables becomes then $O(Cn + K^2N)$ where N is the number of samples of the observed dataset and K is the size of $\mathbf{M}_i$ and $\mathbf{M}_j$ [4].

V. Experiments

We carried out a number of causal inference experiments on a large number of simulated stationary time series characterized by nonlinearity, large dimension and cross-sectional dependencies. The set of 16 synthetic generating processes is detailed in Table I. Two additional linear processes were considered as well. All the processes are in the multivariate Nonlinear Autoregressive (NAR) format

$$\begin{cases} Y_{t+1}[1] &= f_1(Y_t[1],\dots,Y_t[n],Y_{t-1}[1],\dots,Y_{t-1}[n],\dots, \\ &\quad Y_{t-l}[1],\dots,Y_{t-l}[n]) + W[1]_{t+1} \\ \dots \\ Y_{t+1}[n] &= f_n(Y_t[1],\dots,Y_t[n],Y_{t-1}[1],\dots,Y_{t-1}[n],\dots, \\ &\quad Y_{t-l}[1],\dots,Y_{t-l}[n]) + W[n]_{t+1} \end{cases} \tag{4}$$

where $l \geq 0$ is the maximum considered lag, $Y_t = (Y_t[1],\dots,Y_t[n])$ is a n dimensional stationary time series and the covariance of the error vector W is diagonal. Note that not all arguments of the functions $f_i(\cdot)$ are necessarily present and that each of those vector autoregressive processes may be represented by a graph visualizing the conditional distribution of each component of Y_t given the past values $Y_{t-1},\dots,Y_{t-l}$ [1]. From each generating process we obtain several stationary series by changing the seed, the number of series (from $n = 10$ to $n = 50$), the number of observations (from $N = 150$ to $N = 500$), the neighborhood set $\mathcal{N}$ (of random size in the interval $[1, 3]$), the error variance (in the interval $[0.1, 0.3]$) and by selecting a subset of the lagged variables in the right-hand terms of (4). To avoid overfitting, the caD2C algorithm is trained on 1000 time series generated from a subset of 10 processes from Table I and tested on 200 time series generated from the remaining ones. Since the multivariate size of each series goes up to 50 and the maximum lag is $l = 5$, the size of underlying DAGs may go up to 250 nodes.

We benchmarked the caD2C strategy against the original D2C and several state of the art algorithms for causal inference. Note that no algorithm received any information about time, i.e. we did not take advantage of time priority to reduce the search space: this was done on purpose to make the task harder and increase the risk of confounding effects with causes. For the sake of reproducibility we considered algorithms whose implementation is available in R. In particular we consider the following algorithms: Semi-Interleaved HITON-PC local discovery structure learning algorithms (HPC) [24], incremental association MB constraint-based structure learning algorithm (IAMB) [25], the Fast-IAMB version of IAMB (FIAMB), Grow-Shrink (GS) constraint-based structure learning algorithm [26], the PC version implemented in the `pcalg` package (PCalg) and the Granger test (GRA) provided by the `lmtest` package [27]. For each of the 200 test multivariate time series and for each method, we measure the accuracy of the prediction of the causal directionality for 40 edges (about half with existing link and half with no link) and we compute the related Balanced Error Rate (BER)[5]. The adoption of BER, that equally weights errors in sensitivity and specificity, is justified by the unbalancedness of the classification task. A summary of the inference accuracy (distribution of the BERs) of the assessed methods is reported in Figure 4. On the left (right) we report the BER distribution over the time series whose associated DAGs has less (more) than 100 nodes. Some methods were computationally unfeasible for large dimension and are therefore not reported on the right plot. From the summary we see that caD2C outperforms significantly all the state-of-the-art methods (the lower the BER the better). Its average AUC is 0.81. By testing caD2C vs. standard D2C (the second best method) and PC it appears that caD2C outperforms significantly both D2C and P2C for both settings (p-value of the paired t-test < 0.025).

VI. Conclusion and future works

In recent years, we assisted to an adoption of learning techniques for inferring causal structure from data. However, most of those techniques deal with cause-effect pairs and their extension to settings characterised by more than two variables is not evident, e.g. because of confounding effects [6]. We proposed the addition of information-theoretic features in order to take into account the multivariate context. Experimental results showed that the resulting technique is able to outperform state-of-the-art methods in large-variate temporal tasks. Current work consists in benchmarking this technique against other state-of-the-art methods in the CauseMe[7] platform: preliminary

[3]https://github.com/gbonte/D2C

[4]Note that in the code it is possible to approximate the empirical distributions by using only a random sample of the K^2 pairs

[5]The balanced error rate formula is BER = 0.5*(FP/(TN+FP) + FN/(FN+TP)) where FP (FN) stands for the number of False Positives (Negatives) and TP (TN) stands for the number of True Positives (Negatives).

[6]see also chapter 6 in the recent "Cause-effect pairs in machine learning" book whose draft is available in http://causality.chalearn.org/experimental-design

[7]https://causeme.uv.es/

$$
\begin{aligned}
Y_{t+1}[j] &= -0.4\frac{(3-\bar{Y}_t[\mathcal{N}_j]^2)}{(1+\bar{Y}_t[\mathcal{N}_j]^2)} + 0.6\frac{3-(\bar{Y}_{t-1}[\mathcal{N}_j]-0.5)^3}{1+(\bar{Y}_{t-1}[\mathcal{N}_j]-0.5)^4} + W_{t+1}[j] \\
Y_{t+1}[j] &= (0.4 - 2\exp(-50\bar{Y}_{t-5}[\mathcal{N}_j]^2))\bar{Y}_{t-5}[\mathcal{N}_j] + (0.5 - 0.5\exp(-50\bar{Y}_{t-9}[\mathcal{N}_j]^2))\bar{Y}_{t-9}[\mathcal{N}_j] + W_{t+1}[j] \\
Y_{t+1}[j] &= 1.5\sin(\pi/2\bar{Y}_{t-1}[\mathcal{N}_j]) - \sin(\pi/2\bar{Y}_{t-2}[\mathcal{N}_j]) + W_{t+1}[j] \\
Y_{t+1}[j] &= 2\exp(-0.1\bar{Y}_t[\mathcal{N}_j]^2)\bar{Y}_t[\mathcal{N}_j] - \exp(-0.1\bar{Y}_{t-1}[\mathcal{N}_j]^2)\bar{Y}_{t-1}[\mathcal{N}_j] + W_{t+1}[j] \\
Y_{t+1}[j] &= -2\bar{Y}_t[\mathcal{N}_j]I(\bar{Y}_t[\mathcal{N}_j] < 0) + 0.4\bar{Y}_t[\mathcal{N}_j]I(\bar{Y}_t[\mathcal{N}_j] < 0) + W_{t+1}[j] \\
Y_{t+1}[j] &= 0.8\log(1+3\bar{Y}_t[\mathcal{N}_j]^2) - 0.6\log(1+3\bar{Y}_{t-2}[\mathcal{N}_j]^2) + W_{t+1}[j] \\
Y_{t+1}[j] &= (0.4 - 2\cos(40\bar{Y}_{t-5}[\mathcal{N}_j])\exp(-30\bar{Y}_{t-5}[\mathcal{N}_j]^2))\bar{Y}_{t-5}[\mathcal{N}_j] + (0.5 - 0.5\exp(-50\bar{Y}_{t-9}[\mathcal{N}_j]^2))\bar{Y}_{t-9}[\mathcal{N}_j] + W_{t+1}[j] \\
Y_{t+1}[j] &= (0.5 - 1.1\exp(-50\bar{Y}_t[\mathcal{N}_j]^2))\bar{Y}_t[\mathcal{N}_j] + (0.3 - 0.5\exp(-50\bar{Y}_{t-2}[\mathcal{N}_j]^2))\bar{Y}_{t-2}[\mathcal{N}_j] + W_{t+1}[j] \\
Y_{t+1}[j] &= 0.3\bar{Y}_t[\mathcal{N}_j] + 0.6\bar{Y}_{t-1}[\mathcal{N}_j] + \frac{(0.1 - 0.9\bar{Y}_t[\mathcal{N}_j] + 0.8\bar{Y}_{t-1}[\mathcal{N}_j])}{(1+\exp(-10\bar{Y}_t[\mathcal{N}_j]))} + W_{t+1}[j] \\
Y_{t+1}[j] &= \text{sign}(\bar{Y}_t[\mathcal{N}_j]) + W_{t+1}[j] \\
Y_{t+1}[j] &= 0.8\bar{Y}_t[\mathcal{N}_j] - \frac{0.8\bar{Y}_t[\mathcal{N}_j]}{(1+\exp(-10\bar{Y}_t[\mathcal{N}_j]))} + W_{t+1}[j] \\
Y_{t+1}[j] &= 0.3\bar{Y}_t[\mathcal{N}_j] + 0.6\bar{Y}_{t-1}[\mathcal{N}_j] + \frac{(0.1 - 0.9\bar{Y}_t[\mathcal{N}_j] + 0.8\bar{Y}_{t-1}[\mathcal{N}_j])}{(1+\exp(-10\bar{Y}_t[\mathcal{N}_j]))} + W_{t+1}[j] \\
Y_{t+1}[j] &= 0.38\bar{Y}_t[\mathcal{N}_j](1 - \bar{Y}_{t-1}[\mathcal{N}_j]) + W_{t+1}[j] \\
Y_{t+1}[j] &= \begin{cases} -0.5\bar{Y}_t[\mathcal{N}_j] & \text{if} \quad \bar{Y}_t[\mathcal{N}_j] < 1 \\ 0.4\bar{Y}_t[\mathcal{N}_j] \end{cases} \\
Y_{t+1}[j] &= \begin{cases} 0.9\bar{Y}_t[\mathcal{N}_j] + W_{t+1}[j] & \text{if} \quad |\bar{Y}_t[\mathcal{N}_j]| < 1 \\ -0.3\bar{Y}_t[\mathcal{N}_j] + W_{t+1}[j] \end{cases} \\
Y_{t+1}[j] &= \begin{cases} -0.5\bar{Y}_t[\mathcal{N}_j] + W_{t+1}[j] & \text{if} \quad x_t = 1 \\ 0.4\bar{Y}_t[\mathcal{N}_j] + W_{t+1}[j] \end{cases} \\
x_{t+1} &= 1 - x_t, x_0 = 1 \\
Y_{t+1}[j] &= \sqrt{0.000019 + 0.846 * (\bar{Y}_t[\mathcal{N}_j]^2 + 0.3\bar{Y}_{t-1}[\mathcal{N}_j]^2 + 0.2\bar{Y}_{t-2}[\mathcal{N}_j]^2 + 0.1\bar{Y}_{t-3}[\mathcal{N}_j]^2)}W_{t+1}[j]
\end{aligned}
$$

TABLE I

CROSS-SECTIONAL AND TEMPORAL SERIES: $\mathcal{N}_j$ DENOTES THE INDICES OF THE SET OF TIME SERIES WHICH ARE NEIGHBORS OF THE jTH COMPONENT. $\bar{Y}_t[\mathcal{N}_j]$ STANDS FOR THE AVERAGE OF THE VALUE OF THE NEIGHBORING SERIES AT TIME t. THE COVARIANCE MATRIX OF THE GAUSSIAN NOISE VECTOR W IS DIAGONAL.

results are promising for real series. Future work will focus on extending the approach to multi classification tasks, e.g. by extending the labelling in order to take into consideration more structural causal dependencies, like the ancestry or the offspring.

ACKNOWLEDGEMENT

The author acknowledges the support of the project WALINNOV 2017 – N 1710030 - CAUSEL funded by the Walloon Region of Belgium and the project "MACHU-PICCHU: Machine Learning for Predictive and Causal modelling of Churn" funded by INNOVIRIS, Brussels (B).

REFERENCES

[1] M. Eichler, *Causality, statistical perspectives and Applications*. Wiley, 2012, ch. Causal inference in time series analysis.

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

[3] P. Spirtes, C. Glymour, and R. Scheines, *Causation, Prediction and Search*. Berlin: Springer Verlag, 2000.

[4] D. Koller and N. Friedman, *Probabilistic Graphical Models*. The MIT Press, 2009.

[5] D. Lopez-Paz, K. Muandet, B. Schölkopf, and I. O. Tolstikhin, "Towards a learning theory of cause-effect inference." in *ICML*, ser. JMLR Workshop and Conference Proceedings, F. R. Bach and D. M. Blei, Eds., vol. 37. JMLR.org, 2015, pp. 1452–1461. [Online]. Available: http://dblp.uni-trier.de/db/conf/icml/icml2015.html#Lopez-PazMST15

[6] G. Bontempi and M. Flauder, "From dependency to causality: A machine learning approach," *Journal of Machine Learning Research*, vol. 16, pp. 2437–2457, 2015. [Online]. Available: http://jmlr.org/papers/v16/bontempi15a.html

[7] P. Hoyer, D. Janzing, J. Mooij, J. Peters, and B. Scholkopf, "Nonlinear causal discovery with additive noise models." in *Advances in Neural Information Processing Systems*, 2009, pp. 689–696.

[8] P. Daniusis, D. Janzing, J. Mooij, J. Zscheischler, B. Steudel, K. Zhang, and B. Schölkopf, "Inferring deterministic causal relations." in *Proceedings of UAI*, 2010, pp. 143–150.

[9] D. Janzing, J. Mooij, K. Zhang, J. Lemeire, J. Zscheischler, P. Daniusis, B. Steudel, and B. Scholkopf, "Information-geometric approach to inferring causal directions." *Artificial Intelligence*, 2012.

[10] S. Shimizu, P. Hoyer, A. Hyvärinen, and A. Kerminen, "A linear, non-gaussian acyclic model for causal discovery." *Journal of Machine Learning Research*, vol. 7, pp. 2003–2030, 2006.

[11] J. Mooij, O. Stegle, D. Janzing, K. Zhang, and B. Schölkopf, "Probabilistic latent variable models for distinguishing between cause and effect." in *Advances in Neural Information Processing Systems*, 2010.

[12] A. Statnikov, M. Henaff, N. Lytkin, and C. F. Aliferis., "New methods for separating causes from effects in genomics data," *BMC Genomics*, vol. 13, no. S22, 2012.

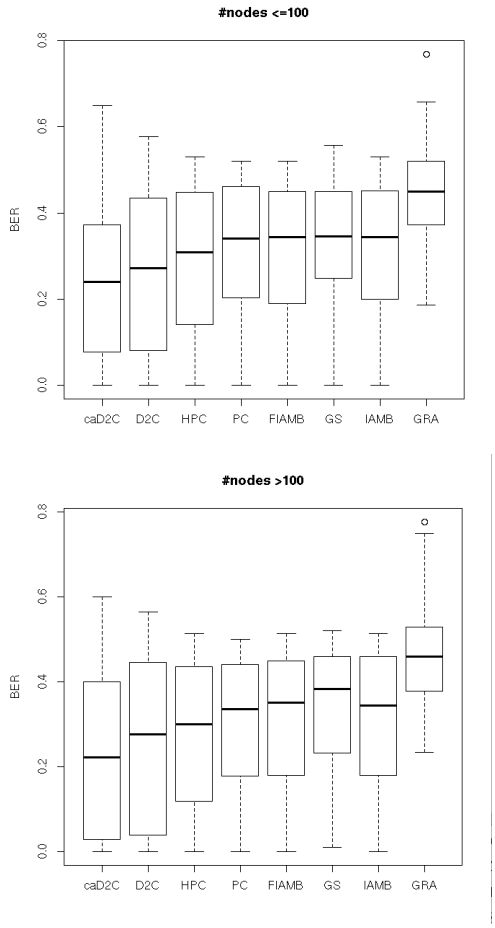

Fig. 4. Distribution of the BER accuracy for the 500 test time series. Above (below) we report the BER distribution over the time series whose associated DAG has less (more) than 100 nodes.

[13] I. Guyon, "Results and analysis of the 2013 ChaLearn cause-effect pair challenge," in *Proceedings of NIPS 2013 Workshop on Causality: Large-scale Experiment Design and Inference of Causal Mechanisms*, 2014.

[14] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97, no. 1-2, pp. 273–324, 1997.

[15] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[16] W. J. McGill, "Multivariate information transmission," *Psychometrika*, vol. 19, 1954.

[17] J. Peters, D. Janzing, and b. Scholkopf, *Elements of causal inference*. The MIT Press, 2017.

[18] Y. Chikahara and A. Fujino, "Causal inference in time series via supervised learning," in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)*, 2018, pp. 2042–2048.

[19] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: John Wiley, 1990.

[20] J. Pellet and A. Elisseeff, "Using markov blankets for causal structure learning," *Journal of Machine Learning Research*, vol. 9, pp. 1295–1342, 2008.

[21] G. Bontempi and P. Meyer, "Causal filter selection in microarray data," in *Proceedings of ICML*, 2010.

[22] G. Bontempi, B. Haibe-Kains, C. Desmedt, C. Sotiriou, and J. Quackenbush, "Multiple-input multiple-output causal strategies for gene selection." *BMC Bioinform.*, vol. 12, p. 458, 2011. [Online]. Available: http://dblp.uni-trier.de/db/journals/bmcbi/bmcbi12.html#BontempiHDSQ11

[23] X.-Y. Liu, J. Wu, and Z.-H. Zhou, "Exploratory undersampling for class-imbalance learning." *IEEE Trans. Systems, Man, and Cybernetics, Part B*, vol. 39, no. 2, pp. 539–550, 2009. [Online]. Available: http://dblp.uni-trier.de/db/journals/tsmc/tsmcb39.html#LiuWZ09

[24] I. Tsamardinos, C. Aliferis, and A. Statnikov, "Time and sample efficient discovery of markov blankets and direct causal relations," in *Proceedings of KDD*, 2003, pp. 673–678.

[25] ——, "Algorithms for large scale markov blanket discovery," in *Proceedings of FLAIRS*, 2003.

[26] D. Margaritis, "Learning bayesian network model structure from data," Ph.D. dissertation, School of Computer Science, Carnegie-Mellon University, Pittsburgh, PA, 2003.

[27] A. Zeileis and T. Hothorn, "Diagnostic checking in regression relationships," *R News*, vol. 2, no. 3, pp. 7–10, 2002. [Online]. Available: https://CRAN.R-project.org/doc/Rnews/